
Article:
Das, K., Pareek, B., Brown, S. orcid.org/0000-0002-4853-9115 et al. (1 more author) (2022) A semi-parametric Bayesian dynamic hurdle model with an application to the health and retirement study. Computational Statistics, 37 (2). pp. 837-863. ISSN 0943-4062

https://doi.org/10.1007/s00180-021-01143-x

A Semi-parametric Bayesian Dynamic Hurdle Model with an Application to the Health and Retirement Study

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

June 7, 2021

Abstract

All developed countries are facing the problem of providing affordable and high quality healthcare in recent years. This is due to the combination of an ageing population and the technological advancement in health science which leads to an increased life expectancy. This will affect the future use of hospital inpatient and outpatient services which in turn will place a significant stress on the economy since most medical services for the elderly are apportioned and funded under a national system. Thus, understanding the demand for healthcare and other key factors influencing the demand is crucial to better serve citizens. Hospital admission is considered to be a key proxy of the demand for healthcare, especially in the context of ageing populations as experienced globally. However, modeling hospital admissions, although very important, is often complicated by zero-inflation, by the covariates with time-varying effects, and by the necessity of borrowing information across individuals. Additionally, the rate of hospital admissions might differ between the group of individuals who have been hospitalized before and the group yet to be hospitalized. Also when individuals are clustered based on their baseline self-assessed health status, the distribution of hospital admissions and its relation to predictors may be quite different across and within different groups. In this paper we propose a unified Bayesian dynamic hurdle model which accommodates these features of the data in a semi-parametric approach. We analyze the data collected by the United States Health and Retirement Study in which the rate of hospital admissions varies across different self-assessed health (SAH) groups. Simulation studies are performed for assessing the usefulness of the proposed model.

Key Words: Bayesian models, Dynamic hurdle model, Hospital admissions, Lasso, Matrix Stick-Breaking Process, Zero-inflated data.

Kiranmoy Das is an Associate Professor, Indian Statistical Institute, Kolkata, India; Bhuvanesh Pareek is an Assistant Professor, Indian Institute of Management, Indore, India; Sarah Brown is a Professor of Economics at University of Sheffield, UK; Pulak Ghosh is a Professor, Department of Decision Sciences & Centre for Public Policy, Indian Institute of Management, Bangalore, India.

1 Introduction

The fact that the world population is ageing 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, the latest estimates of the United Nations predict that between 2015 and 2030 the number of people in the world aged 60 years or more is projected to grow by 56%, from 901 million to 1.4 billion. With respect to the increase in life expectancy, the World Health Organization (2016) reports that the life expectancy in the United States is 79.3 in 2015, compared to 66.6 in 1960, and 46.3 in 1900. As a result, the composition of the population pyramid in the United States is changing, with a widening part at the top for the older age groups. The United States, however, is not alone in facing the implications of an ageing population. It is apparent that with such a demographic change the demand for healthcare, as well as changes in the nature of healthcare demand will be a major concern globally to researchers as well as to policy-makers.

Hospital admission is an important aspect of the demand for healthcare. Approximately, one-third of all healthcare expenditure in OECD (Organization for Economic Co-operation and Development) countries is spent on inpatient services. It is apparent that as chronic conditions become more prevalent in the context of an ageing population, understanding the needs associated with inpatient care and hospital visits becomes increasingly important. Since an ageing population changes the nature of healthcare demand with more emphasis on hospitalization, and thus more challenges are faced by the healthcare systems globally, further analysis into hospital admissions at the individual level seems to be particularly warranted.

However, modeling hospital admissions is challenging since it is often complicated by several non-standard features, e.g., zero-inflation, non-uniform admission rates over time, covariates with time-varying effects, the necessity of borrowing information across individuals for clustering similar individuals. In addition, certain health conditions may lead to groups of individuals having similar hospital admission rates. There has been some research by health economists on estimating healthcare demand (Atella and Deb 2008; Duan et al. 1983; Mukherji et al. 2016). Our paper is concerned with the problem of estimating the demand for healthcare through modeling hospital admissions. It advances the previous studies (Biswas and Das 2019; Biswas et al. 2020; Mukherji et al. 2016) by explicitly developing a unified model which accommodates all the above issues.

1.1 The Health and Retirement Study (HRS)

We use data from the Health and Retirement Study (HRS) conducted by the University of Michigan. This is a longitudinal survey of Americans over the age of 50 years, with a follow-up frequency of every two years (referred to as waves hereafter). In this paper, we use data for 10 waves for 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 Ageing. Our study is based on 2630 individuals who are observed in all the 10 waves, and survived until the end of the study. In addition to the main outcome variable i.e. the number of hospital admissions, information is provided related to the personal and socio-economic characteristics of the respondents, including data related to a range of physical health conditions, as well as the financial health conditions. Table 1 shows all the variables we consider in our analysis, and it also specifies the role of each variable in our model.

Based on some existing studies using the HRS data (Biswas and Das 2019, 2021), we split the set of covariates into two parts, (i) covariates with time-invariant effects on the response, and (ii) the covariates with time-varying effects on the response (see Table 1). We note that some of the covariates related to different chronic diseases usually evolve over time, but for an ageing population like ours, they are almost unchanged over time (as observed in our dataset), and their effects can be assumed to be time-invariant. Additionally, there is information on health insurance: (i) health insurance related to employment (Y/N); (ii) government insurance (Medicaid or Medicare) (Y/N); and (iii) other private health insurance (Y/N). For the older individuals the effects of different types of insurance can be assumed to be time-invariant (Biswas and Das 2021).

The four covariates with time-varying effects on the count of hospital admissions are: (i) body mass index (BMI); (ii) total value of assets; (iii) total value of debt; and (iv) total household income. Note that total financial 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; personal vehicles; net value of any business; and other assets. The measure of total debt includes all mortgages/land contracts; other home loans; and other debt including credit cards. BMI is an important covariate given its known relationship with obesity and poor health conditions. Since obesity can result in many other chronic diseases as the individuals become older, the effect of BMI on the count of hospital admission is time-varying. Three financial covariates, viz, total assets, total debt and total household income are also very important predictors for health conditions as evidenced in a large existing literature exploring the relationship between a range of household financial outcomes and health. For example, Adams et al. (2003), Hurd and Kapteyn (2003), Michaud et al. (2008) generally find a positive association between better health and household wealth over time. There is also a rich literature on exploring the relationship between health and debt. Drentea and Lavrakas (2000) find that both credit card debt and stress regarding debt are inversely associated with good health. Brown et al. (2005) find that unsecured debt is inversely related to psychological wellbeing. Finally, with respect to income, a number of studies have explored the time-varying relationship between income and health. Pelkowski and Berger (2004) use data from the HRS and find that permanent health problems have a significant effect on labor market participation, wages and hours for both men and women. We note that all monetary variables are entered in natural logarithm form and are expressed in 2010 prices.

1.2 Analytical Issues of the HRS Data for Modeling Hospital Admissions

The primary variable of interest in our analysis is the number of hospital admissions since the previous interview. 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. Figure S.1 (in the web-appendix) shows the number of people in the sample and their aggregate count of hospitalization for the 10 waves. In Figure 1 we show the counts of individuals with different numbers of hospital admission over the waves. As is evident, the number of people having zero hospital admission is high as compared with non-zero across the waves, and is highest in waves 1 and 2. The number of non-zero hospital admissions increases in the later waves. While there has been some work on modeling hospital admissions (Atella and Deb 2008; Deb and Trivedi 1997; Mukherji et al. 2016; Winkelmann 2004), we perform a comprehensive study of the HRS data by addressing a number of additional complexities in a unified framework. The major challenges are summarized as follows.

1. There are a considerable proportion of zero counts for hospital admissions in the HRS data, reflecting the fact that a significant proportion of individuals are not admitted to hospital in a given year, at least in the initial waves. Given the considerable amount of zero observations that are observed in the number of hospital admissions at the individual level, existing studies have developed zero inflated approaches for modeling the counts of hospital admissions to account for the excess zeros (Atella and Deb 2008; Deb and Trivedi 1997; Winkelmann 2004). However, previous research does not allow for the ordering information, which is likely to be inherent and very informative in such data: i.e., an individual who was hospitalized 5 times may be regarded as more serious and having a higher probability of being admitted to hospital compared to an individual who was hospitalized only once in the same year. Here, we propose an ordered logistic regression approach, i.e. a proportional odds model, to analyze the number of hospital admissions.

2. Existing literature works on the assumption that the probability of hospital admission is uniform over time. However, in the context of an ageing 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 hospitalized before. 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 the likelihood of requiring repeated hospital care will vary. 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) analyse Hospital Episode Data for a specific region of the United Kingdom, and 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'. In a similar vein, Banerjee et al. (2010) explore persistence in healthcare utilisation using dynamic panel data models. Their analysis of the 2000-2004 US Medical Expenditure Panel Survey (MEPS)

endorses a dynamic modeling approach, with inpatient hospitalization at the initial period found to have a large positive and significant impact on current hospitalization. From a modeling 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. Baetschmann and Winkelmann (2016) developed a dynamic hurdle model to allow the hazard rate for the time to first event (hospital admission, in our case) to differ from the hazard rate from the first to the second, second to the third event, and so forth. We generalize the dynamic hurdle model of Baetschmann and Winkelmann (2016) to accommodate our setting.

3. Finally, Self-Assessed-Health (SAH) which is a categorical variable, is a very good proxy for health risk, where respondents rate their own general health on a response scale from poor to excellent. Idler and Benyamini (1997) showed that SAH is predictive of mortality even after conditioning on objective measures of health. SAH status in the HRS has 5 categories: poor (1), fair (2), good (3), very good (4) and excellent (5). In Figure S.2 (in the web-appendix), through a heatmap, we show the correlation of average counts of individuals with different numbers of hospital admissions across the 10 waves and SAH categories. As expected, the number of non-zero hospital admissions increases as the SAH condition deteriorates.

