JOURNAL OF APPLIED ECONOMETRICS
J. Appl. Econ. (2015)
Published online in Wiley Online Library
(wileyonlinelibrary.com) DOI: 10.1002/jae.2463

ESTIMATING HEALTH DEMAND FOR AN AGING POPULATION: A FLEXIBLE AND ROBUST BAYESIAN JOINT MODEL

ARNAB MUKHERJI,[a]* SATRAJIT ROYCHOUDHURY,[b] PULAK GHOSH[a] AND SARAH BROWN[c]

[a] *IIM Bangalore, India*
[b] *Novartis Pharmaceutical Company, New York, USA*
[c] *Department of Economics, University of Sheffield, UK*

SUMMARY

We analyse two frequently used measures of the demand for health—hospital visits and out-of-pocket health care expenditure—which have been analysed separately in the existing literature. Given that these two measures of health demand are highly likely to be closely correlated, we propose a framework to jointly model hospital visits and out-of-pocket medical expenditure, which allows for the presence of nonlinear effects of covariates using splines to capture the effects of aging on health demand. The findings from our empirical analysis of the US Health and Retirement Survey indicate that the demand for health varies with age. © 2015 The Authors. *Journal of Applied Econometrics* published by John Wiley & Sons Ltd.

Received 3 March 2013; Revised 3 December 2014

Supporting information may be found in the online version of this article.

1. INTRODUCTION

The world population is aging: according to a joint report by the US Department of State and the National Institute on Aging, almost 500 million people worldwide were 65 and older in 2006 (Dobriansky *et al*., 2007). By 2030, this number is expected to increase to 1 billion, or, one in every eight persons will be 65 years old or older. In the USA, life expectancy has increased from 49 years for Americans born in 1900 to 78 years for those born in 2006 (Arias, 2010). Rapid demographic change is expected to lead to an increase in health care spending by 25% by 2030 (Strunk *et al*., 2006; Dobriansky *et al*., 2007). While global aging represents a triumph of medical, social, and economic advances, it also poses tremendous challenges for health systems. It is well understood that aging will change the mix of diseases in favour of chronic conditions for inpatient care and this alone is likely to increase the demand for health care (Strunk *et al*., 2006; Hartman *et al*., 2008). With limited long-term benefits under health care schemes such as Medicare in the USA, such increases in demand will potentially lead to large out-of-pocket medical expenses for the elderly (Wei *et al.*, 2004; Hartman *et al*., 2008). Thus, with aging becoming a worldwide challenge, obtaining reliable estimates for the demand for health care has arguably never been more important than now (Dobriansky *et al*., 2007).

Health economics has traditionally focused on health care demand and Duan *et al.*'s (1982) seminal work on health demand explored different strategies to estimate medical expenditure to address data concerns specific to health cost data. Another metric that is also frequently used to measure health demand is the rate of hospital admissions (Atella and Deb, 2008). Despite the likely relationship

* Correspondence to: Arnab Mukherji, IIM Bangalore, Bannerghatta Road, Bangalore 560076, Karnataka, India. E-mail: arnab@iimb.ernet.in

between hospital visits and medical expenditure, these two measures of health care demand have typically been modelled separately in the existing literature. Furthermore, the probability of needing health increases with age, particularly with the onset of chronic conditions. Hence it is important to understand the effects of aging when modelling the demand for health. Thus managing health demand arguably requires an understanding of hospitalizations as well as medical expenditure in the context of an aging population. In this paper, we analyse the key factors affecting both hospital visits and medical expenditure by developing a novel joint modelling framework, which allows us to reliably study health demand and the correlation between these alternative measures of health demand.

Modelling hospitalizations and medical expenditure requires consideration of a number of complications specific to health data. First, both hospitalization and out-of-pocket expenditure at the individual level usually have a considerable amount of zero observations, which cannot be adequately described by a simple distribution such as a Poisson or a lognormal distribution. For example, Table I shows that 95% of our sample shows no hospital visits and 16% report zero out-of-pocket expenditure in wave 1 of the US Health and Retirement Survey (HRS). Spurious overdispersion occurs due to the presence of these zeros. Recently, Naya *et al*. (2008) compared model fits of a Poisson model and a zero-inflated Poisson (ZIP) model to zero-inflated data and found that a ZIP model gave estimates closer to the true values. Thus we need to modify parametric distributions to incorporate excess zeros in the distributions of the hospitalizations and out-of-pocket medical expenditure. Recent literature (such as Deb and Trivedi, 1997; Winkelmann, 2004; Atella and Deb, 2008) has developed zero-inflated distributions for modelling the count of hospital visits and medical expenditure; however, they are modelled independently. Second, hospital visits and medical expenditure are likely to be correlated with each other over time for the same individual. Accounting for this correlation may lead to a better understanding of health demand. Third, some important individual characteristics, such as age, may have complex nonlinear effects. In addition, the potential nonlinear effects of age could vary with other demographic characteristics, such as gender, resulting in an interaction effect that influences health demand in a nonlinear fashion. Fourth, both the count of hospital visits and medical expenditure are known to be skewed (Liu *et al*., 2010). Although some authors have argued in favour of log transformations to deal with skewness, this can be problematic. Re-transformation presents no problems when errors accord with linearity, normality and homoscedasticity assumptions (Jones, 2000). When any one of these does not hold, re-transformation bias arises on reverting back to the original scale. Since the log-transformed model results in geometric means rather than arithmetic means, log scale predictions will, in general, provide biased estimates of the impact of any explanatory variable on the arithmetic mean (Yu *et al*., 2014).

In this paper, we develop a joint framework for modelling counts of hospital visits and out-of-pocket medical expenditure in an integrated framework to accommodate the aforementioned complications as follows. We model the count of hospital visits made by an individual using a Poisson hurdle model (Mullahy, 1986) and we model out-of-pocket medical expenditure using a semi-continuous model (Liu *et al*., 2010). The Poisson hurdle model (semi-continuous model) consists of two components: a Bernoulli component that models the probability of hospitalization (any positive expense) and a truncated Poisson component (log-normally distributed component) that models the number of hospital visits (amount of money spent) among users. Together, these components accommodate both the high proportion of zeros and the right-skewness of the nonzero events. In addition, we explicitly account for interdependencies between these events by modelling the correlation between these two processes. While the literature on health care demand discusses 'multi-part' models, such as in the original work of (Duan *et al*., 1982) or the more recent work of Liu *et al*. (2008), these differ from our model in a number of ways. These models focus on a single outcome and the multi-part model allows for flexibility in model parameters across subgroups with different demands for health care. For example, Duan *et al*. (1982) focus on how the parameters vary by non-spenders, ambulatory spenders and inpatient spenders; more recently, Liu *et al*. (2008) are interested in the differences between non-spenders,

DOI: 10.1002/jae

outpatient spenders and inpatient spenders. Our model provides a richer specification of health demand that not only captures health care expenditure but also hospital visits within the same joint model with explicitly modelled random effects.

In addition, our sample is drawn from a predominantly aging population and the effects of age on hospital visits and medical expenditure are arguably poorly understood in the existing literature, yet, as argued above, are of utmost policy importance. We thus adopt a semi-parametric approach using spline models to flexibly capture the potentially nonlinear effects of age. This approach not only protects the model from the possible misspecifications of age effects but also allows us to explore whether this nonlinear effect varies across gender. For the distribution of the latent random effects terms of the joint model, a standard assumption is to use a parametric distribution, such as the multivariate normal distribution. The importance of such a choice has received much attention in the joint modelling literature. In particular, it has been shown that a restrictive parametric assumption for this distribution could influence the results (Tsonaka *et al.*, 2009; Naskar and Das, 2006). Thus, in order to protect the derived inferences against potential misspecification effects, we opt for a semi-parametric approach based on a Dirichlet process prior. A similar approach to modelling random effects, but with a single outcome and without splines, has also been proposed (Jochmann and Leon-Gonzalez, 2004). Finally, given our focus on a sample of aging individuals, we develop a discrete time survival model that explicitly allows for dropout and intermittent missing observations within our joint estimation framework that allows the shared random effects to influence the process generating the missing data. The rest of the paper is organized as follows: Section 2 presents the four-part model; Section 3 discusses the HRS data and the results of our empirical analysis; finally, Section 4 concludes.

2. A FOUR-PART ROBUST SEMI-PARAMETRIC JOINT MODEL

Our joint model consists of three components: a semi-parametric Poisson hurdle mixed-effects model for the number of hospitalizations, a semi-parametric semi-continuous model for out-of-pocket medical expenses, and a Dirichlet process for the joint distribution of the latent random effects from the Poisson hurdle and the semi-continuous models.

