ROYAL STATISTICAL SOCIETY
DATA | EVIDENCE | DECISIONS

Journal of the Royal Statistical Society
Statistics in Society
Series A

J. R. Statist. Soc. A (2015)
178, *Part* 2, *pp.* 337–361

Private information in healthcare utilization: specification of a copula-based hurdle model

Peng Shi

University of Wisconsin—Madison, USA

and Wei Zhang

Northern Illinois University, DeKalb, USA

[Received January 2012. Final revision January 2014]

Summary. We study whether individuals' private information on health risk affects their medical care utilization. The presence of such information asymmetry is critical to the optimal payment design in healthcare systems. To do so, we examine the relationship between self-perceived health status and healthcare expenditures. Because of simultaneity, we employ a copula regression to model jointly the mixed outcomes, with the association parameter capturing the residual dependence conditional on covariates. The semicontinuous nature of healthcare expenditures leads to a two-part interpretation of private health information: the hurdle component assesses its effect on the likelihood of using medical care services, and the conditional component quantifies its effect on the expenditures given consumption of care. The methodology proposed is applied to a survey data set of a sample of the US civilian non-institutionalized population to test and quantify the effects of private health information. We find evidence of adverse selection in the utilization of various types of medical care services.

Keywords: Copula; GB2 regression; Healthcare; Hurdle model; Private information

1. Introduction

National health expenditures in the USA have experienced a rapid growth over the past few years. In 2012, the total healthcare spending reached US $2.8 trillion, equivalent to about 17.2% of the grosss domestic product in comparison with US $1.4 trillion in the year 2000 (Centers for Medicare and Medicaid Services: `www.cms.gov`). High rising medical bills could reflect the inefficiency of the current healthcare delivery system or the poor design of its payment mechanism and thus lead to the on-going healthcare reform. According to the special report of the Council of Economic Advisers, 'The economic case for health care reform' (The Whitehouse: `www.whitehouse.gov`), the most important source of inefficiency is private or asymmetric information, i.e. one party in a transaction has more information than the other.

In this study, we aim to test the existence of private information from the consumer's side and to quantify its effect on the utilization of medical care services. The foundations of the theory of markets with asymmetric information were established by three famous economists, George A. Akerlof, Joseph E. Stiglitz and A. Michael Spence, who shared the 2001 Nobel Memorial Prize in Economic Sciences. One stylized form of information asymmetry is adverse selection, where 'bad' products or services are selected in a transaction because of private information

Address for correspondence: Peng Shi, School of Business, University of Wisconsin—Madison, Madison, WI 53706, USA.
E-mail: pshi@bus.wisc.edu

 0964–1998/15/178337

that is held by one party. In the health insurance and medical care market, adverse selection arises when individuals who know that they are likely to have high care costs are inclined to seek more insurance. Insurance providers cannot distinguish to the fullest extent healthy and unhealthy consumers, and they charge premiums according to average costs. Consequently, good risk individuals are squeezed out of the market, and higher premiums are charged by insurance companies in response (see Akerlof (1970) and Rothschild and Stiglitz (1976)).

On the basis of theoretical prediction, high risk individuals are more likely to choose high coverage and to consume more care (Rothschild and Stiglitz, 1976). Thus, a positive coverage–risk relationship suggests adverse selection. To test adverse selection in a healthcare market, the majority of the current literature examines the relationship between individuals' insurance coverage and their utilization of healthcare services, i.e. the effect of additional insurance coverage on healthcare utilization (for example, see Holly *et al*. (1998), Vera-Hernández (1999) and Riphahn *et al*. (2003) among others). We also refer to Cutler and Zeckhauser (2000) for an excellent review on adverse selection in the health insurance market. However, the positive coverage–risk correlation could also be explained by the incentive of overutilizing healthcare services for individuals with additional insurance coverage. This effect is known as moral hazard and is another form of asymmetric information. In this sense, the evidence of asymmetric information that was found in the above studies represents a combined effect of adverse selection and moral hazard. Separating adverse selection from moral hazard is still a challenge in economic studies (see Dionne *et al*. (2011, 2013)).