Considering SAH simply as a covariate is rather restrictive and does not use the inherent information offered by this covariate. For example, people in the same SAH category might be more similar over time across the important time-varying covariates, while at the same time people across different categories may also have similarity across predictors leading to a "local" and "global" similarity. To address this issue, we first "group" individuals based on their baseline SAH, and model the hospital admissions for each group over time. For the count of hospital admissions, different SAH groups have some similarities as is evident in Figure S.3 (in the web-appendix). We note that for our sample data, SAH conditions change for less than 10% individuals, and hence "grouping" with respect to the baseline SAH condition is not problematic in our case. Then the relation of the time-varying predictors with the response variables across SAH groups are first expressed as polynomial functions over time. Next, to address the "local" and "global" similarity aspect, a Matrix Stick-Breaking Process (MSBP) prior (Dunson et al. 2008) is considered on the coefficients of the polynomial functions used for modeling the time-varying effects of the four covariates mentioned earlier. By specifying an MSBP prior distribution on the coefficients we allow: (i) multiple shrinkage of a large set of model parameters; and (ii) local and global clustering by borrowing information within and across different groups (Das et al. 2021).

The rest of this paper is organized as follows. In Section 2, we propose a zero-inflated dynamic hurdle proportional odds model for the count of hospital admissions. In Section 3, we propose different shrinkage priors for the model parameters. The joint posterior density and computational details are provided in Section 4. The results for the HRS data analysis are presented and discussed in Section 5. Finally Section 6 concludes.

2 The Proposed Model

Let r denote the baseline SAH group (with $r = 1$ for "poor SAH", and $r = 5$ for "excellent SAH"). Further, let Y_{irt} be the count of hospital admissions at wave t ($t = 1, 2, \ldots, T$) for the i-th individual ($i = 1, 2, \ldots, N$) belonging to the r-th baseline SAH group. Thus, $Y_{irt} = 0$ indicates that the i-th individual from the r-th SAH category was not admitted to hospital at the t-th wave. As expected, we observe excess zeros in Y_{irt}.

For each observed response Y_{irt} we define:

$$Y_{irt} \begin{cases} = 0, & \text{with probability} = 1 - p_{irt}, \\ \sim G(Y_{irt}), & \text{with probability} = p_{irt}, \end{cases} \tag{1}$$

where p_{irt} is the probability of non-zero hospital admission, i.e., $p_{irt} = P(Y_{irt} > 0)$, and G is the distribution of the count of hospital visits conditional on non-zero hospitalization. Based on equation (1), we obtain the following probability distribution for Y_{irt}:

$$Y_{irt} \sim (1 - p_{irt}) 1_{[Y_{irt}=0]} + p_{irt} G(Y_{irt} | Y_{irt} > 0). \tag{2}$$

2.1 The Model for p_{irt}

For modeling the proportion of non-zeros we consider a Bayesian Probit model; and thus express p_{irt} as: $\Phi^{-1}(p_{irt}) = \mathbf{x}_{it}^T \boldsymbol{\delta} + \eta_i$, where Φ denotes the cumulative distribution function of a standard normal variable, $\mathbf{x}$ is the set of all covariates, and $\boldsymbol{\delta}$ is the vector of regression coefficients. We note that the individual-specific random effects i.e. the η_is capture the temporal dependence among the measurements from the same individual. For capturing temporal dependence, an auto-regressive term is typically introduced in a linear model. However, we capture such dependence through random effects, and that reduces the computational burden (Das and Daniels 2014). We also allow the probabilities of the non-zero responses (and hence the zero responses) to vary over the waves for the obvious reason.

For the Bayesian computation, we develop a data-augmentation technique following Albert and Chib (1993). We define a latent continuous random variable, say Z_{irt}, so that,

$$Y_{irt} \begin{cases} = 0, & \text{for } \; Z_{irt} < 0, \\ \neq 0, & \text{for } \; Z_{irt} \geq 0. \end{cases} \tag{3}$$

Then, we develop a linear random effects model for Z_{irt} as follows:

$$Z_{irt} = \mathbf{x}_{it}^T \boldsymbol{\delta} + \eta_i + \epsilon_{it}, \tag{4}$$

where the residual errors ϵ_{it} are iid $N(0, 1)$. A Gibbs sampler algorithm similar to the one proposed in Albert and Chib (1993) can be developed for Bayesian computation. Details of the Gibbs sampler are given in Section 4.

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

Table 2 illustrates that the distributions of the count of hospital admissions, i.e., $G(Y_{irt}|Y_{irt} > 0)$ are different for (i) the individuals who visit the hospital for the first time at the t-th wave and; (ii) those who have been hospitalized before the t-th wave. We use a proportional odds model with a dynamic hurdle component (Baetschmann and Winkelmann 2016) for the first scenario, and use a simple proportional odds model for the second scenario. The proportional odds model preserves the ordering information (Brant 1990).

Case:(i): Individuals being hospitalised for the first time

Let t_i be the wave where the i-th individual visits hospital for the first time. Note that since the waves are of length 2 years, wave t_i corresponds to the time interval $[2(t_i - 1), 2t_i]$. Our dynamic hurdle model is based on the assumption that in this time interval, the total count for hospital admissions depends on the time for the first hospital admission. Suppose the first hospital admission for individual i occurs at time T_i; for $2(t_i - 1) < T_i < 2t_i$. Note that T_i is also a random variable. Now, the joint distribution of Y_{irt_i} and T_i can be written as:

$$Pr(Y_{irt_i} = k, T_i) = Pr(Y_{irt_i} = k|T_i)f_1(T_i) = Pr[Y_{irt_i}(T_i, 2t_i) = k - 1]f_1(T_i);$$

where $Y_{irt_i}(T_i, 2t_i)$ denotes the number of hospital visits for the i-th individual in the t_i-th wave within the time interval $(T_i, 2t_i)$, and f_1 denotes the probability density function (pdf) of the time to the first hospital visit. The marginal distribution of Y_{irt_i}, thus, can be obtained as:

$$Pr(Y_{irt_i} = k) = \int_{2(t_i-1)}^{2t_i} Pr(Y_{irt_i}(T_i, 2t_i) = k - 1)f_1(T_i)dT_i. \tag{5}$$

We consider T_i as the time to elapse before the event of interest (hospital admission) and thus needs to be modeled as a survival endpoint. Both parametric and semiparametric models are available in the literature for modeling time to event. Commonly used parametric models include the Exponential and Weibull (Carroll 2003; Zhang 2016), which are attractive in their simplicity and can be easily interpreted. Hence, we model this waiting time by a Weibull distribution. This is a fully parametric approach (an alternative to the semi-parametric Cox model) for modeling the waiting time. Specifically, we consider f_1 as the density of a Weibull distribution with parameters η_1 and η_2, and thus, $T_i \sim \text{Weibull}(\eta_1, \eta_2)$.

The cumulative density function (cdf) of $Y_{irt_i}(T_i, 2t_i)$ is modeled by the following proportional odds model (McCullagh 1980) which can capture dependence among the ordered categories of SAH, time, and the count of hospital admissions:

$$\text{logit}(Pr(Y_{irt_i}(T_i, 2t_i) \leq k)) = \sum_{j=1}^{J} \alpha_{jkr}(t_i)x_{ijt_i} + \sum_{j'=1}^{J'} \beta_{j'kr}z_{ij't_i} + b_i, \tag{6}$$

for $k = 1, 2, \ldots, K$. In our application, the maximum number of hospital admissions for any individual at any wave is 15, and therefore, we consider K=15. Note that for each k, the variable

of interest (i.e. count of hospital admissions) is now binary (less than k, or higher than k), and then we model the log-odds by the linear mixed model given in (6). We consider J covariates with time-varying effects; the effect of the j-th covariate on the log-odds at wave t_i for the r-th baseline SAH group is $\alpha_{jkr}(t_i)$. Additionally, we consider J' covariates with time-invariant effects. Note that the individual-specific random effects, i.e. the b_is capture the temporal correlation among the log-odds for the same individuals at different waves (t_i), and also for different values of k.

For modeling the dependence among the non-zero response probabilities and the log odds, we jointly model (η_i, b_i) by a Bivariate Normal (BN) density with mean vector=$\mathbf{0}$ and $\text{var}(\eta_i) = \sigma^2_\eta$, $\text{var}(b_i) = \sigma^2_b$, and $\text{correlation}(\eta_i, b_i) = \rho$.

Finally, for modeling the time-varying coefficients $\alpha_{jkr}(t)$, we use penalized splines which are smooth and flexible functions (Ruppert et al. 2003) as follows:

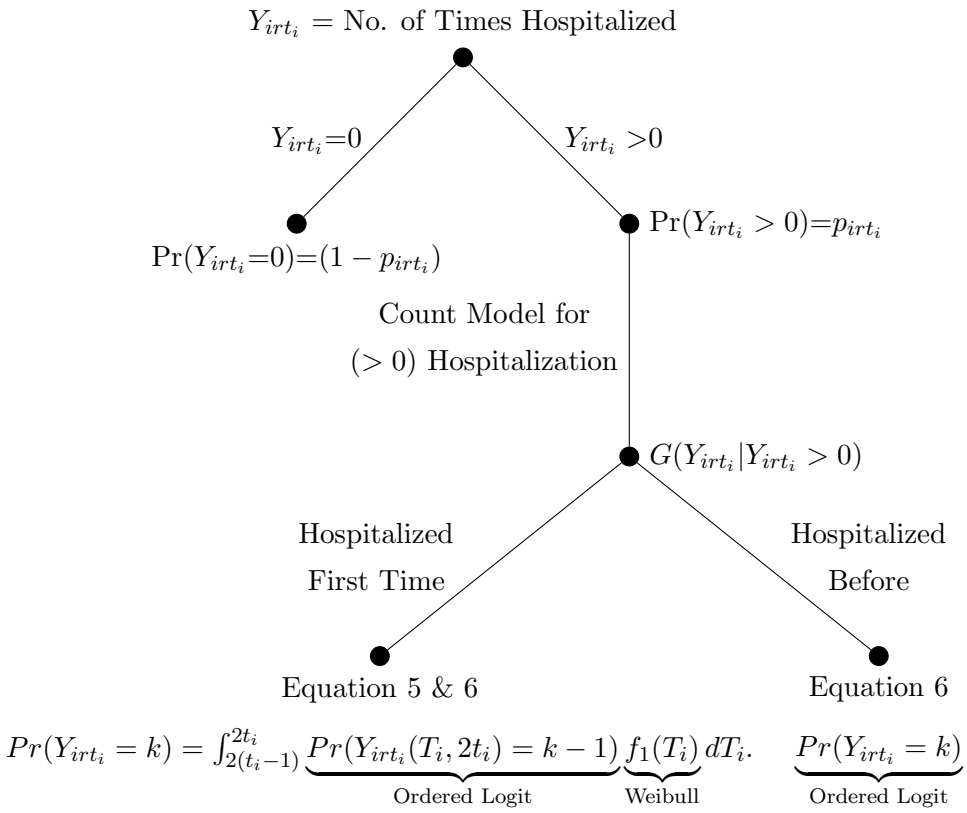
$$\alpha_{jkr}(t_i) = b_{jkr0} + b_{jkr1}t_i + b_{jkr2}t_i^2 + ... + b_{jkrg}t_i^g + \sum_{s=1}^{S} c_{jkrs}(t_i - \mathcal{T}_s)_+^g, \tag{7}$$