2.1. The Poisson Hurdle Model for the Count of Hospital Visits

The Poisson hurdle model is a two-component mixture model consisting of a point mass at zero followed by a truncated Poisson for the nonzero observations (Mullahy, 1986). The hurdle model, with independent and identically distributed responses, is given by

$$\begin{aligned} \Pr(Y_i = 0) &= 1 - p, \quad 0 \leq p \leq 1 \\ \Pr(Y_i = k) &= p \frac{\mu^k e^{-\mu}}{k!(1 - \mathrm{e}^{-\mu})}, \quad k = 1, \ldots, \infty, : 0 < \mu < \infty \end{aligned} \tag{1}$$

where Y_i denotes the response for individual $i = 1, \ldots, m$, and μ is the mean for an untruncated Poisson distribution. As the zeros and nonzero counts are modelled uniquely, the hurdle model accommodates both an excess number of zeros and a right-skewed distribution for the positive counts. With a large probability mass at zero (such as in our data; see Tables I and II) we have skewed data that may be spuriously suggestive of over-dispersion. Su *et al.* (2009) have shown, both theoretically as well as through simulations, that bias will be induced if skewness due to a large amount of zeros is not modelled by a zero-inflated distribution (see also Naya *et al.*, 2008). In comparison, a standard Poisson regression would have to compromise between the effect of having excess zeros that would tend to lower the Poisson mean, while large nonzero values would tend to increase it. Separating these

DOI: 10.1002/jae

competing effects are handled appropriately in a two-part framework.[1] The expected count under the Poisson hurdle model is given by $\mathrm{E}(Y) = p\mu/(1-\mathrm{e}^{-\mu})$.

In health services research, p is known as the *usage probability*, i.e. the probability of using services at least once. When $(1-p) > \mathrm{e}^{-\mu}$, the data are zero-inflated relative to an ordinary Poisson; when $(1-p) < e^{-\mu}$ there is zero deflation (i.e. fewer than expected zeros). In the extremes, $p = 0$ or 1. When $p = 1$, there are no zero counts and the model reduces to a truncated Poisson, and when $p = 0$ there are no users (i.e. all counts equal zero) and the model is degenerate at zero. Typically, one assumes that p lies strictly between 0 and 1, so that all individuals have a nonzero probability of usage and are, therefore, considered 'potential' users even if they do not actually use health services during the study period. A special case of equation (1) is the ZIP model, which is often used in this context (Lambert, 1992). The ZIP model consists of a degenerate distribution at zero mixed with an untruncated Poisson distribution:

$$P(Y_i = 0) = (1-p) + p\mathrm{e}^{-\mu}, \quad 0 < p < 1 \tag{2}$$

$$P(Y_i = k) = p\frac{\mu^k \mathrm{e}^{-\mu}}{k!}, \quad k = 1, \ldots, \infty, : 0 < \mu < \infty \tag{3}$$

Note that the ZIP model can be rewritten as a hurdle model with mixing probability $\theta = p(1-\mathrm{e}^{-\mu})$. Unlike the hurdle model, which accommodates zero deflation as well as zero inflation, the ZIP allows only for zero inflation and thus allows for greater flexibility (Neelon *et al.*, 2010). Let Y_{ij}^H be the count of the number of hospital stays reported by the ith individual in the jth wave, $i = 1, 2, \ldots, m$; $j = 1, 2, \ldots, n$, where m represents the number of individuals in the study, and n is the total number of waves over which the individual is surveyed. Depending on whether an individual is hospitalized or not, a large number of zeros is observed in Y_{ij}^H. Also, let X_{ijk} be the kth covariate for individual i at time j; such covariates include baseline and time-varying variables.

Each individual's total count of hospital visits is determined simultaneously by needing some health care (p_{ij}) as well as the level of care needed given that the person needs care λ_{ij}. Given that these are jointly determined, and that the determinants of either may or may not be relevant for the other, we consider simultaneous modelling of both λ_{ij} and p_{ij}. The hurdle model can be extended to accommodate covariates and random effects as follows:

$$\begin{aligned} p\left(y_{ij}^H|\phi_i\right) &= \left(1-p_{ij}^H\right)1\left[y_{ij}^H = 0\right] + p_{ij}^H \mathrm{Tpois}\left(y_{ij}^H; \mu_{ij}^H\right)1\left[y_{ij}^H > 0\right] \\ \mathrm{logit}\left(p_{ij}^H\right) &= \mathbf{X}_{ij1}^T\boldsymbol{\beta}_1^p + \mathbf{Z}_{ij1}^T\boldsymbol{b}_{i1} + f^p\left(W_{ij}\right) \\ \log\left(\mu_{ij}^H\right) &= \mathbf{X}_{ij2}^T\boldsymbol{\beta}_1^\lambda + \mathbf{Z}_{ij2}^T\boldsymbol{b}_{i2} + f^\lambda\left(W_{ij}\right) \end{aligned} \tag{4}$$

where $p_{ij}^H = \Pr(Y_{ij}^H > 0)$, $\mathrm{Tpois}(y_{ij}^H; \mu_{ij}^H)$ denotes a truncated Poisson distribution with parameter μ_{ij}^H. $\mathbf{X}_{ij1}, \mathbf{X}_{ij2}$ are the vectors of covariates corresponding to the fixed effects and $\mathbf{Z}_{ij1}, \mathbf{Z}_{ij2}$ are the vectors of covariates corresponding to the random effects. Note that the zero-state and the Poisson state do not need to have the same set of covariates. The b_{i1} and b_{i2} are the random individual effects on p_{ij} and λ_{ij}, respectively. We will discuss the distribution of the random individual effects later.

[1] A trinomial distribution is an alternative to the standard Poisson regression, particularly when there are limited positive values. While this is attractive, it still fails to allow for the two-part process and thus would attract similar concerns to the standard Poisson regression as in Naya *et al.* (2008) and Su *et al.* (2009). Consequently, we specify the more flexible two-part Poisson hurdle specification.

DOI: 10.1002/jae

In many situations, such as our application, the effect of some covariates, viz. W_{ij} on p_{ij}^H and μ_{ij}^H, may not be linear. Thus the effects of those covariates can be modelled by unspecified nonparametric functions $f^p(W_{ij})$ and $f^\lambda(W_{ij})$. These unknown smoothing functions reflect the nonlinear effects of the covariate. However, these functions only represent the population averages for a single population.

We now consider a modified model for multiple subgroups. Instead of fitting one nonparametric smoothing spline for the entire sample, we use multiple nonparametric smoothing splines for different subgroups within one model. We consider

$$\begin{aligned}\text{logit}\left(p_{ij}^H\right) = \mathbf{X}_{ij1}^T\boldsymbol{\beta}_1^p + \mathbf{Z}_{ij1}^T\boldsymbol{b}_{i1} + f_1^p(W_{ij})d_{ij1}^p + f_2^p(W_{ij})d_{ij2}^p \\ + \ldots + f_L^p(W_{ij})\left(1 - d_{ij1}^p - d_{ij2}^p - \ldots - d_{ij(L-1)}^p\right)\end{aligned} \tag{5}$$

$$\begin{aligned}\log\left(\mu_{ij}^H\right) = \mathbf{X}_{ij2}^T\boldsymbol{\beta}_1^\lambda + \mathbf{Z}_{ij2}^T\boldsymbol{b}_{i2} + f_1^\lambda(W_{ij})d_{ij1}^\lambda + f_2^\lambda(W_{ij})d_{ij2}^\lambda \\ + \ldots + f_L^\lambda(W_{ij})\left(1 - d_{ij1}^\lambda - d_{ij2}^\lambda - \ldots - d_{ij(L-1)}^\lambda\right)\end{aligned} \tag{6}$$

where d_{ijk}; $k = 1, 2, \ldots, L$ are indicator variables for multiple populations. With L populations, the first group is indicated by ($d_{ij1} = 1$, $d_{ij2} = 0, \ldots, d_{ij(L-1)} = 0$), the second group is indicated by ($d_{ij1} = 0$, $d_{ij2} = 1, \ldots, d_{ij(L-1)} = 0$) and the last group is indicated by ($d_{ij1} = 0$, $d_{ij2} = 0, \ldots, d_{ij(L-1)} = 0$). The $f_1, f_2, \ldots, f_L$ are their respective nonparametric smoothing splines. So far we have described a general structure for the model. For the analysis of our sample of aging individuals presented in Section 3, we have two groups—male ($d_{ij1} = 1$) and female ($d_{ij1} = 0$)—and we have $f(\text{age})$. The explicit expression for the model in the context of our data and the relevant covariates are discussed in detail in Section 3 below.[2]

2.2. Semi-Continuous Out-of-Pocket Medical Expenditure Model

In this section, a semi-continuous model for longitudinal data on out-of-pocket medical expenditure is introduced. Since in some years the individual may not have incurred any medical expenditure, this type of data has a mix of zeros and positive continuous observations. To formulate the model, let y_{ij}^M be the medical expenditure of individual i at year j. Let R_{ij} be a random variable denoting annual medical expenditure, where

$$R_{ij} = \begin{cases}0, & \text{if } y_{ij}^M = 0 \\ 1, & \text{if } y_{ij}^M > 0\end{cases} \tag{7}$$

with conditional probabilities

$$\Pr(R_{ij} = r_{ij}) = \begin{cases}1 - p_{ij}^M, & \text{if } r_{ij} = 0 \\ p_{ij}^M, & \text{if } r_{ij} = 1\end{cases}$$

For these semi-continuous data, we introduce an analogous semi-continuous model consisting of a degenerate distribution at zero and a positive continuous distribution, such as a lognormal (LN), for the nonzero values:

$$f\left(y_{ij}^M|\mathbf{p}_i^M\right) = \left(1 - p_{ij}^M\right)^{1-r_{ij}}\left\{p_{ij}^M \times \text{N}(\log(y_{ij}^M); \mu_{ij}^M, \sigma^2)\right\}^{r_{ij}}$$

[2] Further details on the spline formulation are available in the online Appendix (supporting information).