To provide evidence of adverse selection, we adopt an alternative approach by looking directly into the relationship between individuals' self-perceived health and their medical care expenditures. Self-assessed health reveals an individual's private information on her health risk (Doiron *et al*., 2008). Thus, when conditional on observed covariates, any residual correlation with healthcare expenditures indicates adverse selection, i.e. individuals hold private information regarding their health risk which affects the utilization of healthcare services. The simultaneity of self-perceived health and care utilization motivates a joint modelling framework using copulas. More interestingly, the semicontinuous nature of healthcare expenditures leads to a decoupling result, where one copula is used to capture the dependence in the hurdle component whereas another copula is used for the conditional component. The resulting two-part framework has a very intuitive interpretation: the hurdle part examines the effect of hidden health risk on the probability of using care, and the conditional part examines the effect of private health information on the amount of expenditures given the consumption of medical care services.

The copula model demonstrates several advantages in this context. First, in our study, self-assessed health status is a nominal categorical variable, and healthcare expenditure is a semicontinuous variable, characterized by a fraction of 0s and skewed positive values. Copula regression ensures flexibility in jointly modelling discrete and mixed outcomes. Second, the copula model relaxes the strong distributional assumptions that are often used in a system of equations and hence easily accommodates the skewness and heavy tails of healthcare expenditures. Third, treating self-assessed health as endogenous, the copula model addresses the endogeneity issue by specifying the conditional joint distribution of healthcare expenditure and self-perceived health of each individual in a simultaneous modelling framework. Therefore, the partial effects and, thus, adverse selection can be identified through functional forms. Lastly, because of the less restrictive assumptions, the copula model can take a variety of forms and many standard models are covered in a unified framework, which simplifies the process of specification and inference.

In the empirical analysis, we apply the methodology proposed to a data set from the Medical Expenditure Panel Survey (MEPS). We limit our sample to individuals who have some form of private insurance and no form of public health insurance. In the USA, health insurance has

historically not been mandatory, and public insurance programmes are designed to assist special groups of people, such as people aged 65 years and older, people with certain disabilities or low income families and individuals. Because of the above nature of public health insurance, the private insurance market better serves our purpose of examining adverse selection. By examining individuals with only private insurance, we can focus on a more homogeneous population in the empirical test. It should be pointed out that by excluding individuals with public insurance one could underestimate the effect of asymmetric information in the entire health insurance market. It is worth mentioning that the recently enacted Patient Protection and Affordable Act (or 'Obamacare') will overhaul the US healthcare system. However, our data period precedes this Act.

We find that individuals hold private health information that is positively related to both the likelihood of using care and the size of healthcare expenditures given care consumption. To quantify a combined effect from the two components, we further calculate the average treatment effect for each pair of self-perceived health. The results demonstrate the significant effect of asymmetric health information on medical care utilization.

We acknowledge the significant body of empirical work on adverse selection in the health economics literature. Despite existing analyses, our work contributes to the literature in the following aspects. First, current empirical studies have centred on one important implication of adverse selection, i.e. individuals with worse health choose more generous plans. As pointed out earlier, this positive risk–coverage relationship represents a combined effect of adverse selection and moral hazard. In contrast, our direct examination of the effect of private information isolates the dimension of moral hazard. Second, to investigate adverse selection, existing studies have focused on the utilization of health services. Supplementing the literature, we look into not only the likelihood of using care but also the associated health expenditures, and then we quantify the overall effect of private health information on care consumption. Third, in studies based on a specific market section, it is difficult to extrapolate current findings to other populations. In contrast, we examine a probability sample from participants in the entire private health insurance market. Thus, our findings provide evidence for the prevalence of asymmetric health information in the USA among privately insured individuals. Furthermore, the study is of interest to several different constituencies, including microeconomists in general, health statisticians, policy makers and health practitioners, and even the general public with regard to the costs of provision of medical services. Given the common features of the health insurance system across the world and the effects of adverse selection on healthcare consumption, our methodology and results are also applicable to the global market of healthcare services.