where $(x)_+^g = x^g I(x > 0)$ and $(\mathcal{T}_1 < \mathcal{T}_2 < ... < \mathcal{T}_S)$ is a fixed set of knots, typically selected based on evenly spaced sample quantiles (Ruppert et al. 2003). The optimal order g and the optimal number of knots (S) are obtained using the Deviance Information Criteria (DIC).

Case:(ii): Individuals who have been hospitalized before

Here, the t-th wave indicates the wave after the i-th individual visited hospital for the first time. For such waves, we model G by a proportional odds model. The cdf of Y_{irt} is also modeled by the proportional odds model given in equation (6).

Below, we graphically depict our model to aid understanding.



3 Proposed Prior Structures

3.1 Matrix Stick-Breaking Priors

Recall that our model in (7) is very flexible since the regression coefficients depend on SAH group (r) and the count of hospital admissions (k). Thus, we have a large number of coefficients, and hence we need a shrinkage prior for estimating these coefficients. We also note that different SAH groups can be "similar" in the sense that some of the regression coefficients for these groups over time can be exactly the same, or can take similar values. To achieve this, we implement the Matrix Stick-Breaking Process (MSBP) prior proposed in Dunson et al. (2008). We use the MSBP as a shrinkage prior where the parameters for all J covariates with time-varying effects are shared across different r and k.

Define a $1 \times (g+1)$ vector of polynomial coefficients from (7), $\mathbf{b}_{jkr} = [b_{jkr0}, b_{jkr1}, \ldots, b_{jkrg}]$, for $j = 1, 2, \ldots, J$. Define $J(g+1) \times 1$ vector $b_{kr} = [b_{1kr}, b_{2kr}, \ldots, b_{Jkr}]^T$. We assume

$$\begin{aligned} \mathbf{b}_{kr} &\sim G_{kr}^{(b)} = \sum_{h=1}^{N_b} \pi_{krh}^{(b)} \delta_{\xi_{kh}^{(b)}}; \quad k = 1, 2, \ldots, K; \quad r = 1, 2, \ldots, 5; \\ \xi_{kh}^{(b)} &\sim G_k^{(b)}; \quad G_k^{(b)} \sim N(\mu_0, \Sigma_0), \end{aligned} \tag{8}$$

where δ_x denotes a point mass at the $J(g+1) \times 1$ vector x. Define $\Xi^b = \left(\xi_{kh}^{(b)}\right)$ a matrix of order $K \times N_b$, the rows of which correspond to the parameters with the base distribution $G_k^{(b)}$; and the

columns correspond to the "clusters". The stick-breaking weights $\pi_{krh}^{(b)}$ are defined as follows:

$$\pi_{krh}^{(b)} = V_{krh}^{(b)} \prod_{s'<h} (1 - V_{krs'}^{(b)}); \;\; V_{krh}^{(b)} = U_{kh}^{(b)} W_{rh}^{(b)},$$
$$U_{kh}^{(b)} \overset{\text{iid}}{\sim} \text{Beta}(1, \delta_1^{(b)}); \;\; W_{rh}^{(b)} \overset{\text{iid}}{\sim} \text{Beta}(1, \delta_2^{(b)}). \tag{9}$$

The stick-breaking weights $\pi_{krh}^{(b)}$ control the dependence among the distributions $G_{kr}^{(b)}$. Note that we partition $\pi_{krh}^{(b)}$ into two components, $U_{kh}^{(b)}$ and $W_{rh}^{(b)}$, which allocate the vector of the polynomial coefficients from the k number of hospital visits and the r-th health status group to the h-th cluster. We need to take $V_{krN_b}^{(b)} = 1$, for all k, r; to make $G_{kr}^{(b)}$ a valid probability measure (Dunson et al. 2008).

Note that in the above prior, we consider the coefficients for the polynomial part of (7) for all J covariates with time-varying effects on the number of hospital admissions. These coefficients depend on SAH group r, and on the number of hospital admissions k, thus resulting in a large number of regression coefficients. By considering the MSBP prior on this set of parameters, we encourage the "exchange of information" across different SAH groups (r); and also perform a global shrinkage by allocating the coefficients for different k and r to the same cluster. Such a prior specification, as noted in Das et al. (2021), is advantageous since it provides a (prior) matching probability for $\boldsymbol{b}_{kr}$ and $\boldsymbol{b}_{kr'}$, for $r \neq r'$. Additionally, it increases the (prior) matching probability for $\boldsymbol{b}_{kr}$ and $\boldsymbol{b}_{kr'}$ given that they match for some other $k' \neq k$, i.e. $P(\boldsymbol{b}_{kr} = \boldsymbol{b}_{kr'} | \boldsymbol{b}_{k'r} = \boldsymbol{b}_{k'r'}) > P(\boldsymbol{b}_{kr} = \boldsymbol{b}_{kr'})$.

For the set of spline coefficients $\mathbf{c}_{kr} = [c_{1kr}, c_{2kr}, \ldots, c_{Jkr}]^T$, where $c_{jkr} = [c_{jkr1}, c_{jkr2}, \ldots, c_{jkrS}]$, we similarly specify the following MSBP prior:

$$\mathbf{c}_{kr} \sim G_{kr}^{(c)}; \;\; k = 1, 2, \ldots, K; \;\; r = 1, 2, \ldots, 5;$$
$$G_{kr}^{(c)} = \sum_{h=1}^{N_c} \pi_{krh}^{(c)} \delta_{\xi_{kh}^{(c)}}; \;\; \pi_{krh}^{(c)} = V_{krh}^{(c)} \prod_{s'<h} (1 - V_{krs'}^{(c)}); \;\; V_{krh}^{(c)} = U_{kh}^{(c)} W_{rh}^{(c)},$$
$$U_{kh}^{(c)} \overset{\text{iid}}{\sim} \text{Beta}(1, \delta_1^{(c)}); \;\; W_{rh}^{(c)} \overset{\text{iid}}{\sim} \text{Beta}(1, \delta_2^{(c)}),$$
$$\xi_{kh}^{(c)} \sim G_k^{(c)}; \;\; G_k^{(c)} \sim N\left(\mu, \frac{1}{\lambda^*} I\right); \;\; \lambda^* \sim \text{Gamma}(\kappa_1, \kappa_2). \tag{10}$$

We note that the above prior is very similar to the priors proposed in (8) and (9), except the prior for $G_k^{(c)}$. 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 $G_k^{(c)}$, and then placing a Gamma prior on the penalty parameter λ^* (Das and Daniels 2014).

3.2 Prior for the Time-Invariant Regression Coefficients

Since we have a large number of covariates with time-invariant effects on our response variable, we incorporate a shrinkage prior for these coefficients. Our approach is similar to the hierarchical

Bayes representation of Das (2016), and Park and Casella (2008). We consider the Lasso type shrinkage prior by specifying a Laplace prior on the coefficients. The hierarchical representation is based on the fact that the Laplace distribution is a scale mixture of normal distributions with an exponential scale distribution. Define $\boldsymbol{\beta}_{kr} = (\beta_{1kr}, \cdots, \beta_{J'kr})^T$. We add sparsity to the prior (i.e. consider a zero-inflated prior distribution) and consider the following hierarchical representation:

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

Here δ_0 is a point mass probability measure at 0. Note that the parameters $\pi_{j'}$s are updated from the data. The above specification shrinks the regression coefficients towards 0. 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 following Park and Casella (2008) for shrinkage.

4 The Joint Likelihood and Computational Details

Note that the joint likelihood function of the longitudinal responses Y_{irt}, and latent variables Z_{irt} can be expressed as follows:

$$L = \prod_{i,t}\int\int [I(Z_{irt} < 0)\phi(Z_{irt}|Y_{irt} = 0, \boldsymbol{\delta}, \eta_i) + I(Z_{irt} \geq 0)\phi(Z_{irt}|Y_{irt} \neq 0, \boldsymbol{\delta}, \eta_i)G(Y_{irt}|b_i)] \times f(\eta_i, b_i)d\eta_i db_i,$$

where $G(Y_{irt}|b_i)$ is the density described in Section 2.2 and $f(\eta_i, b_i)$ is the bivariate normal density for the random effects. Here, $I(A)$ denotes an indicator variable which takes value 1 if the event A occurs. The joint posterior distribution is obtained by multiplying the corresponding prior components to L. Note that for $\boldsymbol{\delta}$ (Section 2.1), we consider a multivariate normal prior with mean vector=0 and covariance matrix=$\sigma^2_\delta I$.