DOI: 10.1002/jae

$$\begin{aligned}\text{logit}(p_{ij}^{M}) = {} & \mathbf{X}_{ij}^{T}\boldsymbol{\beta}_{1}^{Mp} + \mathbf{Z}_{ij1}^{T}\boldsymbol{b}_{i3} \\ & + h_{1}^{p}(W_{ij})e_{ij1}^{p} + h_{2}^{p}(W_{ij})e_{ij2}^{p} + \dots \\ & + h_{L}^{p}(W_{ij})\left(1 - e_{ij1}^{p} - e_{ij2}^{p} - \dots - e_{ij(L-1)}^{p}\right)\end{aligned} \tag{8}$$

$$\log(Y_{ij}^{M} | Y_{ij}^{M} > 0) \sim N\left(\mu_{ij}^{M}, \sigma^{2}\right) \tag{9}$$

$$\begin{aligned}\mu_{ij}^{M} = {} & \mathbf{X}_{ij}^{T}\boldsymbol{\beta}_{1}^{M\lambda} + \mathbf{Z}_{ij2}^{T}\boldsymbol{b}_{i4} \\ & + h_{1}^{\lambda}(W_{ij})e_{ij1}^{\lambda} + h_{2}^{\lambda}(W_{ij})e_{ij2}^{\lambda} \dots \\ & + h_{L}^{\lambda}(W_{ij})\left(1 - e_{ij1}^{\lambda} - e_{ij2}^{\lambda} - \dots - e_{ij(L-1)}^{\lambda}\right)\end{aligned} \tag{10}$$

where r_{ij} is an indicator as defined above, μ_{ij}^{M} and σ^{2} are the mean and variance of $\log\left(y_{ij}^{M}\right)$, respectively. The interpretation of e_{ijk} is the same as d_{ijk} in the ZIP model and the nonparametric spline function $h(.)$ is also defined in a similar fashion. The model given by equations (9) and (10) is a semiparametric counterpart of the correlated two-part model proposed by Olsen and Schafer (2001); a gamma or log-skew-normal distribution may also be used to model the nonzero values.

2.3. Modelling Missing Observations

Recently there has been a great deal of interest in modelling longitudinal data subject to dropout. Biased inferences result when dropout is related to either the unobserved value or the underlying response process and this is not modelled properly. These dropout data mechanisms have been referred to as informative dropout mechanisms (Little and Rubin, 1987). When the probability of dropping out is related to the underlying response process for an individual, the dropout data mechanism is said to be informative (Wu and Carroll, 1988). This type of dropout has been modelled by introducing random effects that are shared between the model for the repeated measures and the model for the missing data mechanism. Various authors have proposed shared random-effect models for longitudinal data subject to informative dropout and we also take the same route. However, our scenario is more complicated as we observe two distinct dropout patterns in the data: one type of individual can be regarded as 'rotational dropout', i.e. they drop out intermittently and return to the survey in later waves. This is opposed to individuals who, once they drop out, never return to the survey, who we classify as permanent dropout. To account for these two different types of dropout processes, we develop a discrete-time competing risk model, which is essentially a multinomial logistic model. The proposed discrete-time logit model can be extended by incorporating random effects varying across individuals. This effect takes into account unobserved heterogeneity and dependence between the different dropout processes.

Let, V_{ij} be the missing value indicator, which takes the value 0 if individual i is observed, 1 if individual i exhibits intermittent missing values and 2 if individual i permanently dropped out. Thus V_{ij} is a multinomial response vector with $v\ (=3)$ categories. The probability model for the multinomial logit V_{ij} can be written as follows:

$$p_{vij} = \text{prob}\left(V_{ij} = v\right) = \begin{cases} \frac{1}{1+\sum_{v=1}^{2}\exp\left(\pi_{ijv}^{\omega}\right)} & v = 0 \\ \\ \frac{\exp\left(\pi_{ijv}^{\omega}\right)}{1+\sum_{v=1}^{2}\exp\left(\pi_{ijv}^{\omega}\right)} & v = 1,\ 2 \end{cases}$$

DOI: 10.1002/jae

The dropout model is assumed to be conditionally dependent on past and current values of the dependent variables. The regression for the intermittent missing value is given by

$$\pi_{ij1}^{\omega} = \zeta_1^P y_{i,j-1}^H + \zeta_2^P y_{i,j}^H + u_i \tag{11}$$

and the same for the permanent dropout is given by

$$\pi_{ij2}^{\omega} = \chi_1^P y_{i,j-1}^H + \chi_2^P y_{i,j}^H + \lambda_2 u_i$$

Here, u_i is the random effect corresponding to the intermittent dropout and follows the normal distribution. The λ_2 is the connection of the random effects between intermittent and failure time. The π_{ijv}^{ω} is the hazard of an event of type v occurring at time t for an individual i (Elashoff *et al.*, 2008).

In the traditional approach to the problem of competing risks, where the occurrence of an event removes the individual from the risk of other events, each event is analysed separately, while all other events are treated as censored. In our formulation, however, the two risks are correlated through the sharing of the random effects u_i. This is important, as an individual who is intermittently dropping out may be more prone to permanently dropout and thus the two risks may be correlated. We have taken the censored cases as the reference category.[3]

2.4. The Latent Random-Effects Distribution: Dirichlet Process Priors

Without loss of generality, we assume that all $\mathbf{b}_{ik}$ in equations (5), (6), (9) and (10) are $r \times 1$ unobserved vectors. Let $\mathbf{b}_i = \left(\mathbf{b}_{i1}^\top, \mathbf{b}_{i2}^\top, \mathbf{b}_{i3}^\top, \mathbf{b}_{i4}^\top\right)^\top$, $i = 1, \ldots, m$, is a $4r \times 1$ vector representing the random effects for the ith individual. To allow for the correlation structure between repeated observations for the same individual taken over different years and also to account for uncertainty in the probability distributions of the random effects, usually one takes a multivariate normal distribution. Recent work shows the importance of explicitly modelling the correlation structure (Su *et al.*, 2009).

However, in an aging population the subjects' responses may result in increased heterogeneity in the population. In addition, the endpoints are skewed and thus a parametric normal distribution may be restrictive for the latent random effects. Thus, instead of a normal distribution, we employ a Dirichlet process (DP) prior based on a stick-breaking scheme (Ferguson, 1973; Blackwell and MacQueen, 1973; Antoniak, 1974; Sethuraman, 1994) that makes fewer assumptions about the distribution function.

To proceed, we assume latent variables $\mathbf{b}_i$ are drawn from an arbitrary distribution G, where G has a DP prior, denoted by $\mathbf{b}_i \sim \mathrm{DP}(a G_0)$, $G_0 \sim \mathtt{N}_{4r}(\mathbf{0}, \boldsymbol{\Sigma})$ and a is an unknown concentration parameter. Usually a uniform prior is assumed for a. Thus the DP prior is essentially a distribution on the space of distributions and parametrized by a known base distribution G_0 and by a positive concentration parameter a that represents variability around G_0. The G_0 can be viewed as the 'mean' distribution in the space of distributions covered and a is a measure of the 'variance' of realizations of G around G_0. Hjort *et al.* (2010) provide a thorough review of the DP literature. Formally our model for $\mathbf{b}_i$ can be hierarchically expressed as

$$\begin{aligned} \mathbf{b}_i | G &\overset{\text{i.i.d.}}{\sim} G,\ i = 1, \ldots, m, \\ G | a, G_0 &\sim \mathtt{DP}(a G_0), \quad \text{with } G_0 = \mathtt{N}_{4r}(\mathbf{0}, \boldsymbol{\Sigma}) \end{aligned} \tag{12}$$

[3] In our application presented in Section 3 below, in the case of both intermittent and permanent missing values, past and current values of the dependent variables are found to be statistically significant. For brevity, these results are not reported in the paper but are available on request.