The rest of the paper is structured as follows. Section 2 describes the MEPS data and the characteristics of response and explanatory variables. Section 3 introduces the bivariate copula-based hurdle modelling framework and discusses model inference and interpretation. Section 4 summarizes the empirical analysis when applying the methodology to the MEPS data. Section 5 concludes the paper and discusses future research. Some technical details are delegated to Appendix A.

2. Data

2.1. The survey

To examine the private information regarding health risk in medical care utilization, this study considers a data set from the MEPS. Conducted by the Agency for Healthcare Research and Quality, which is a unit of the Department of Health and Human Services, the MEPS is a set of large-scale surveys of families and individuals, their medical providers and employers across the

USA. It currently consists of two major components: the household component providing data for a probability sample of families and individuals, and the insurance component collecting data from a sample of employers on employer-based health insurance plans. We focus on the household component, which is a nationally representative sample of the US civilian non-institutionalized population. Data are collected through an overlapping panel design where measures are taken over 2 years in five rounds of interviews for each panel drawn from the respondents to the National Health Interview Survey. Thus, we are looking at essentially a cross-sectional data set though the MEPS was initially designed to be a long panel survey.

The data contain records on individuals' utilization of medical care, including the uses of various types of services and the associated expenditures between interviews. In the survey, each respondent is asked to evaluate his or her health status, which reveals the inherent unobservable health risk, and thus one's private information, in healthcare utilization. In addition, the data set provides detailed information on person level demographic characteristics, socio-economic status, health conditions, health insurance coverage and employment, among others that could affect both the consumption of medical care and self-judgement of health status. We use the survey data for the year 2008. Our study considers a subsample of non-elderly adults (of ages ranging between 18 and 64 years). Our final sample includes 9737 individuals.

2.2. Data characteristics

To quantify the utilization of heathcare services, this study looks into the expenditures of three categories of medical services: office-based visits, hospital-based visits and emergency room (ER) visits. The expenditures that are associated with office-based visits are composed of payments to doctors including primary care physicians and specialists, and to non-doctors such as psychologists, nurse practitioners and social workers, for events that take place in office-based settings and clinics. The expenditures that are associated with hospital-based events combine the cost for both hospital out-patient visits and in-patient hospital stays. The ER expenditures add the facility expenses and separately billed doctor expenses together. The total expenditures aggregate the utilization of these three types.

A notable feature of the expenditure variables is that they all exhibit a semicontinuous property with a significant fraction of 0s. In our sample, 73% of individuals have at least one office-based visit, 18% go to hospitals at least once and 10% have visits to ERs. Even for the aggregated expenditures, there are over 10% zero utilization. The semicontinuous nature in expenditure variables motivates the bivariate two-part modelling framework in Section 3.

Another observation made from the preliminary analysis is that, for those individuals with positive healthcare expenditures, the distributions of expenditure variables are right skewed with long tails. The skewness and fat tails will be addressed by heavy-tailed regression techniques in this study. To emphasize, we exhibit in Fig. 1 the histograms of positive expenditures related to visits to the office, hospital, ER and the total utilization. Though individuals have higher frequency in office-based visits, it is not surprising to see from Fig. 1 that the costs for hospital and ER visits are much higher instead, given that there is positive consumption for these medical care services. In fact, the average annual expenditures are US $1222, $6312 and $1525 for office-based, hospital and ER visits respectively.

Self-assessed health is an ordinal categorical outcome, being either 'excellent', 'very good', 'good', 'fair' or 'poor'. Table 1 displays the frequency and percentage of self-assessed health status. A majority of sampled individuals perceive themselves as being in a relatively good health condition, with 30.93% rated excellent and 34.85% rated very good. Around 9% of individuals believe that their health status is below average, among which 1.42% are credited as poor. In