We denote the full conditional distribution of a parameter, say A, conditional on all the other parameters by $A|-$. Following the data-augmentation technique proposed in Albert and Chib (1993), we develop a Gibbs sampler for sampling the latent variables Z_{irt} and the coefficients $\boldsymbol{\delta}$, as follows:

(a) If $Y_{irt} > 0$, then $Z_{irt}|- \sim N^+(x^T_{it}\boldsymbol{\delta} + \eta_i, 1)$, where N^+ stands for a truncated normal density which is truncated at the left by 0.

(b) If $Y_{irt} = 0$, then $Z_{irt}|- \sim N^-(x^T_{it}\boldsymbol{\delta} + \eta_i, 1)$, where N^- stands for a truncated normal density which is truncated at the right by 0.

(c) $\boldsymbol{\delta}|- \sim N\left((\sum_{i=1}^{N}\sum_{t=1}^{T} x^T_{it}X_{it} + \frac{I}{\sigma^2_\delta})^{-1}x^T_{it}Z_{irt}, (\sum_{i=1}^{N}\sum_{t=1}^{T} x^T_{it}x_{it} + \frac{I}{\sigma^2_\delta})^{-1}\right)$.

Next, we develop the MCMC algorithm for the posterior computations based on the MSBP priors for the coefficients given in equation (7). We discuss the computational details for the

coefficients from the polynomial part, and computations are similar for the coefficients from the spline part.

Following Dunson et al. (2008), and Das et al. (2021) we introduce latent variables $R_{kr}^{(b)}$ from multinomial distributions with respective probabilities $(\pi_{krh}^{(b)})$. We define the following binary dummy variables for $r = 1, \ldots, 5$, $k = 1, \ldots, K$, and $h = 1, \ldots, N_b$:

$u_{krh} \sim \text{Bernoulli}(U_{kh}^{(b)})$, $w_{krh} \sim \text{Bernoulli}(W_{rh}^{(b)})$.

Now define $R_{kr}^{(b)} = min\,(h : 1 = u_{krh} = w_{krh})$. Recall that $R_{kr}^{(b)}$'s are distributed as multinomial distributions. We let $R_{kr}^{(b)}$ designate which $\xi_{kh}^{(b)}$ to choose as $\mathbf{b}_{kr}$. Hence $\mathbf{b}_{kr}$'s are determined by $R_{kr}^{(b)}$ and $\xi_{th}^{(b)}$. Once we sample $R_{kt}^{(b)}$ and $\xi_{kh}^{(b)}$, we can find the values of $\mathbf{b}_{kr}$'s. The forms of the full conditional distributions are given below.

1. The full conditional distribution of $R_{kr}^{(b)}$ is multinomial with

$$P(R_{kr}^{(b)} = h|-) \propto \pi_{krh}^{(b)} \times L. \tag{11}$$

2. The conditional for (u_{krh}, w_{krh}) sets $u_{krh} = w_{krh} = 1$ when $h = R_{kr}^{(b)}$; and otherwise samples with probabilities $p_{uw} = P(u_{krh} = u, w_{krh} = w)$, for $h = 1, 2, \ldots, R_{kr}^{(b)} - 1$, where
$p_{00} = \frac{U_{kh}^{(b)}(1-W_{rh}^{(b)})}{1-U_{kh}^{(b)}W_{rh}^{(b)}}$, $p_{10} = \frac{(1-U_{kh}^{(b)})(1-W_{rh}^{(b)})}{1-U_{kh}^{(b)}W_{rh}^{(b)}}$, $p_{01} = \frac{(1-U_{kh}^{(b)})W_{rh}^{(b)}}{1-U_{kh}^{(b)}W_{rh}^{(b)}}$.

3. The full conditional distributions for $U_{kh}^{(b)}$ for $h < N_b$ are given as:

$$U_{kh}^{(b)}|- \sim \text{Beta}\left(1 + \sum_{t:R_{kr}^{(b)} \geq h} u_{krh}, \delta_1^{(b)} + \sum_{t:R_{kr}^{(b)} \geq h} (1 - u_{krh})\right). \tag{12}$$

Similarly for $h < N_b$,

$$W_{rh}^{(b)}|- \sim \text{Beta}\left(1 + \sum_{k:R_{kr}^{(b)} \geq h} w_{krh}, \delta_2^{(b)} + \sum_{k:R_{kr}^{(b)} \geq h} (1 - w_{krh})\right). \tag{13}$$

4. By considering Gamma(1,1,) priors for $\delta_1^{(b)}$ and $\delta_2^{(b)}$, the following full conditionals are obtained.

$$\delta_1^{(b)}|- \sim \text{Gamma}\left(K(N_b - 1) + 1, 1 - \sum_{k=1}^{K}\sum_{h=1}^{N_b-1} log(1 - U_{kh}^{(b)})\right). \tag{14}$$

Similarly,

$$\delta_2^{(b)}|- \sim \text{Gamma}\left(T(N_b - 1) + 1, 1 - \sum_{r=1}^{5}\sum_{h=1}^{N_b-1} log(1 - W_{rh}^{(b)})\right). \tag{15}$$

We note that the truncation of the MSBP to finite N_b (and also for N_c) is conducted 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 $\mathbf{b}_{kr}$, the expected approximation error $= \left[1 - \frac{1}{(1+\delta_1^{(b)})(1+\delta_2^{(b)})}\right]^{N_b-1}$, which requires knowing $\delta_1^{(b)}$ and $\delta_2^{(b)}$. Following Dunson et al. (2008),

we specify independent Gamma(1,1) priors for $\delta_1^{(b)}$ and $\delta_2^{(b)}$, and we run the MCMC algorithm for about 15% of its total length and use the posterior means of $\delta_1^{(b)}$ and $\delta_2^{(b)}$ to determine if the expected approximation error is below 0.01. The same approach is used to choose N_c. A Similar approach is also proposed in Chatterjee et al. (2016). All computations are performed in R.

For estimating the regression coefficients, we run 52,000 MCMC iterations, and discard the first 2,000 iterations as "burn-in". Then, to reduce the inherent auto-correlation among the MCMC iterations, we thin the chains by saving every 10-th iteration. The regression coefficients are estimated by their respective sample means. We report an estimate as zero only if for more than 95% of the MCMC iterations the values are in the interval (-0.01,0.01). Since for some of the coefficients we use Lasso-type shrinkage priors, we end up with some coefficients which are estimated as zero.

5 Data Analysis

5.1 Prior Specification and Diagnostics

For λ^* in (10), we specify a Gamma (2,3.5) prior. For $\pi_{j'}$ in Section 3.2, we consider a Beta (1.5,2.5) prior, and for λ^2 and σ^2 we consider Inverse Gamma (1,2) and Inverse Gamma (2,3) priors, respectively. For σ_η^2, σ_b^2 and ρ, we consider Inverse Gamma (2,3.5), Inverse Gamma (1.5,2.5), and Uniform (-1,1) priors, respectively. We note that the specifications of the prior parameters are made following the literature, and these priors have minimal effect on the estimated parameter values. Table 3 shows the results from a sensitivity analysis, and we notice that choice for the hyper-parameters has little effect on the final estimates.

Since there are a good number of covariates, we check if there is any multicollinearity. At each wave, we compute the variance inflation factor (VIF) for each covariate. We notice that the computed VIFs are smaller than 7, indicating the absence of multicollinearity. For the HRS dataset, Biswas and Das (2019) reported similar results.

We use DIC for selecting the optimal order of the spline function in (7), and also for the optimal number of knots. We compare the fit using the conditional deviance information criteria (DIC) proposed in Celeux et al. (2006). This DIC is based on the conditional likelihood $l(\mathbf{Y}|\boldsymbol{\gamma})$ and is given by DIC$=-4\mathbf{E}[logl(\mathbf{Y}|\boldsymbol{\gamma})] + 2logl(\mathbf{Y}|\hat{\boldsymbol{\gamma}})$, where $\boldsymbol{\gamma}$ denotes the vector of random effects, and $\hat{\boldsymbol{\gamma}}$ is the corresponding estimate (posterior mean). The smallest DIC is achieved for a second order (g=2) spline with 3 knots (S=3). Also for truncating the MSBP prior, we get $N_b = N_c = 20$, and this choice specifies the expected approximation error less than 0.01, the commonly used threshold (Chatterjee et al. 2016). The convergence of the MCMC is assessed by computing the scale reduction factors (Das and Daniels 2014). For our analysis, these factors are all smaller than 1.1. Also Figure S.4 (in the web-appendix) shows the trace plots for some of the regression coefficients for which the LASSO-type shrinkage prior is used. These plots indicate a good convergence of the chains, and the scale reduction factors smaller than 1.1 also indicates the same.

5.2 Model Comparison

We consider the following competing models, and compare their performances for our HRS dataset. We select the model with the best predictive power.

- Model-1: This is our proposed model. We use a dynamic hurdle model with MSBP as a shrinkage prior as discussed in Section 3.

- Model-2: Here, we use a dynamic hurdle model for handling the excess zeros. However, for the coefficients in equation (7) we use a simple Dirichlet Process (DP) prior. Specifically, for $j = 1, 2, \ldots, J$; and for $r = 1, 2, \ldots, 5$; we define the vector $b_k = [b_{1k1}, b_{2k1}, \ldots, b_{Jk1}, b_{1k2}, b_{2k2}, \ldots, b_{Jk2}, \ldots, b_{1k5}, b_{2k5}, \ldots, b_{Jk5}]^T$. Then, we assume: $b_k | G_k^{(b)} \sim G_k^{(b)}$, $G_k^{(b)} \sim DP(\alpha_k, G_k^{(0)})$. This prior will cluster all the coefficients from all the time-varying predictors across all different SAH groups. For our computation, we consider $\alpha_k = 1$, and the base distribution $G_k^{(0)}$ is taken as a multivariate normal with mean vector=0, and the covariance matrix is taken as an identity matrix. A Similar DP prior is considered for the spline coefficients as well.