DOI: 10.1002/jae

Sethuraman (1994) provided an explicit characterization of G in terms of a stick-breaking construction, where G is represented as an infinite mixture of discrete atoms m_h with probabilities w_h $\left(\sum_{h=1}^{\infty} w_h = 1\right)$. In our context, the m_h are drawn i.i.d. from $G_0 \sim \mathrm{N}_{4r}(\mathbf{0}, \boldsymbol{\Sigma})$. For w_h, imagine a probability stick of unit length and break off a portion $w_1 = \pi_1$, where π_1 is drawn from a beta distribution, Beta$(1, a)$. The length of the remaining stick is $(1 - \pi_1)$. Let π_2 be another independent draw from the same beta distribution, representing the portion of the remaining probability stick that is broken off. Thus $w_2 = \pi_2(1 - \pi_1)$ denotes the probability associated with the second independent draw m_2 from G_0. Continuing, we obtain

$$G = \sum_{h=1}^{\infty} w_h \delta_{m_h}; \text{ with } w_h = \pi_h \prod_{l=1}^{h-1}(1 - \pi_l), \text{ for } h = 1, 2, \ldots, \infty$$

$$\text{where } \pi_h | \nu \sim \text{Beta}(1, a), \text{ and } m_h \overset{\text{i.i.d.}}{\sim} G_0$$

Here, δ_{m_h} denotes a discrete distribution with all its probability mass at m_h. For all values of $a (a \approx 1)$ the first four or five m_h account for 99% of the distribution G, while for a large value of $a (a \approx 10)$, 99% of the distribution of G is accounted for by the first 50 m_h's (Hjort *et al.*, 2010). Because of this fact G can be reduced to a truncated DP by truncating at a large number R.[4]

3. EMPIRICAL ANALYSIS

3.1. Data

In order to explore the relationship between out-of-pocket medical expenditure and hospitalizations, we use data from the University of Michigan's Health and Retirement Study (HRS). The HRS is a longitudinal survey of Americans over the age of 50, with a follow-up frequency of 2 years, and is designed to provide multidisciplinary data to understand the challenges of aging. In this paper, we use data for eight waves from the 1931–41 cohort—the HRS cohort. Baseline observations for the HRS cohort begin in 1992 when individuals were between 52 and 62 years of age and were near retirement. The data we use are maintained by RAND's Center for the Study of Aging and has been comprehensively cleaned and documented (St Clair *et al.*, 2009). Our estimation is based on sampling from 1929 individuals, 1227 of whom we observe in all eight waves. All individuals are observed in wave 1 and at least two other waves.

For our outcome measures, we use the number of hospital visits made since the previous interview, which is based on responses to the following question: How many different times were you a patient in a hospital overnight in the last 12 months? On the other hand, the total out-of-pocket medical expenses variable (OOPMD) covers total medical costs for all medical services since the previous interview and excludes all costs that were reimbursed or paid through insurance. It covers four groups of services, namely: hospital/nursing; doctor/outpatient/dental; prescription drugs; and home health care/special services. Our focus on out-of-pocket medical expenditure, namely medical expenditure minus that covered by insurance, follows numerous contributions to the existing health economics literature (such as Palumbo, 1999; Finkelstein and McKnight, 2008; and, more recently, Goda *et al.*, 2013; Goldman and Maestas, 2013) and reflects the argument that it is a measure of the actual financial burden of medical expenditure incurred by individuals, as well as a significant source of financial risk. It is important to acknowledge, however, that type of insurance status has an

[4] Details on Bayesian inference and computation are presented in the online Appendix.

DOI: 10.1002/jae

Table I. Distribution of outcomes

	Count of hospital visits				Out-of-pocket medical expense			
	%	Non-zeros			%	Non-zeros		
HRS wave	Zeros	Mean	Min.	Max.	Zeros	Mean	Min.	Max.
1	94.67	1.12	1	2	16.00	1075	8	15,928
2	94.20	1.62	1	6	18.12	1437	9	34,687
3	92.86	1.30	1	2	15.00	1201	15	13,400
4	90.48	1.08	1	2	15.08	1269	3	21,600
5	81.97	1.45	1	6	12.30	1463	25	18,080
6	80.87	1.27	1	2	11.30	2310	5	29,780
7	88.70	1.15	1	2	10.43	2414	100	52,400
8	81.25	1.24	1	3	9.82	1971	15	15,100

Table II. Summary statistics of response and predictors

Variables	Mean	SD	Min.	Max.
Time invariant				
Education: GED or higher? (gedplus_i)	0.88	0.33	0.00	1.00
Is female? (female_i)	0.47	0.50	0.00	1.00
95–100% of pre 65 years covered (covered)	0.86	0.35	0.00	1.00
10–60% of pre-65 years covered (mostlycovered)	0.06	0.24	0.00	1.00
60–95% of pre-65 years covered (partlycovered)	0.05	0.23	0.00	1.00
Time varying				
Count of hospital visits (rhsptim)	0.15	0.48	0.00	6.00
OOPMD (roopmd)	1391.30	3158.36	0.00	52400.00
Age (Age)	60.30	5.37	50.00	73.00
Has no difficulty in dressing (rdress)	0.84	0.37	0.00	1.00
Self-reported expectation of living 10+ years (rliv)	70.13	24.51	0.00	100.00
Self-reported health: Δ in current wave (rshlth)	−0.95	0.88	−2.00	2.00
Self-reported health: Δ in previous wave (rshlthc)	0.06	0.71	−2.00	4.00
Does health limit work? (rwork)	0.05	0.22	0.00	1.00
Respondents BMI (rbmi)	26.87	4.39	15.30	51.50
Respondent has at least one of chronic conditions (rchronic)	0.50	0.50	0.00	1.00

important influence on out-of-pocket expenditure and that the extent to which we are able to control for such factors is somewhat limited.[5]

Table I presents summary statistics for both outcome measures. The count of hospital stays exhibits increasing frequency of missing observations in later waves; additionally, there is a high but declining fraction of the sample in each wave with zero hospital visits. This change in health demand over time is also captured through narrower ranges of outcomes in earlier waves than in later waves, emphasizing the important effects of aging. Thus, while in wave 1 approximately 95% of the sample did not visit a hospital, by the last wave this proportion had declined to 81%. This high frequency of zeros supports the use of a Poisson hurdle model for hospital visits. Similarly, OOPMD also shows significant zero inflation, suggesting that treating it as a continuous variable would be problematic. As people age, the frequency of hospital visits rises, and so does OOPMD. We see this in Table I: as the frequency of zeros declines, the average OOPMD rises from \$1075 to \$1971.

Descriptive statistics for the baseline and time-varying covariates are presented in Table II. Information on many factors, such as gender, education, insurance coverage and functional independence,

[5] Total medical expenditure is an alternative cost measure available in the HRS; however, it is only available in waves 1–6 of the HRS. This would limit the length of the panel and would arguably lead to an overestimate of the burden of medical expenditure incurred at the individual level.

DOI: 10.1002/jae

Table III. Between and within variation in time-varying variables

Time-varying continuous variables		Mean	SD	Min.	Max.
OOPMD (roopmd)	Overall	1391.3	3158.36	0	52400
	Between		803.33		
	Within		2855.54		
Age (Age)	Overall	60.3	5.37	50	73
	Between		2.7		
	Within		0.22		
Expectation of living 10 or more years (rliv)	Overall	70.13	24.51	0	100
	Between		18.68		
	Within		14.6		
Respondents BMI (rbmi)	Overall	26.87	4.39	15.3	51.5
	Between		4.4		
	Within		1.24		

is available in the HRS. We control for being female as well as for whether the individual's education is at the level of the General Education Diploma (GED) or higher. We also control for the age of the respondent; in almost any aging study, the age of the respondent is an important predictor of health outcomes (Strunk *et al.*, 2006; Wei *et al.*, 2004); hence it is believed that the age of the respondent is predictive of his/her health care demand and out-of-pocket medical expenditure. For each continuous variable Table III reports the cross-sectional and time variation that we exploit.[6] While the extent of the variation changes from variable to variable, we again note the substantial portion of the variation that is due to time variation within individuals over time rather than across individuals.