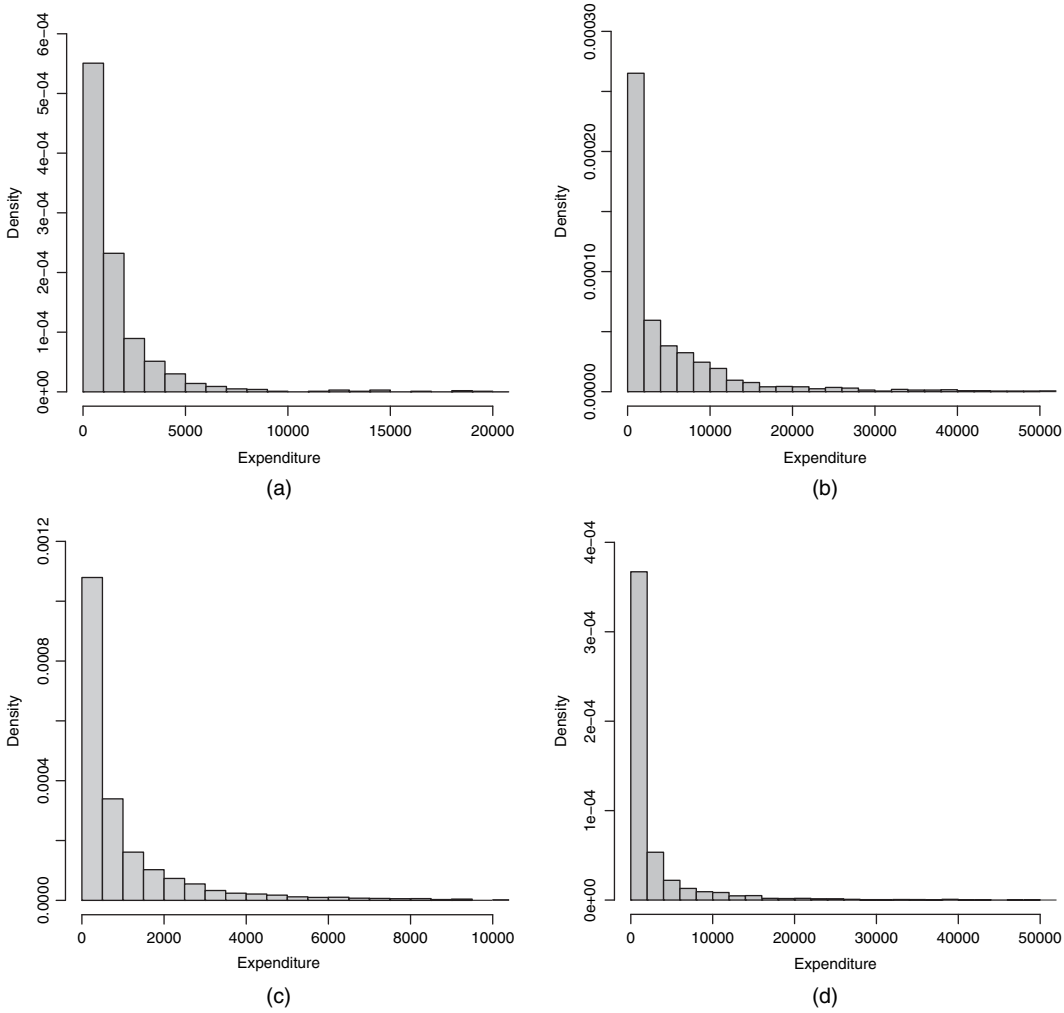

Fig. 1. Histograms of positive healthcare expenditures: (a) office based; (b) hospital; (c) ER; (c) total

general, the distribution between five categories provides enough observations to examine their relative effects on healthcare utilization.

To obtain some knowledge about the relationship between the two responses, we exhibit also in Table 1 the average expenditures by health status for each type of care services. As expected, individuals with better self-perceived health on average consume less healthcare services. For example, the poor health group spends $\$3174 - \$579 = \$2595$ more than the excellent health group on office-based visits. Note that this difference is based on a coarse sample average without purging off the effects of explanatory variables. More discussion on the partial effect of self-assessed health on care utilization is given in Section 4.3.