- Model-3: Here, we consider a dynamic hurdle model for handling the excess zeros. However, for the proportional odds model in equation (6), we take $\alpha_{jkr}(t_i) = \alpha_j(t_i)$, for all $j = 1, 2, \ldots, J$. Then we model the time-varying coefficients $\alpha_j(t_i)$ as follows: $\alpha_j(t_i) = b_{j0} + b_{j1}t_i + b_{j2}t_i^2 + ... + b_{jg}t_i^g + \sum_{s=1}^{S} c_{js}(t_i - \mathcal{T}_s)_+^g$, and consider an independent $N(0, \sigma^2)$ prior (with $\sigma^2 = 100$) for each regression coefficient.

- Model-4: Here, for handling the excess zeros, we consider a zero-inflated distribution given in equation (2). However, we assume that $Y_{irt} | Y_{irt} > 0 \sim Poisson(\lambda_{irt})$ (i.e. G is a Poisson distribution). Further, we consider the following linear model: $log(\lambda_{irt}) = x_{it}^T \theta_r$. This is the traditional hurdle model. And then for the coefficients θ_{jr}, $j = 1, 2, \ldots, J + J'$; we consider the following MSBP prior:

$$\begin{aligned} \theta_{jr} &\sim G_{jr} = \sum_{h=1}^{N} \pi_{jrh} \delta_{\xi_{jh}}; \\ \xi_{jh} &\sim N(0, \sigma^2). \end{aligned} \tag{16}$$

  We define the stick-breaking weights π_{jrh} as:

$$\begin{aligned} &\pi_{jrh} = V_{jrh} \prod_{s'<h} (1 - V_{jrs'}); \;\; V_{jrh} = U_{jh} W_{rh}, \\ &U_{jh} \overset{\text{iid}}{\sim} \text{Beta}(1, \delta_1); \;\; W_{rh} \overset{\text{iid}}{\sim} \text{Beta}(1, \delta_2). \end{aligned} \tag{17}$$

.

- Model-5: Here for handling the excess zeros, we consider the traditional hurdle model similar to Model 4. However, for the coefficients θ_{jr}, we consider a Dirichlet Process prior as follows: $\theta_{jr}|G_{jr} \sim G_{jr}$, $G_{jr} \sim DP(\alpha_0, G_0)$. We take $\alpha_0 = 1$, and the base distribution G_0 is taken as $N(0, 10)$.

- Model-6: This is the simplest model we consider. For handling the excess zeros we assume that $Y_{irt}|Y_{irt} > 0 \sim Poisson(\lambda_{it})$. Further, we consider the following linear model: $log(\lambda_{it}) = x_{it}^T\theta$. For each θ_j we consider a $N(0, 100)$ prior distribution.

Model selection is quite challenging in our case because the competing models are non-nested, and have different structures as well as the difficulty in integrating out the latent random effects to achieve the marginal model selection criteria. To tackle the above-mentioned problems we compute $P(\mathbf{Y}_i|\mathbf{Y}_{-i})$, which is the posterior predictive density of $\mathbf{Y}_i$ (the observed responses for the i-th individual) conditional on the observed dataset with a single data point (the one for the i-th individual) deleted. This value is known as the Conditional Predictive Ordinate (Gelfand et al. 1992, Das et al. 2021).

For the i-th individual, the CPO statistic is defined as:

$$\begin{aligned} \text{CPO}_i &= P(\mathbf{Y}_i|\mathbf{Y}_{-i}) \\ &= E_\Theta[P(\mathbf{Y}_i|\mathbf{Y}_{-i}, \Theta)], \end{aligned}$$

where $-i$ denotes the exclusion of the data for individual i, and Θ stands for the set of all model parameters. The expectation above is taken with respect to the posterior distribution of Θ given the cross-validated data $\mathbf{Y}_{-i}$. For individual i, the CPO_i can be obtained from the MCMC iterations by computing the following weighted average:

$$\widehat{\text{CPO}}_i = \left(\frac{1}{M} \sum_{m=1}^{M} \frac{1}{f(\mathbf{Y}_i|\Theta^{(m)})} \right)^{-1},$$

where M is the number of MCMC iterations. Here, $\Theta^{(m)}$ denotes the parameter samples at the mth iteration. A higher CPO value indicates a better fit. A useful summary statistic of the CPO_i is the logarithm of the pseudo-marginal likelihood (LPML), defined as $\text{LPML} = \sum_{i=1}^{N} \log(\widehat{CPO}_i)$. The Model with a higher LPML value has the better predictive power (Das et al. 2021).

We fit the above six competing models for the HRS dataset, and compute the LPML values. Table 4 summarizes the results, and we can see that the proposed model (Model I) provides the highest LPML. This illustrates that our proposed model has the best predictive power among the other competing models. Also the simplest model (Model 6) results in the lowest LPML value, indicating the worst fit. In a simulation study (given in the web-appendix), we show that an MSBP prior also performs better (higher LPML value) than the group-specific models, and also than a common model for all the groups (Table S.5). In Table 4, we show the computational time for

all the six competing models. We notice that the simpler models (with an ordinary hurdle model, and a simple DP or no shrinkage prior) are computationally less expensive, and a dynamic hurdle component makes a model computationally expensive. However, we think it is still worthwhile since it provides some interesting results.

5.3 Inference on the Covariates With Time-Invariant Effects

In Table 5, we show the effects of all 19 covariates on the probability of non-zero hospital visits. The estimates of the regression coefficients and the 95% posterior credible intervals (based on the MCMC) are provided in this table. In our analysis, we report a covariate as significant if the posterior credible interval (95%) for the corresponding regression coefficient does not contain zero. Among the chronic health conditions, cancer and strokes are found to have significant effects on the probability of a non-zero hospital visit, which accords with intuition since such serious conditions are associated with frequent hospitalization. Health insurance related to employment and other (private) health insurance are both found to have significant effects, which may reflect better quality healthcare associated with such insurances relative to government health insurance typically provided to low income groups and older individuals. Among the financial covariates total financial assets are found to be significant indicating that the financially affluent individuals are more likely to be hospitalized.

Next, in order to highlight the flexibility of our approach, we summarize the results for the covariates with time-invariant effects from equation (6) for different SAH categories and for different values of k. In Table 6, we present the corresponding coefficient estimates, and 95% credible intervals (based on the MCMC iterations) for the people belonging to the "good" SAH group (r=3) with 1 or less hospital admissions (k=1). With respect to the chronic health conditions, the findings accord with intuition, the three particularly severe conditions, namely, cancer, heart problems and strokes are found to be significant.

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=4. Note that here there are some additional important chronic diseases namely, blood pressure, lung problems, and arthritis. Additionally, smoking and alcohol consumption play significant roles on hospital admission for this group. Interestingly, government insurance is found to be significant here, indicating that medicaid and/or medicare are particularly important for the "poor" SAH group with a moderate to higher number of hospital admissions. Coefficients for the covariates gender, education and psychological problems are mostly (more than 95% of the times) in between (-0.01,0.01), and hence we report those estimates as zero. Table 6 also summarizes the results for the "excellent" SAH group with k=1. In addition to the chronic health conditions of cancer, heart problems and strokes, smoking is also significant here. Interestingly, education level is found to be an important predictor for this

case.

5.4 Inference on the Covariates with Time-Varying Effects

We now turn to the plots (Figures 2-5) for the time varying coefficients for these four covariates, namely: BMI, total financial assets, total financial debt and total household income. Focusing initially on BMI, it is apparent from Figure 2 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 SAH 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 the five 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 sizes 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 waves 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 onwards 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 total household income, the observed patterns tend to follow those associated with financial assets, with an upwards trend generally observed for all 5 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.

5.5 Grouping Behaviors

We want to investigate the "grouping" behavior of the MSBP priors for the sets of parameters $\boldsymbol{b}_{kr}$, and $\boldsymbol{c}_{kr}$. We define $\boldsymbol{\psi}_{kr} = [\boldsymbol{b}_{kr}, \boldsymbol{c}_{kr}]$, and then compute the posterior probabilities, $Pr(\boldsymbol{\psi}_{kr} = \boldsymbol{\psi}_{kr'}|\text{Data})$. The size of each box in Figure 6 is proportional to the corresponding posterior matching probabilities. The boxes on the diagonal correspond to probability= 1. In Figure 6, we notice that for k=1, there are higher matching probabilities between the 'fair' and 'good' SAH groups, and between the 'very good' and 'excellent' groups. This reflects that when the number of hospital admissions is relatively low, then the effects of the covariates BMI, total financial assets, total debt, and total household income are similar for the SAH groups 'fair' and 'good', and for 'very good' and 'excellent'. However, for k=3, we notice that the effects of those variables are similar for the SAH groups 'poor' and 'fair', and for 'good', 'very good', and 'excellent' groups. This highlights the usefulness of MSBP priors which can automatically cluster the SAH groups with respect to the effects of a group of covariates.

6 Discussion

The rate of hospital admissions at a particular time point differs between the individuals who have been hospitalized previously and the individuals who are yet to be hospitalized. This feature is observed in our HRS dataset, and hence we develop a dynamic hurdle proportional odds model for modeling the count of hospital admissions for older people. Unlike the traditional zero-inflated Poisson, or zero-inflated negative binomial model, the proportional odds model captures the natural ordering related to the count of hospital admissions. Additionally, there is an inherent 'group' structure among the individuals with respect to SAH status and the count of hospital admissions. It is not possible to handle such groups 'manually', and hence we use MSBP priors which can automatically handle such inherent groups, and also cluster the model parameters accordingly. Our analysis highlights how the effects of different predictors change over time for different baseline SAH status. The time-invariant effects of different chronic diseases and types of health insurance for different baseline SAH status are also reported. Additionally, through a simulation study (details are given in the web-appendix) we show that our proposed modelling provides estimates with lower bias, shorter credible intervals with acceptable coverage probabilities (Table S.6 in the web-appendix).

It is important to acknowledge, however, that our approach suffers from a few limitations. First, we assume no missingness in our data. In practice, 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). Second, we consider the baseline SAH status for grouping the individuals. However, the covariate SAH is indeed time-varying, and hence the composition of the groups should vary across the waves. A more flexible model (similar to the one proposed in Das et al. 2021) can potentially handle such dynamic group structure. These possible extensions remain as avenues for our future research.