In the context of the USA, it is important to control for health insurance. In particular, at age 65, US citizens become eligible for Medicare, a public insurance program. Hence, post age 65, insurance coverage in our sample is universal. Thus there is a discontinuity in access to health insurance at age 65 such that health care usage is expected to rise, especially if individuals were previously uninsured (see, for example, Card *et al.*, 2008, 2009). We thus include a set of controls for insurance histories of individuals prior to turning 65 years of age. It is important to acknowledge that there is heterogeneity in health insurance status post age 65, which our approach does not allow for. [7] For example, some individuals purchase Medicare supplements, some individuals may hold private insurance [8] and some may be eligible for Medicare and Medicaid, a health care programme for those with low income. Given data limitations, since we are unable to control for such considerations, our approach essentially assumes that insurance status post age 65 is homogeneous. Thus it is important to acknowledge that our estimated age effects may be contaminated by closely related omitted variables (such as type of insurance).

A key aspect of aging is a loss of functional abilities (muscular strength, ventilatory capacity, incontinence or cardiovascular output); however, the rate of this decay varies with lifestyle and environmental factors (Wei *et al.*, 2004). The variables used to capture functional independence are whether the individual reports that they experience no difficulty in dressing and whether their health limits their ability to work. We also control for the respondent's BMI. Additionally, data on each individual's self-reported health status in the current and past wave, distinguishing between excellent, very good, good, fair and poor health, is used as it is known to be predictive of health status (McGee *et al.*, 1999). Specifically, we include variables capturing the change in self-assessed health between

[6] The online Appendix shows time plots of categorical variables that we use in our analysis.

[7] We are very grateful to a referee for highlighting this important point.

[8] Interestingly, in our sample, 1108 individuals reported having non-public health insurance after age 65, which in the vast majority of cases was related to their employment. 99% of these individuals reported having previously held non-public health insurance prior to age 65. For these individuals, we assume that the nature of their insurance cover is the same in the pre and post age 65 periods.

DOI: 10.1002/jae

the current and previous waves, where positive (negative) values indicate a deterioration (improvement) in self-reported health between waves.[9] The HRS also includes information on each respondent's expectation of being alive for the next 10 years or more on a 0–100 scale; this is known to predict mortality (Hurd and McGarry, 2002). We include this variable to explore the influence of expectations on health-seeking behaviour. To capture the long-term state of the respondent's health, we include a dummy variable for chronic conditions which equals one if the individual has ever had any of the following chronic conditions: (i) high blood pressure or hypertension; (ii) diabetes or high blood sugar; (iii) cancer or a malignant tumour of any kind except skin cancer; (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 ischaemic attack (TIA); (vii) emotional, nervous or psychiatric problems; and (viii) arthritis or rheumatism.

3.2. Model Specifications and Results

Before discussing our results, we first compare our model with some alternative models to test the quality of model fit that characterizes our model. To compare alternative models, we compute $P(Y_i|Y_{-i})$, which is the posterior predictive distribution of Y_i conditional on the observed data with a single data point deleted. This value is known as the conditional predictive ordinate (CPO) and has been widely used for model diagnostics and assessment (Gelfand *et al.*, 1992). For the ith individual, the CPO statistic under model $M_l : 1 \leq l \leq L$ is defined as

$$\text{CPO}_i = P(Y_i|Y_{-i}) = E_{\underset{\sim}{\theta l}}\left[P\left(Y_i|\underset{\sim}{\theta_l}, Y_{-i}\right)\right] \tag{13}$$

where $-i$ denotes the exclusion of individual i from the sample. The $\underset{\sim}{\theta_l}$ is the set of parameters of the M_l and $P\left(Y_i|\underset{\sim}{\theta_l}\right)$ is the sampling density of the model evaluated at the ith observation. The preceding expectation is taken with respect to the posterior distribution of the model parameter $\underset{\sim}{\theta_l}$ given the cross-validated data, Y_{-i}. Also, note that equation (13) is based on the fact that $\frac{1}{P(Y_i|Y_{-i})} = \int \frac{1}{P(Y_i|\theta)}\pi(\theta|Y)d\theta$. For individual i, the CPO_i can be obtained from the Markov chin Monte Carlo samples by computing the following weighted average:

$$\widehat{\text{CPO}_i} = \left(\frac{1}{M}\sum_{m=1}^{M}\frac{1}{P\left(Y_i|\theta_l^{(m)}\right)}\right)^{-1} \tag{14}$$

where M is the number of simulations and $\theta_l^{(m)}$ denotes the parameter samples at the mth iteration, i.e. they are the draws from the posterior that uses the whole sample. A large 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

$$\text{LPML} = \sum_{i=1}^{n}\log(\widehat{\text{CPO}_i}) \tag{15}$$

[9] It is important to acknowledge that such health measures may be endogenous. We have also explored the approach taken by Terza *et al.* (2008), where we include self-assessed health measures and the generalized residuals associated with modelling self-assessed health. Such an approach is, however, constrained by the shortage of objective measures of health in the HRS to use in modelling self-assessed health. Given such issues, we present the findings related to self-assessed health.

DOI: 10.1002/jae

Greater LPML values represent a better fit. The LPML is well defined under the posterior predictive density, where it is computationally stable. We compare the following models using the LPML values: Model 1, the four-part model proposed in this paper, the results from which are discussed below; Model 2, a four-part model where each part is modelled independently without random effects; Model 3, a four-part model with correlated random effects in a multivariate normal distribution; and Model 4, a four-part model with robust random effects but no age splines or interactions. The LPML values for Models 1–4 are −5405.7, −7198.4, −6201.8 and −61332.4, respectively. Thus Model 1 has the highest LPML value, suggesting that it has the best fit amongst the alternative models. The large difference in the LPML values between our proposed model and the alternative models indicates the presence of a nonlinear age effect and the importance of DP for our analysis.

We formulate an empirical version of the four-part model discussed above to be applied to the HRS data as follows. Equations (16) and (17) present the zero-inflated semi-continuous component of the model that seeks to explain hospital stays. The same covariates are allowed to differentially impact on the propensity for visiting a hospital (in equation (16)) and the count of such visits made (in equation (17)):

$$\begin{aligned}\text{logit}\left(p_{ij}^{H}\right) = {} & \beta_{11}^{p} + \beta_{12}^{p} t_{ij} + \beta_{13}^{p}\text{gedplus}_i + \beta_{14}^{p}\text{female}_i + \beta_{15}^{p}\text{rhlthlm}_{ij} + \beta_{16}^{p}\text{rnodiff}_{ij} \\ & + \beta_{17}^{p}\text{rlive}_{ij} + \beta_{18}^{p}\text{coh}_{ij} + \beta_{19}^{p}\text{coh}_{i,j-1} + f_1^{p}(age_{ij})\text{female}_i \\ & + f_2^{p}(\text{age}_{ij})(1-\text{female}_i) + b_{i1}\end{aligned} \tag{16}$$

$$\begin{aligned}\log\left(\mu_{ij}^{H}\right) = {} & \beta_{11}^{\lambda} + \beta_{12}^{\lambda} t_{ij} + \beta_{13}^{\lambda}\text{gedplus}_i + \beta_{14}^{\lambda}\text{female}_i + \beta_{15}^{\lambda}\text{rhlthlm}_{ij} + \beta_{16}^{\lambda}\text{rnodiff}_{ij} \\ & + \beta_{17}^{\lambda}\text{rlive}_{ij} + \beta_{18}^{\lambda}\text{coh}_{ij} + \beta_{19}^{\lambda}\text{coh}_{i,j-1} + f_1^{\lambda}(\text{age}_{ij})\text{female}_i \\ & + f_2^{\lambda}(\text{age}_{ij})(1-\text{female}_i) + b_{i2}\end{aligned} \tag{17}$$

Similarly, equations (18) and (19) are the two components of the semi-continuous hurdle model for out-of-pocket medical expenses incurred. For both the Poisson hurdle model and the semi-continuous model, age is allowed to flexibly affect both the propensity and the level of health care demand through a smoothing spline that is allowed to vary by gender:

$$\begin{aligned}\text{logit}\left(p_{ij}^{M}\right) = {} & \beta_{11}^{M_p} + \beta_{12}^{M_p} t_{ij} + \beta_{13}^{M_p}\text{gedplus}_i + \beta_{14}^{M_p}\text{female}_i + \beta_{15}^{M_p}\text{rhlthlm}_{ij} + \beta_{16}^{M_p}\text{rnodiff}_{ij} \\ & + \beta_{17}^{M_p}\text{rlive}_{ij} + \beta_{18}^{M_p}\text{coh}_{ij} + \beta_{19}^{M_p}\text{coh}_{i,j-1} + h_1^{p}(\text{age}_{ij})e_{ij1}^{p} \\ & + h_2^{p}\left(\text{age}_{ij}\right)\left(1-e_{ij1}^{p}\right) + b_{i3}\end{aligned} \tag{18}$$

$$\begin{aligned}\log(\mu_{ij}^{M}) = {} & \beta_{11}^{M_\lambda} + \beta_{12}^{M_\lambda} t_{ij} + \beta_{13}^{M_\lambda}\text{gedplus}_i + \beta_{14}^{M_\lambda}\text{female}_i + \beta_{15}^{M_\lambda}\text{rhlthlm}_{ij} + \beta_{16}^{M_\lambda}\text{rnodiff}_{ij} \\ & + \beta_{17}^{M_\lambda}\text{rlive}_{ij} + \beta_{18}^{M_\lambda}\text{coh}_{ij} + \beta_{19}^{M_\lambda}\text{coh}_{i,j-1} + h_1^{M_\lambda}(\text{age}_{ij})e_{ij1}^{M_\lambda} \\ & + h_2^{M_\lambda}(\text{age}_{ij})(1-e_{ij1}^{M_\lambda}) + b_{i2}\end{aligned} \tag{19}$$