In addition to the above two response variables, the survey data contain detailed information on the factors that could affect the utilization of medical care and the assessment of health status. The explanatory variables for healthcare utilization and self-assessed health are selected mainly following Dowd *et al.* (1991), Goldman *et al.* (1995) and Deb *et al.* (2006). These covariates are grouped into two categories: socio-economic and demographic characteristics and health

Table 1. Distribution of self-assessed health status and the associated average expenditure

Health	*Health status*			*Average expenditure ($)*			
	Description	*Frequency*	*%*	*Office*	*Hospital*	*ER*	*Total*
1	Excellent	3012	30.93	579	614	108	1301
2	Very good	3393	34.85	846	958	116	1920
3	Good	2480	25.47	1056	1464	183	2704
4	Fair	714	7.33	1395	2622	335	4353
5	Poor	138	1.42	3174	7003	773	10950
Overall		9737	100.00	890	1188	156	2235

Table 2. Description and sample mean of explanatory variables

Variable	*Description*	*Overall*	*Results for the following self-assessed health scores:*				
			1	*2*	*3*	*4*	*5*
age	Age in years	41.04	38.25	41.22	42.65	45.11	47.39
female	1 if the person is female	0.52	0.49	0.54	0.54	0.53	0.51
married	1 if the person if married	0.63	0.61	0.63	0.65	0.63	0.69
hispanic	1 if the person is Hispanic	0.18	0.16	0.17	0.21	0.26	0.27
black	1 if the person is black	0.16	0.14	0.15	0.18	0.18	0.12
asian	1 if the person is Asian	0.08	0.09	0.08	0.08	0.07	0.08
edu	Number of years of education	13.63	14.04	13.87	13.25	12.46	11.91
familysize	Family size	3.07	3.15	3.03	3.04	3.04	3.08
income	Annual income	79.75	89.09	80.94	72.72	63.63	55.88
msa	1 if a metropolitan statistical area	0.87	0.88	0.87	0.86	0.84	0.87
northeast	1 if residence is in the north-east	0.16	0.17	0.16	0.15	0.14	0.12
midwest	1 if residence is in the mid-west	0.22	0.22	0.22	0.23	0.21	0.20
southern	1 if residence is in the south	0.36	0.35	0.37	0.36	0.36	0.41
limitation	1 if the person has a physical limitation	0.15	0.07	0.12	0.20	0.35	0.67
chronic	Number of chronic conditions	1.34	0.68	1.21	1.79	2.66	3.89
smoke	1 if the person smokes	0.17	0.12	0.16	0.21	0.23	0.34

condition indicators. For the former group, we include the age of the respondent, denoted by *age*, as a continuous covariate. The binary variables *female* and *married* indicate the gender and marital status of the respondent respectively. Ethnicity is controlled by three indicators, *hispanic*, *black* and *asian*. Residence information is distinguished by whether the respondent is from a metropolitan statistical area, *msa*, and other geographic indicators. Other socio-economic variables include annual gross income, *income*, number of school years, *edu*, and size of the family, *familysize*. Three covariates are employed to measure the health condition of the respondent: whether the individual has any physical limitation, *limitation*, the number of chronic conditions, *chronic*, and whether the individual is a smoker, *smoke*.

The description and sample mean of the explanatory variables are described in Table 2 for the overall sample and by self-perceived health status. An average person in our sample is an individual of age 41 years who has 14 years of education and has a family of three members.

Over half of individuals are female and 63% of them are married. Survey respondents are well diversified in their ethnicity, with 18% Hispanic, 16% black and 8% Asian. The populations are widely distributed over the country and, among them, about 87% live in a metropolitan statistical area and over a third are from the southern USA. When comparing by self-assessed health, we note a significant difference in many socio-economic characteristics. For example, a self-assessed healthier individual tends to be a younger and wealthier person. Also as expected, the health conditions are highly correlated with an individual's self-perception of his or her health status. Note that, compared with the entire population, the MEPS sampling frames reflect an oversampling of minorities, including blacks and Hispanics.

3. Methodology

Our interest is to examine to what extent self-assessed health and healthcare utilization are related, which testifies to the presence of private information on individuals' health risk when consuming medical care services. In the following presentation, we use y_{i1} and y_{i2} to denote the utilization of healthcare services and the self-assessed health status of individual i respectively.