Web-Appendix: Some additional plots for the HRS data analysis, and the details of the simulation studies (with results summarized in tables) are given in the web-appendix.

References

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

Albert JH, Chib S (1993) Bayesian analysis of binary and polychotomous response data. Journal of the American statistical Association 88: 669–679.

Ando A, Modigliani F (1963) The" life cycle" hypothesis of saving: Aggregate implications and

tests. The American Economic Review 53: 55–84.

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.

Baetschmann G, Winkelman R (2016) A Dynamic Hurdle Model for Zero-Inflated Count Data. Communications in Statistics-Theory and Methods 46: 7174–7187.

Banerjee R, Ziegenfus JY, Shah ND (2010) Impact of discontinuity in health insurance on resource utilization. BMC Health Services Research 10: 1–10.

Biswas J, Das K (2019) A Bayesian approach of analysing semi-continuous longitudinal data with monotone missingness. Statistical Modelling 20: 148–170.

Biswas J, Ghosh P, Das K (2020) A semi-parametric quantile regression approach to zero-inflated and incomplete longitudinal outcomes. Advances in Statistical Analysis 104: 261–283.

Biswas J, Das K (2021) A Bayesian quantile regression approach to multivariate semi-continuous longitudinal data. Computational Statistics 36: 241–260.

Brant R (1990) Assessing proportionality in the proportional odds model for ordinal logistic regression. Biometrics 46: 1171–1178.

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

Carroll KJ (2003) On the use and utility of the Weibull model in the analysis of survival data. Controlled clinical trials 24: 682–701.

Celeux G, Forbes F, Robert CP, Titterington DM (2006) Deviance information criteria for missing data models. Bayesian analysis 1: 651–673.

Chatterjee A, Venkateswaran P, Das K (2016) Simultaneous State Estimation for Clustered Based Wireless Sensor Networks. IEEE Transactions on Wireless Communications 15: 7985–7995.

Daniels MJ, Hogan JW (2008) Missing data in longitudinal studies: Strategies for Bayesian modeling and sensitivity analysis. CRC press.

Das K (2016) A Semiparametric Bayesian Approach for Joint Modeling of Longitudinal Trait and Event Time. Journal of Applied Statistics 43: 2850–2865.

Das K, Daniels MJ (2014) A semiparametric approach to simultaneous covariance estimation for bivariate sparse longitudinal data. Biometrics 70: 33–43.

Das K, Ghosh P, Daniels MJ (2021) Modeling Multiple Time-Varying Related Groups: A Dynamic Hierarchical Bayesian Approach with an Application to the Health and Retirement Study. Journal of the American Statistical Association: doi.org/10.1080/01621459.2021.1886105.

Deb P, Trivedi PK (1997) Demand for medical care by the elderly: a finite mixture approach. Journal of Applied Econometrics 12: 313–336.

Drentea P, Lavrakas PJ (2000) Over the limit: the association among health, race and debt. Social science and medicine 50: 517-529.

Duan N, Manning WG, Morris CN, Newhouse JP (1983) A comparison of alternative models for the demand for medical care. Journal of business & economic statistics 1: 115–126.

Dunson D, Xue Y, Carin L (2008) The matrix stick-breaking process: Flexible Bayes meta-analysis. Journal of American Statistical Association 103: 317–327.

Gelfand AE, Dey D, 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.

Hurd M, Kapteyn A (2003) Health, wealth, and the role of institutions. Journal of Human Resources 3: 386–415.

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

Ishwaran H, James LF (2002) Approximate Dirichlet Process Computing in Finite Normal Mixtures. Journal of Computational and Graphical Statistics 11: 508–532.

McCullagh P (1980) Regression models for ordinal data. Journal of the Royal Statistical Society: Series B (Methodological) 42: 109–127.

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

Mukherji A, Roychoudhury S, Ghosh P, Brown S (2016) Estimating health demand for an aging population: A flexible and robust Bayesian joint model. Journal of Applied Econometrics 31: 1140–1158.

Pelkowski JM, Berger MC (2004) The impact of health on employment, wages, and hours worked over the life cycle. Quarterly Review of Economics and Finance 44: 102–121.

Park T, Casella G (2008) The Bayesian Lasso. Journal of American Statistical Association 103: 681–686.

Ruppert D, Wand MP, Carroll RJ (2003) Semiparametric Regression. Cambridge University Press: New York (2003).

Sethuraman J (1994) A constructive definition of Dirichlet priors. Statistica Sinica 4: 639-650.

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

Westbury LD, Syddall HE, Simmonds SJ, Cooper C, 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.

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

Zhang Z (2016) Parametric regression model for survival data: Weibull regression model as an example. Annals of Translational Medicine 4: 484.

Table 1: Variables used in HRS data analysis. The role of each variable in our modeling is also given.

Name	Type	Role (in our model)
No. of hospital admissions	Count	Outcome
Self-assessed health	Categorical	Group
Total assets	Continuous	Covariate with time-varying effect
Total debt	Continuous	Covariate with time-varying effect
Total household income	Continuous	Covariate with time-varying effect
Body mass index	Continuous	Covariate with time-varying effect
Smoking habit	Binary	Covariate with time-invariant effect
Alcohol consumption	Binary	Covariate with time-invariant effect
Gender	Binary	Covariate with time-invariant effect
Education level	Binary	Covariate with time-invariant effect
Hypertension	Binary	Covariate with time-invariant effect
Diabetes	Binary	Covariate with time-invariant effect
Cancer	Binary	Covariate with time-invariant effect
Lung problem	Binary	Covariate with time-invariant effect
Heart problem	Binary	Covariate with time-invariant effect
Stroke	Binary	Covariate with time-invariant effect
Arthritis	Binary	Covariate with time-invariant effect
Psychological problem	Binary	Covariate with time-invariant effect
Employment health insurance	Binary	Covariate with time-invariant effect
Government health insurance	Binary	Covariate with time-invariant effect
Private health insurance	Binary	Covariate with time-invariant effect

Table 2: The distribution of the average number of hospital admissions depending on the time of the first hospital admission in the HRS data.

	Waves									
Wave for the first admission	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	0.60	0.43	0.32	0.28	0.26	0.26	0.38
3	0	0	-	0.57	0.45	0.28	0.18	0.23	0.20	0.21
4	0	0	0	-	0.37	0.39	0.24	0.23	0.21	0.20
5	0	0	0	0	-	0.36	0.42	0.26	0.31	0.27
6	0	0	0	0	0	-	0.38	0.38	0.41	0.31
7	0	0	0	0	0	0	-	0.37	0.58	0.34
8	0	0	0	0	0	0	0	-	0.41	0.42
9	0	0	0	0	0	0	0	0	-	0.49

Table 3: Results from the sensitivity analysis for the HRS dataset.

Coefficient	Prior	Estimate
λ^*	Gamma(2,3.5)	1.72
-	Gamma(1,4)	1.69
-	Gamma(1.5,2.5)	1.71
σ^2	Inverse Gamma (2,3)	1.54
-	Inverse Gamma (1,3.5)	1.51
-	Inverse Gamma (0.01,0.01)	1.48
σ_η^2	Inverse Gamma (2,3.5)	1.22
-	Inverse Gamma (1,2.5)	1.19
-	Inverse Gamma (0.01,0.01)	1.23
ρ	Uniform(-1-1)	0.38
-	Beta(1,2)	0.37
-	Beta(1.3,2.8)	0.40

Table 4: LPML values and Computational Times (CT) for different competing models in the HRS data analysis.

-	Model-1	Model-2	Model-3	Model-4	Model-5	Model-6
LPML	**-137.5**	-198.6	-258.4	-185.3	-234.8	-326.4
CT	20.3 min	18.5 min	16.3 min	5.3 min	4.8 min	2.6 min

Table 5: Estimates and 95% credible intervals (based on MCMC iterations) for the coefficients corresponding to all covariates for the probit model 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 6: Estimates and 95% credible intervals (based on MCMC iterations) for the regression coefficients corresponding to the covariates with time-invariant effects in the HRS data.

	SAH group=Good, k=1		SAH group=Poor, k=4		SAH group=Excellent, k=1	
Covariate	Estimate	95% CI	Estimate	95% CI	Estimate	95% CI
Blood pressure	0.064	(-1.33,0.87)	1.89*	(0.84,4.02)	0.53	(-2.15,1.73)
Diabetes	0.041	(-2.02,1.76)	0.67	(-1.54,2.51)	0.72	(-1.95,1.23)
Cancer	4.58*	(2.65,5.93)	3.88*	(1.55,6.01)	5.91*	(3.50,6.99)
Lung problem	1.16	(-0.29,2.05)	1.53*	(1.03,2.24)	1.13	(-2.05,2.86)
Heart problem	3.56*	(1.32,5.68)	2.66*	(1.13,5.61)	2.54*	(1.12,4.65)
Stroke	10.32*	(7.96,13.30)	7.19*	(3.36,9.14)	8.11*	(5.84,10.71)
Arthritis	0.29	(-1.31,2.04)	1.27*	(0.59,3.66)	0	-
Psychological problem	0.02	(-2.11,1.19)	0	-	0	-
Employment Insurance	0.18	(-2.74,1.55)	0.26	(-0.56,0.88)	0.31	(-0.94,0.75)
Government Insurance	3.63	(-1.49,6.26)	2.58*	(1.39,4.43)	0	-
Other Private Insurance	0.03	(-1.49,1.14)	0.16	(-1.29,0.94)	0.83	(-1.04,1.88)
Smoking	0.05	(-2.79,0.68)	1.19*	(0.78,2.18)	1.59*	(0.96,3.18)
Alcohol consumption	0.007	(-0.99,1.06)	2.34*	(1.39,4.11)	0.06	(-0.39,0.51)
Gender	0.58	(-3.61,2.47)	0	-	0	-
Education level	1.12	(-2.44,2.83)	0	-	1.27*	(0.94,2.86)