Finally, in equations (16)–(19), the random effects $\mathbf{b_i} = (b_{i1}, b_{i2}, b_{i3}, b_{i4})$ are jointly modelled as a DP ($aG_0 \equiv N_4(0, \Sigma)$) and $a \sim \text{Uniform}(0.4, 10)$. To fully specify the Bayesian model, we assign weakly informative conjugate priors for the parameters. For each aggregate-level coefficient, we assume a normal density prior of $\text{N}(0, 100)$. For the variance parameters, we assume inverse-gamma (IG) priors of $\text{IG}(2.01, 1.01)$, giving rise to a prior mean of 1 and a prior variance of 100. Lastly, we take an inverse-Wishart prior for the variance–covariance matrix by assuming

DOI: 10.1002/jae

Table IV. Poisson hurdle model for hospital visits

	Parameter	Mean	95% credible interval
Logit: p^H			
Intercept	$\beta_{11}^{M_p}$	**2.62**	(1.04, 4.33)
Wave	$\beta_{12}^{M_p}$	**1.70**	(0.90, 2.69)
Education: GED or higher?	$\beta_{13}^{M_p}$	0.43	(−2.20, 1.39)
Is female?	$\beta_{14}^{M_p}$	**0.20**	(0.015, 1.67)
95–100% of pre-65 years covered	$\beta_{15}^{M_p}$	−0.39	(−2.32, 1.48)
10–60% of pre-65 years covered	$\beta_{16}^{M_p}$	−0.06	(−2.05, 1.90)
60–95% of pre-65 years covered	$\beta_{17}^{M_p}$	−0.02	(−1.92, 1.88)
Has no difficulty in dressing	$\beta_{18}^{M_p}$	**−1.12**	(−2.68, −0.86)
Self-reported expectation of living 10+ years	$\beta_{19}^{M_p}$	**−0.70**	(−1.13, −0.2)
Self-reported health: Δ in current wave	$\beta_{20}^{M_p}$	**0.70**	(0.19, 1.4)
Self-reported health: Δ in previous wave	$\beta_{21}^{M_p}$	**0.10**	(0.02, 1.01)
Does health limit work?	$\beta_{22}^{M_p}$	−0.18	(−2.18, 1.69)
Respondents BMI	$\beta_{23}^{M_p}$	**1.03**	(0.34, 2.27)
Respondent has at least one chronic condition	$\beta_{24}^{M_p}$	−0.22	(−1.86, -0.1)
Log: μ^H			
Intercept	$\beta_{11}^{M_\lambda}$	−1.03	(−2.60, 0.50)
Wave	$\beta_{12}^{M_\lambda}$	0.13	(−0.02, 0.29)
Education: GED or higher?	$\beta_{13}^{M_\lambda}$	−0.15	(−0.93, 0.59)
Is female?	$\beta_{14}^{M_\lambda}$	−0.18	(−0.70, 0.38)
95–100% of pre-65 years covered	$\beta_{15}^{M_\lambda}$	**0.35**	(0.16, 1.55)
10–60% of pre-65 years covered	$\beta_{16}^{M_\lambda}$	**0.89**	(0.12, 1.35)
60–95% of pre-65 years covered	$\beta_{17}^{M_\lambda}$	**0.70**	(0.19, 1.48)
Has no difficulty in dressing	$\beta_{18}^{M_\lambda}$	**−0.40**	(−1.05, −0.16)
Self-reported expectation of living 10+ years	$\beta_{19}^{M_\lambda}$	0.00	(−0.007, 0.008)
Self-reported health: Δ in current wave	$\beta_{20}^{M_\lambda}$	**0.43**	(0.09, 0.73)
Self-reported health: Δ in previous wave	$\beta_{21}^{M_\lambda}$	0.16	(−0.08, 0.41)
Does health limit work?	$\beta_{22}^{M_\lambda}$	0.61	(−0.002, 1.22)
Respondents BMI	$\beta_{23}^{M_\lambda}$	−0.06	(−0.10, −0.002)
Respondent has at least one chronic condition	$\beta_{24}^{M_\lambda}$	**0.81**	(0.33, 1.31)

$\Sigma^{-1} \sim$ Wishart$(4, 0.1\mathrm{I}_4)$, where I_4 is the 4×4 identity matrix. Each component of this multi-part joint model with robust random effects captures important aspects of demand for health care.[10]

The estimates for the two-part Poisson hurdle model given by equations (16) and (17) are reported in Table IV. The top panel reports the determinants of the propensity for hospital stays, while the bottom panel presents the determinants of the count of hospital stays conditional on stays. It is apparent that flexibility to differentially affect the logit and log portions is important, with many variables behaving

[10] See online Appendix for further details.

DOI: 10.1002/jae

differentially in the two components. There are some exceptions to this: for example, education and health limiting work have no effect on either the logit or the log portion. Similarly, respondents who are able to dress independently are less likely to visit a hospital and visit less frequently conditional on visiting a hospital, indicating the importance of functional abilities. Finally, a deterioration in self-reported health between the current and past waves is associated with an increased propensity to visit the hospital as well as an increase in the count of hospitalizations conditional on there being at least one hospitalization. Other variables, such as being female, having a higher self-reported probability of being alive for at least another 10 years, the lagged change in self-reported health status and the respondent's BMI, affect only the logit portion, with no effect on the log portion. The three insurance history variables, on the other hand, have no effect on the propensity of hospitalization, but are associated with an increase in the count of hospital visits conditional on visiting a hospital. Similarly, having ever had at least one chronic condition affects the log portion of the model but not the propensity to visit hospital. Thus our flexible modelling framework identifies the various different ways that these variables affect the different parts of the demand for hospital visits.

Table V reports estimates from the semi-continuous model for out-of-pocket medical expenditure (OOPMD). A number of interesting differences with the Poisson hurdle model are noted. First, the propensity for any OOPMD is unaffected by education, changes in self-assessed health status and insurance histories covering 10–60% of the individual's pre-65 years. Similarly, a higher self-reported expectation of being alive for the next 10 years is associated with not only a lower propensity of incurring OOPMD but also a lower amount of expenditure if it is positive. While the middle level of insurance coverage prior to 65 has no effect on OOPMD, high levels of coverage (95–100% pre-65 years covered) are associated with a lower propensity of positive OOPMD, but no effect on the level conditional on incurring any expenditure. For low levels of health insurance coverage (10–60%), we find that there is no effect on the propensity for positive OOPMD; however, conditional on positive expenditure, respondents with low levels of health insurance tend to experience higher levels of OOPMD. Being female, or having health conditions that limit work affect the OOPMD distribution in a similar fashion—they both increase the conditional out-of-pocket expenditure on medical care, but have no effect on the propensity of positive OOPMD. Having no difficulty in dressing, higher BMI, and having ever had at least one chronic condition, have a much more complex effect on the OOPMD distribution. These variables reduce the propensity of experiencing positive OOPMD; however, conditional on incurring expenditure, these variables are associated with an increase in OOPMD.