3.1. Joint modelling using copulas

What complicates our analysis is the semicontinuity and heavy tails of healthcare expenditures and the endogeneity of self-assessed health status. Although health status has been found to be one important factor driving the utilization of medical care services in health economics studies, the endogeneity issue of self-assessed health status has rarely been raised by researchers; for example, see Manning *et al.* (1982), Windmeijer and Santos Silva (1997) and Van Ourti (2004) among others. Self-perceived health is endogenous because it is jointly determined with healthcare utilization. The simultaneity is attributed by the existence of unobserved variables that affect y_{i1} and y_{i2} simultaneously. In this application, such an unobserved factor could be an individual's inherent health risk (see Jürges (2007)), which will affect both medical care consumption and individuals' self-judgement when conditioning on the observed covariates that are available in the survey data. Thus, the evidence of simultaneity is in line with the presence of private health information. Essentially, to detect private information regarding individuals' health risk, we look into the residual dependence between y_{i1} and y_{i2}, i.e. whether unobservable health risk, reflected by self-assessed health status, affects healthcare utilization after controlling for the effects of observable variables.

We deal with the simultaneity of (y_{i1}, y_{i2}) by adopting a full information approach, where the joint distribution of (y_{i1}, y_{i2}) is specified through a parametric copula and the inference for the resulting model is implemented through a likelihood-based method. See Trivedi and Zimmer (2007) for a general introduction to copula modelling in econometrics. The modelling framework that is most relevant to this study is the copula regression technique, where distributions are specified conditionally on a set of regressors. Copula regressions have been extensively employed in various applied disciplines, with an emerging trend on discrete outcomes. Some recent work includes Smith (2003), Zimmer and Trivedi (2006) and Lo and Wilke (2010) in economics, Song *et al.* (2009), Chen (2010) and Madsen and Fang (2011) in statistics, and Frees and Valdez (2008), Shi (2012) and Shi *et al.* (2012) in insurance. In this work, copulas are used for modelling multivariate variables involving both discrete and mixed outcomes.

Using a parametric copula $H(\cdot)$, the joint cumulative distribution function of y_{i1} and y_{i2} could be expressed as

$$F(y_{i1}, y_{i2}|\mathbf{x}_i) = H\{F_1(y_{i1}|\mathbf{x}_{i1}), F_2(y_{i2}|\mathbf{x}_{i2})|\mathbf{x}_i; \boldsymbol{\theta}\} \tag{1}$$

where F_1 and F_2 denote the cumulative distribution functions for the marginals of y_{i1} and y_{i2} respectively. The covariate set $\mathbf{x}_i = (\mathbf{x}_{i1}', \mathbf{x}_{i2}')'$ contains observable explanatory variables for each margin. Vector $\boldsymbol{\theta}$ summarizes association parameters in the copula. In this formulation, the association parameter in the copula describes the dependence between y_{i1} and y_{i2} with the effects of observable covariates purged off. Using u_1 and u_2 to denote the arguments of the bivariate copula $H(\cdot)$, the relationship between the copula and dependence is reflected by the Fréchet–Hoeffding bounds inequality, i.e. $\max\{u_1 + u_2 - 1, 0\} \leqslant H(u_1, u_2) \leqslant \min\{u_1, u_2\}$. The lower or upper bound is attained if one is respectively a decreasing or increasing transformation of the other. Between the two extreme cases, many forms of dependences including linear and non-linear are described by the copula (see Joe (1997) and Nelsen (2006) for more details on copulas and dependence modelling).

3.2. Two-part framework

The above copula framework is not readily applicable in this study because of the semicontinuous healthcare expenditure, which is characterized by a significant portion of 0s and right-skewed continuous positive values. We adopt a bivariate hurdle formulation to specify the statistical process for each of the two sets of configurations of outcomes:

(a) $y_{i1} = 0$ and $y_{i2} = j$, $j = 1, \ldots, 5$, i.e. individual i does not have any healthcare consumption;
(b) $y_{i1} > 0$ and $y_{i2} = j$, $j = 1, \ldots, 5$, i.e. individual i has positive healthcare consumption.

Consistent with the standard hurdle model, the two-part framework reinforces the economic theory that the choice of healthcare utilization and the subsequent expenditures are two different processes with the choice of healthcare services mainly determined by patients whereas the subsequent expenditures are largely influenced by medical providers.