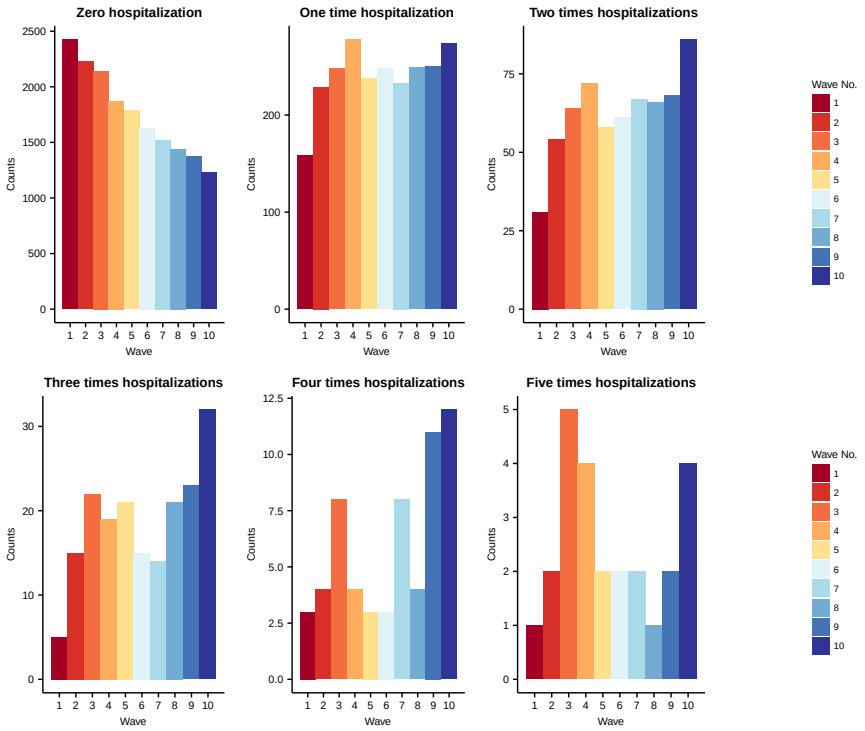


Figure 1: Bar charts comparing the counts of individuals with different numbers of hospitalizations across 10 waves.

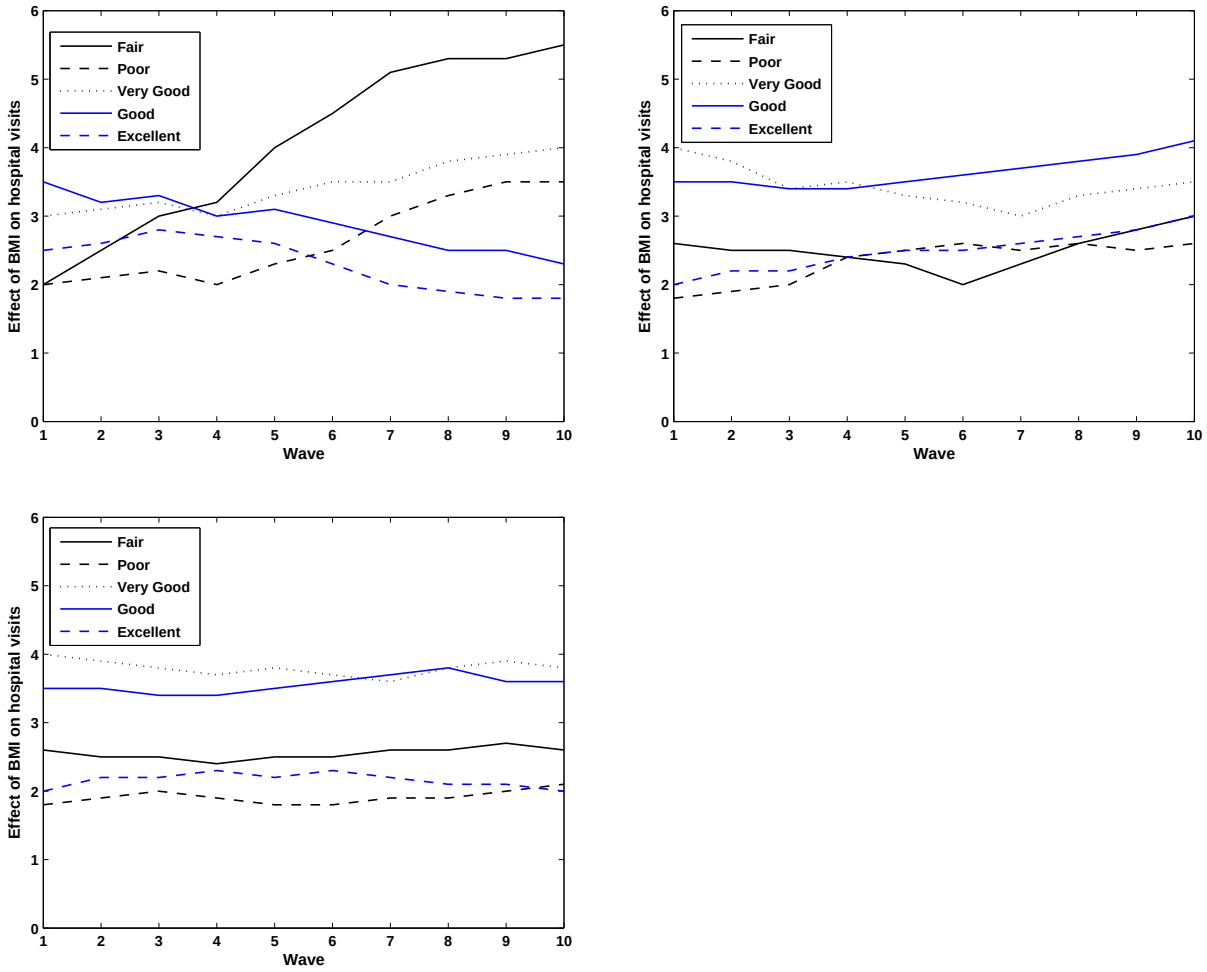


Figure 2: Time varying effect of BMI for 5 different groups for k=2, 4 and 6, respectively.

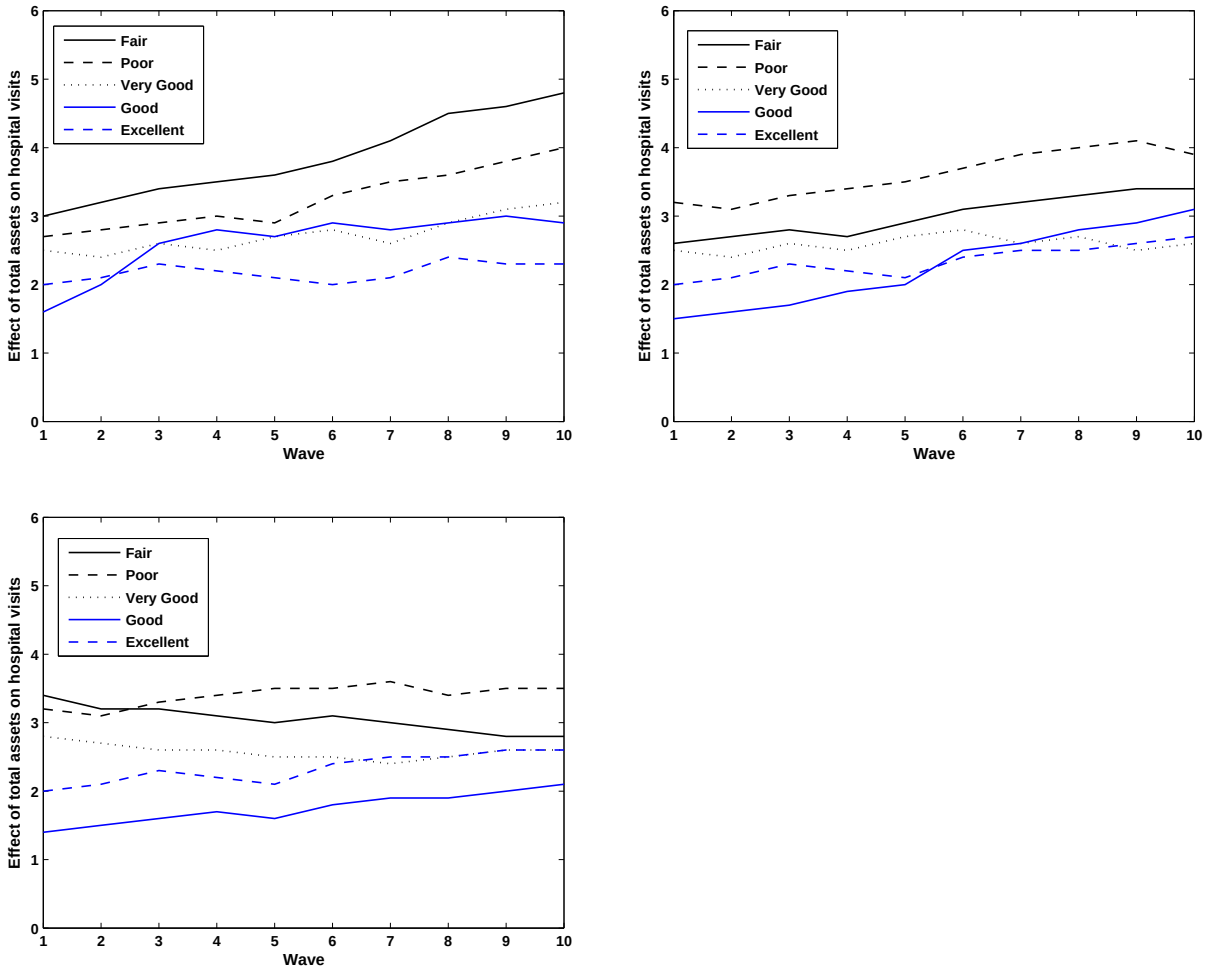


Figure 3: Time varying effect of the total assets for 5 different groups for k=2, 4 and 6, respectively.