With a large number of variables affecting multiple parts of the four-part model, it seems natural to expect significant correlation across the random effects from each of the components. Some examples of these variables are being female (affecting the logit part of the Poisson model, and the log part of the two-part model), having no difficulty dressing (affecting all four parts of the model), and the self-reported probability of living another 10 years or more (also affecting all four parts of the model). Table VI presents estimates for the correlation coefficients across the four components of the model. Three of the correlation coefficients between the random effects are non-zero; these are the correlation between the random effects of the logit and log components of the Poisson hurdle sub-model (negative), the correlation between the random effects of the log portion of the Poisson hurdle model and the logit portion of the semi-continuous hurdle model (negative), and the correlation between the random effects of the log portion of the Poisson hurdle and log of the semi-continuous model (positive). The first, being negative, suggests that individuals with larger unobserved effects on the propensity of hospitalization tend to have lower unobserved effects on the conditional count of hospital visits. While statistically significant, the correlation coefficient is much smaller (-0.24) than the correlation between the random effects from the conditional count of hospital visits from the Poisson hurdle model and the random effects from the conditional OOPMD component of the semi-continuous model (0.61). The high correlation between the unobserved components of the conditional count of hospital

DOI: 10.1002/jae

Table V. Two-part model for out-of-pocket medical expenses

	Parameter	Mean	95% credible interval
Logit: p^M			
Intercept	$\beta_{11}^{M_p}$	−0.03	(−1.95,2.04)
Wave	$\beta_{12}^{M_p}$	−0.20	(**−2.03**, −0.01)
Education: GED or higher?	$\beta_{13}^{M_p}$	0.10	(−2.07,1.92)
Is female?	$\beta_{14}^{M_p}$	0.03	(−1.90,1.87)
95–100% of pre-65 years covered	$\beta_{15}^{M_p}$	**−0.20**	(**−1.41**, −0.07)
60–95% of pre-65 years covered	$\beta_{17}^{M_p}$	−0.03	(−0.93,0.98)
10 − 60% of pre 65 years covered	$\beta_{16}^{M_p}$	−0.10	(−1.91,1.90)
Has no difficulty in dressing	$\beta_{18}^{M_p}$	**−0.12**	(**−2.16**, −0.07)
Self-reported expectation of living 10+ years	$\beta_{19}^{M_p}$	**−1.01**	(**−1.96**, −0.25)
Self-reported health: Δ in current wave	$\beta_{20}^{M_p}$	0.40	(−0.89,0.95)
Self-reported health: Δ in previous wave	$\beta_{21}^{M_p}$	0.10	(−0.4,0.89)
Does health limit work?	$\beta_{22}^{M_p}$	0.30	(−1.01,1.00)
Respondents BMI	$\beta_{23}^{M_p}$	**−1.21**	(**−2.02**, −0.19)
Respondent has at least one of chronic conditions	$\beta_{24}^{M_p}$	**−0.28**	(**−1.09**, −0.08)
Log: μ^M			
Intercept	$\beta_{11}^{M_\lambda}$	−0.20	(−1.05,1.8)
Wave	$\beta_{12}^{M_\lambda}$	**0.20**	(0.05,1.31)
Education: GED or higher?	$\beta_{13}^{M_\lambda}$	−0.30	(−1.12,1.00)
Is female?	$\beta_{14}^{M_\lambda}$	**0.40**	(0.15,1.17)
95–100% of pre-65 years covered	$\beta_{15}^{M_\lambda}$	−0.20	(**−1.05**, −0.01)
10–60% of pre-65 years covered	$\beta_{16}^{M_\lambda}$	**0.70**	(0.02,1.06)
60–95% of pre 65 years covered	$\beta_{17}^{M_\lambda}$	0.10	(−1.01,0.89)
Has no difficulty in dressing	$\beta_{18}^{M_\lambda}$	**0.22**	(0.02,1.2)
Self–reported expectation of living 10+ years	$\beta_{19}^{M_\lambda}$	**−0.23**	(**−0.48**, −0.03)
Self-reported health: Δ in current wave	$\beta_{20}^{M_\lambda}$	0.10	(−0.88,1.09)
Self-reported health: Δ in previous wave	$\beta_{21}^{M_\lambda}$	0.40	(−0.3,1.8)
Does health limit work?	$\beta_{22}^{M_\lambda}$	**0.70**	(0.05,1.17)
Respondents BMI	$\beta_{23}^{M_\lambda}$	**0.50**	(0.02,1.22)
Respondent has at least one of chronic conditions	$\beta_{24}^{M_\lambda}$	**1.10**	(0.29,2.01)

Table VI. Correlation between random effects across models

	Mean	95% credible interval
Corr. between logit and log of ZIP	**−0.24**	(−0.63, −0.05)
Corr. between logit of ZIP and logit of semi-continuous	0.04	(−0.95, 0.89)
Corr. between log of ZIP and logit of semi-continuous	**−0.23**	(−1.96, −0.08)
Corr. between logit of ZIP and log of semi-continuous	0.01	(−0.08, 0.26)
Corr. between log of ZIP and log of semi-continuous	**0.61**	(0.13, 1.35)
Corr. between logit and log of semi-continuous	0.03	(−0.19,0.16)

DOI: 10.1002/jae

visits and the conditional OOPMD is expected as unobserved factors that determine hospital visits are likely to be closely related to unobserved factors that explain OOPMD. The modestly negative correlation (−0.23) between the log portion of the Poisson hurdle model and the logit of the semi-continuous model is also interesting in that it suggests that unobserved factors that explain the conditional count of hospital visits are negatively associated with the unobserved factors that explain the propensity to experience positive OOPMD.

Finally, we analyse the effect of age on health care demand and how it varies as people age and with gender. Figure 1 plots the effect of aging on each component of the four-part model. Figure 1 shows that the demand for health care varies significantly across a person's life and across gender; note this is not apparent from the baseline effects in the regression tables. The first quadrant of the figure shows that there is a large difference in the baseline levels of demand for health care with women having a higher propensity for making any hospital visits. This finding accords with the existing literature (see, for example, Briscoe, 1987; Bago d'Uva, 2005; Koopmans and Lamers, 2007). For women, the baseline demand for health care does not change until the age of 40, after which it rises linearly until the age of 60. After the age of 60, further aging appears to have almost no additional impact on the propensity to use hospital facilities. Men, on the other hand, have no change in the baseline propensity to visit a hospital until the age of almost 60. Thereafter, the propensity to visit a hospital at least once increases exponentially. The conditional demand for health care in terms of the count of visits behaves somewhat differentially—women visit more frequently over their entire lifetime, while men maintain their baseline rates of hospitalization almost until the age of 60. Thereafter, men start visiting a hospital more frequently than they had in the past. However, the increase is slower than the increase in the conditional counts observed for women.

Similarly, with OOPMD, we find that women are more likely to incur expenditure and they also tend to incur higher expenditure than men at each stage of the life cycle. From the age of 40, the propensity to incur expenditure rises rapidly until the age of 60 and, thereafter, it increases at a much more modest rate for women. For men, there is no change in the baseline propensity of incurring OOPMD until the

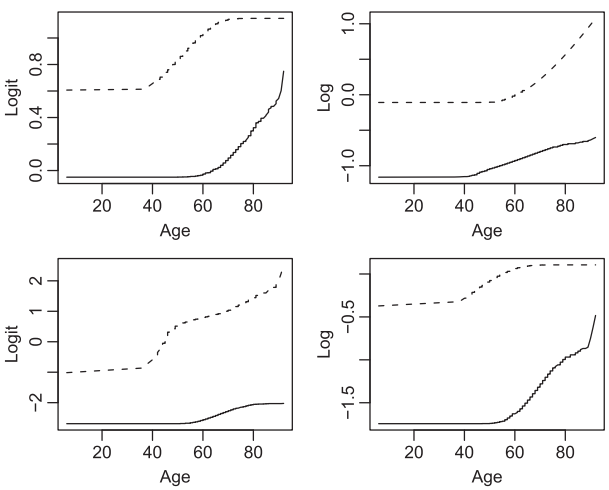

Figure 1. Nonlinear effects of aging for each part of the 4PM. On each plot the x-axis measures age in years. The plotted line is the logit of p_i; the top two plots capture the gender effects of the Poisson hurdle model, where the top left plot captures the difference in the propensity for any hospital visit and no hospital visit for women (dotted line) and men (continuous line). The top right plot captures the conditional count of hospital visits. The bottom two plots capture gender effects in a semi-continuous model, with the bottom left plot capturing gender differences in propensity for any OOPMD, while the bottom right plot captures the gender difference in conditional OOPMD

DOI: 10.1002/jae

age of 60. Thereafter, there is a modest increase in the propensity of incurring any OOPMD. In terms of OOPMD expenditure conditional on positive expenditure, it is clear that women incur substantially higher costs throughout their lifetime than men, with a modest increase after the age of 40. Consistent with the Poisson hurdle model, men have a much lower level of baseline conditional OOPMD expenditure until the age of 60. After the age of 60, conditional OOPMD expenditure increases very rapidly and the gap between male and female medical expenditure declines rapidly, but does not fully go away.

4. CONCLUSION

In this paper, we analyse health care demand for an aging population using a Bayesian semi-parametric joint modelling framework. We incorporate a number of interesting adaptations to this joint model to ensure that our model is appropriate for this application as well as being robust and allowing us to flexibly estimate a key covariate for an aging population, namely the effects of age itself. In the Bayesian framework, we allow for zero inflation, which a key characteristic for both hospital stays and out-of-pocket medical expenditure (Duan *et al.*, 1982; Olsen and Schafer, 2001; Liu *et al.*, 2010). Thus our four-part model differentially captures the propensities for usage as well the levels of use across these two measures of health care demand. This enables us to uncover complex patterns of correlations across a range of covariates and at different portions of the distribution of each outcome. Using DP priors to specify random effects for each participant allows us to reliably estimate health care demand after accounting for unobserved heterogeneity. Finally, the correlation across the components allows us to borrow information across the two measures of health care demand to better understand the co-movement in our joint model in a way that has not been previously applied to health care demand.