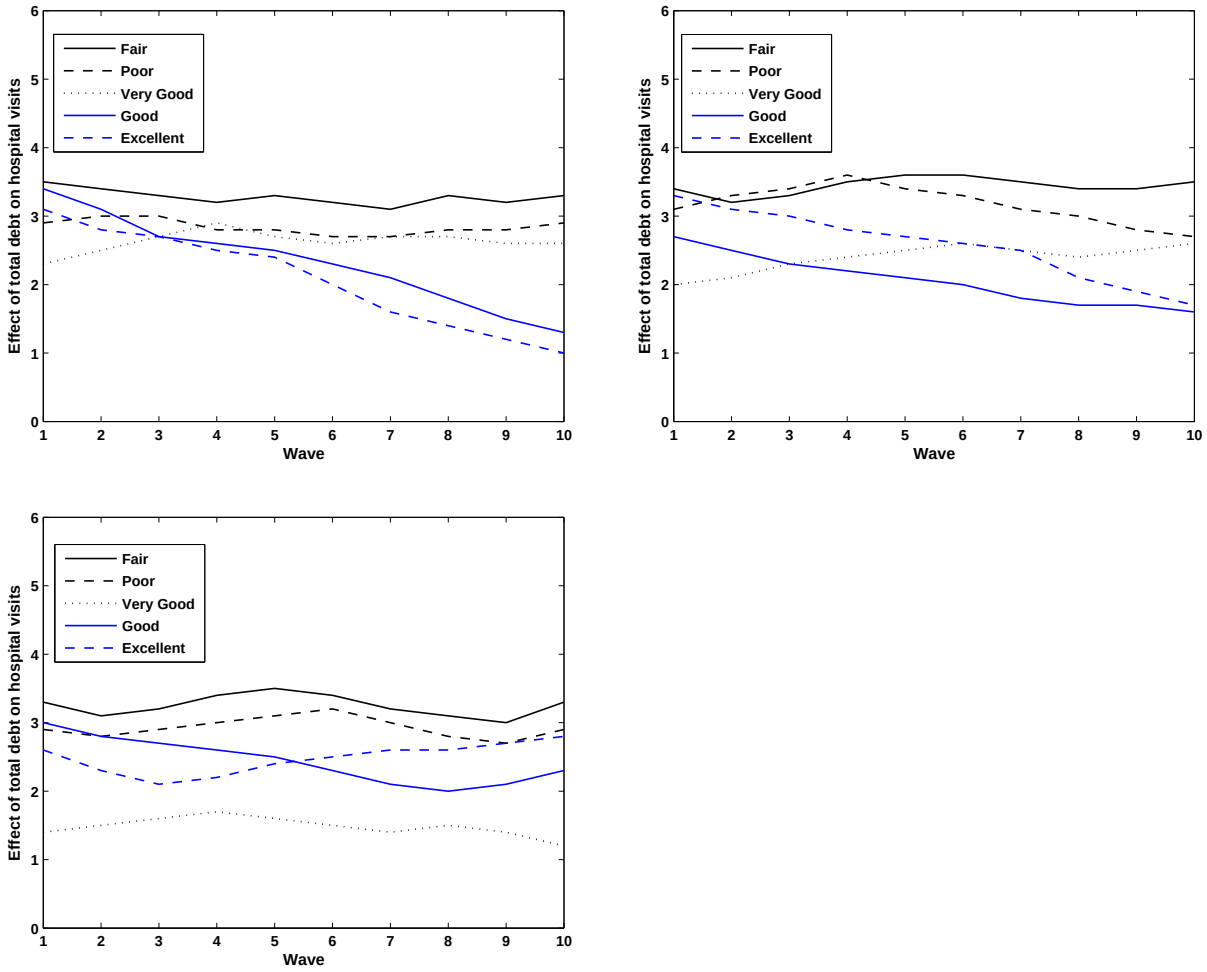


Figure 4: Time varying effect of the total debt for 5 different groups for k=2, 4 and 6, respectively.

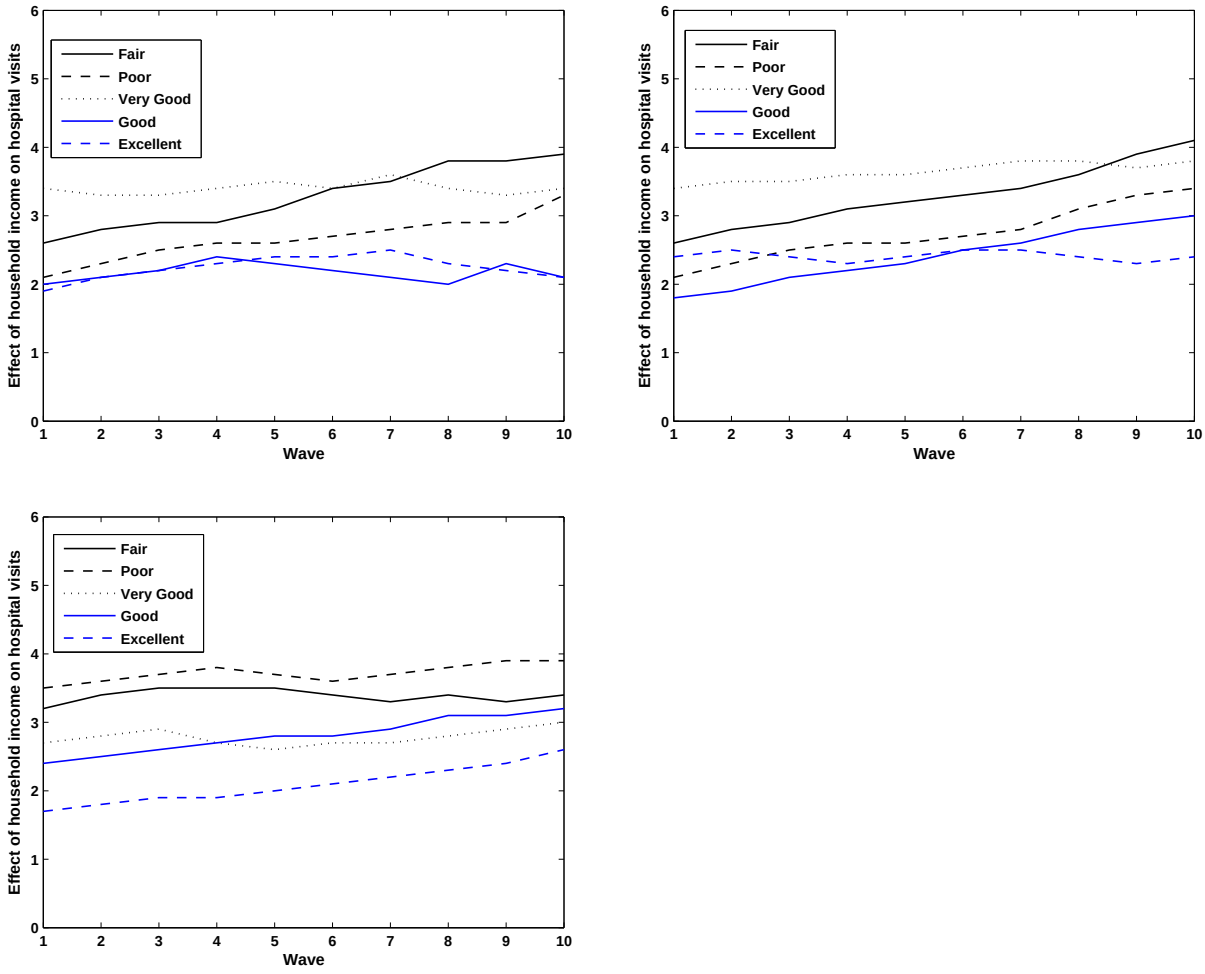


Figure 5: Time varying effect of the total household income for 5 different groups for k=2, 4 and 6, respectively.

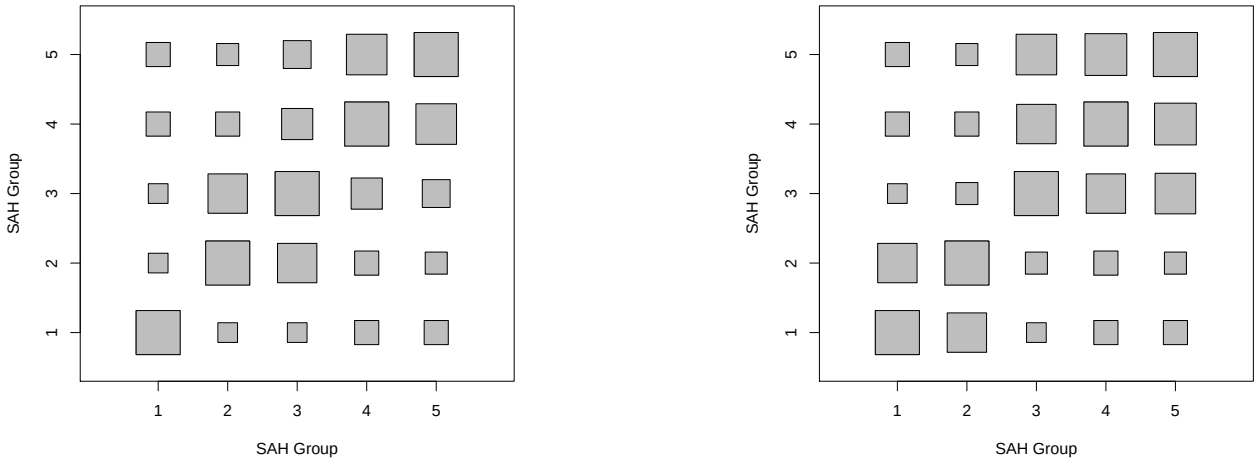


Figure 6: Matching probabilities across different SAH groups for when the number of hospital admissions (k) are 1 (left figure) and 3 (right figure), respectively.