The four-part model allows us to capture a number of important aspects relating to how aging influences health care demand. Age splines and their interaction with gender allow us to ascertain that at younger ages health care demand is higher for women, while after the age of 60 health care demand for men increases very rapidly. This affects both hospital visits and out-of-pocket medical expenditure. These findings have different implications: for example, with increased aging, there is need for greater profiling of men as they near 60, which has implications for the health sector, while greater out-of-pocket medical expenses will have important implications for the financial planning of individuals and households as well as for the design of health insurance systems. We hope that our findings will stimulate further research into this area of economics, which is clearly set to increase in terms of its policy relevance in the future.

ACKNOWLEDGEMENTS

We are very grateful to the editor and two referees for excellent comments and detailed suggestions. We would also like to thank Daniel Gray for careful research assistance and Arne Risa Hole and Jennifer Roberts for valuable comments on an earlier version. Pulak Ghosh was supported by grant SR/S4/MS:648/10 from the Department of Science and Technology, Government of India. We are grateful to the National Institute on Aging for supporting the Health and Retirement Study (HRS) (U01 AG09740) to make such high-quality data available for public access.

REFERENCES

Antoniak CE. 1974. Mixtures of Dirichlet processes with applications to nonparametric problems. *Annals of Statistics* **2**: 1152–1174.

Arias E. 2010. *National Vital Statistics Reports*, United States Life Tables, 2006. National Center for Health Statistics, Hyattsville: MD, USA.

DOI: 10.1002/jae

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.

Bago d'Uva T. 2005. Latent class models for use of primary care: evidence from a British panel. *Health Economics* **14**: 873–892.

Blackwell D, MacQueen JB. 1973. Ferguson distributions via polya urn schemes. *Annals of Statistics* **1**: 353–355.

Briscoe ME. 1987. Why do people go to the doctor? Sex differences in the correlates of GP consultation. *Social Science and Medicine* **25**: 507–513.

Card D, Dobkin C, Maestas N. 2008. The impact of nearly universal insurance coverage on health care utilization and health: evidence from Medicare. *American Economic Review* **98**: 2242–2258.

Card D, Dobkin C, Maestas N. 2009. Does Medicare save lives? *Quarterly Journal of Economics* **124**: 597–636.

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

Dobriansky P, Suzman R, Hodes R. 2007. Why population aging matters: a global perspective. http://www.nia.nih.gov/sites/default/files/WPAM.pdf [21 March 2015].

Duan N, Newhouse JP, Morris CN, Manning WG. 1982. A comparison of alternative models for the demand for medical care, *Report R-2754-HHS,* RAND, Santa Monica, CA.

Elashoff R, Li G, Li N. 2008. An approach to joint analysis of longitudinal measurements and competing risks failure time data. *Biometrics* **64**: 762–771.

Ferguson TS. 1973. A Bayesian analysis of some nonparametric problems. *Annals of Statistics* **1**: 209–230.

Finkelstein A, McKnight R. 2008. What did Medicare do? The initial impact of Medicare on mortality and out of pocket medical spending. *Journal of Public Economics* **92**: 1644–1668.

Gelfand AE, Dey DK, Chang H. 1992. Model determination using predictive distributions with implementation via sampling-based methods. In *Bayesian Statistics*, Bernardo JM, Berger JO, Dawid AP, Smith AFM (eds). Oxford University Press: Oxford; 147–167.

Goda GS, Shoven JB, Slavov SN. 2013. Does widowhood explain gender differences in out-of-pocket medical spending among the elderly? *Journal of Health Economics* **32**: 647–658.

Goldman D, Maestas N. 2013. Medical expenditure risk and household portfolio choice. *Journal of Applied Econometrics* **28**: 527–550.

Hartman M, Catlin A, Lassman D, Cylus J, Heffler S. 2008. U.S. health spending by age, selected years through 2004. *Health Affairs* **27**: w1–w12.

Hjort NL, Holmes C, Muller P, Walker SG. 2010. *Bayesian Nonparametrics*. Cambridge University Press: Cambridge, UK.

Hurd MD, McGarry K. 2002. The predictive validity of subjective probabilities of survival. *Economic Journal* **112**: 966–985.

Jochmann M, Leon-Gonzalez R. 2004. Estimating the demand for health care with panel data: a semiparametric Bayesian approach. *Health Economics* **13**: 1003–1014.

Jones AM. 2000. Health econometrics. In *Handbook of Health Economics*, Part A, Culyer AJ, Newhouse JP (eds), Vol. 1. Elsevier: Amsterdam; 265–344.

Koopmans GT, Lamers LM. 2007. Gender and health care utilization: the role of mental distress and help-seeking propensity. *Social Science and Medicine* **64**: 1216–1230.

Lambert D. 1992. Zero-inflated Poisson regression, with an application to defects in manufacturing. *Technometrics* **34**: 1–14.

Little RJA, Rubin DB. 1987. *Statistical Analysis with Missing Data*. Wiley: New York.

Liu L, Conaway MR, Knaus WA, Bergin JD. 2008. A random effects four-part model, with application to correlated medical costs. *Computational Statistics and Data Analysis* **52**: 4458–4473.

Liu L, Strawderman RL, Cowen ME, Shih YCT. 2010. A flexible two-part random effects model for correlated medical costs. *Journal of Health Economics* **29**: 110–123.

McGee DL, Liao Y, Cao G, Cooper RS. 1999. Self-reported health status and mortality in a multiethnic US cohort. *American Journal of Epidemiology* **149**: 41–46.

Mullahy J. 1986. Specification and testing of some modified count data models. *Journal of Econometrics* **33**: 341–365.

Naskar M, Das K. 2006. Semiparametric analysis of two-level bivariate binary data. *Biometrics* **62**: 1004–1013.

Naya H, Urioste JI, Chang YM, Rodrigues-Motta M, Kremer R, Gianola D. 2008. A comparison between Poisson and zero-inflated Poisson regression models with an application to number of black spots in Corriedale sheep. *Genetics Selection Evolution* **40**: 379–394.

Neelon BH, O'Malley AJ, Normand ST. 2010. A Bayesian model for repeated measures zero-inflated count data with application to outpatient psychiatric service use. *Statistical Modelling* **10**: 421–439.

DOI: 10.1002/jae

Olsen M K, Schafer J L. 2001. A two-part random-effects model for semicontinuous longitudinal data. *Journal of the American Statistical Association* **96**: 730–745.

Palumbo M G. 1999. Uncertain medical expenses and precautionary saving near the end of the life cycle. *Review of Economics Studies* **66**: 395–421.

Sethuraman J. 1994. A constructive definition of Dirichlet priors. *Statistica Sinica* **4**: 639–650.

St Clair P, Blake D, Bugliari D, Chien S, Hayden O, Hurd M, Ilchuk S, Kung F, Miu A, Panis C, Pantoja P, Rastegar A, Rohwedder S, Roth E, Carroll J, Zissimopoulos J. 2009. RAND HRS Data Version L, Santa Monica, CA.

Strunk BC, Ginsburg PB, Banker MI. 2006. The effect of population aging on future hospital demand. *Health Affairs* **25**: w141–w149.

Su L, Tom BDM, Farewll VT. 2009. Bias in 2-part mixed models for longitudinal semicontinuous data. *Biostatistics* **10**: 374–389.

Terza JV, Basu A, Rathouz PJ. 2008. Two-stage residual inclusion estimation: addressing endogeneity in health econometric modeling. *Journal of Health Economics* **27**: 531–543.

Tsonaka R, Verbeke G, Lesaffre E. 2009. A semi-parametric shared parameter model to handle nonmonotone nonignorable missingness. *Biometrics* **65**: 81–87.

Wei Y, Ravelo A, Wagner TH, Barnett PG. 2004. The relationships among age, chronic conditions, and healthcare costs. *American Journal of Managed Care* **10**: 909–916.

Winkelmann R. 2004. Health care reform and the number of doctor visits: an econometric analysis. *Journal of Applied Econometrics* **19**: 455–472.

Wu MC, Carroll RJ. 1988. Estimation and comparison of changes in the presence of informative right censoring by modeling the censoring process. *Biometrics* **44**: 175–188.

Yu B, O'Malley AJ, Ghosh P. 2014. Linear mixed models for multiple outcomes using extended multivariate skew-t distributions. *Statistics and Its Interface* **7**: 101–111.

DOI: 10.1002/jae