The hurdle component examines the relationship between self-perceived health and the choice of care. Using a parametric copula to accommodate such a relationship, the joint probability mass function for the hurdle part can be described as (see Appendix A for details)

$$\begin{aligned} G^0(y_{i1} = 0, y_{i2} = j|\mathbf{x}_i) &= H^0\{G_1^0(0|\mathbf{x}_{i1}), G_2^0(j|\mathbf{x}_{i2}); \theta^0\} - H^0\{G_1^0(0|\mathbf{x}_{i1}), G_2^0(j-1|\mathbf{x}_{i2}); \theta^0\} \\ G^0(y_{i1} > 0, y_{i2} = j|\mathbf{x}_i) &= G_2^0(j|\mathbf{x}_{i2}) - H^0\{G_1^0(0|\mathbf{x}_{i1}), G_2^0(j|\mathbf{x}_{i2}); \theta^0\} - G_2^0(j-1|\mathbf{x}_{i2}) \\ &\quad + H^0\{G_1^0(0|\mathbf{x}_{i1}), G_2^0(j-1|\mathbf{x}_{i2}); \theta^0\} \end{aligned} \tag{2}$$

for $j = 1, \ldots, 5$. Here, G^0 is the joint probability mass function defined for the pair of discrete random variables in the hurdle part, where the first marginal is a binary outcome indicating whether the individual consumes medical care services and the second marginal is an ordinal categorical outcome representing the scale of self-assessed health status. G_1^0 and G_2^0 are the cumulative distribution functions for the two discrete outcomes; for example, these can be the logit and ordered logit formulation respectively. The function $H^0(\cdot; \theta^0)$, joining the marginals, is a parametric copula with dependence parameter θ^0.

The conditional component examines the relationship between self-perceived health and the amount of healthcare expenditure given positive care consumption. The marginals in the conditional part of our model appear to be mixed outcomes, involving both continuous and discrete responses. The joint probability density function is shown to have the following copula representation (see Appendix A for details):

$$\begin{aligned} g^+(y_{i1}, y_{i2}|y_{i1} > 0, \mathbf{x}_i) = g_1^+(y_{i1}|y_{i1} > 0, \mathbf{x}_{i1})[&h_1^+\{G_1^+(y_{i1}|y_{i1} > 0, \mathbf{x}_{i1}), G_2^+(y_{i2}|y_{i1} > 0, \mathbf{x}_{i2}); \theta^+\} \\ &- h_1^+\{G_1^+(y_{i1}|y_{i1} > 0, \mathbf{x}_{i1}), G_2^+(y_{i2} - 1|y_{i1} > 0, \mathbf{x}_{i2}); \theta^+\}], \end{aligned} \tag{3}$$

where g^+ is the joint probability density function defined for the pair of mixed outcomes in the conditional part, the continuous healthcare expenditure and the discrete self-assessed health, given that there is healthcare consumption. g_1^+ and G_1^+ denote the density and distribution functions of the positive healthcare expenditure respectively, and G_2^+ represents the cumulative distribution function of $y_{i2}|y_{i1}>0$. Assuming that $H^+(\cdot;\theta^+)$ is a copula that joins the marginals, $h_1^+(\cdot;\theta^+)$ is then defined as

$$h_1^+(u_1,u_2;\theta^+)=\frac{\partial}{\partial u_1}H(u_1,u_2;\theta^+),$$

where θ^+ is the dependence parameter that is associated with copula $H^+(\cdot;\theta^+)$.

Skewed outcomes are often encountered in healthcare studies and regression techniques based on generalized distributions have been proposed to accommodate skewness and long tails (see, for example, Manning *et al.* (2005) and Liu *et al.* (2010)). Supplementing the existing literature, we consider the generalized beta distribution of the second kind, GB2, for positive healthcare expenditures. The GB2 distribution was introduced by McDonald (1984) and its applications have been found extensively in the economics literature. Some examples include McDonald (1987), McDonald and Butler (1987) and Cummins *et al.* (1990). Despite the flexibility of the GB2 distribution in modelling skewed outcome data, its application in regression analysis is still sparse. Details on GB2 regression are provided in Appendix B.

In our application, we assume that the probability distribution of $y_{i1}|y_{i1}>0$ follows the GB2 distribution, i.e. $g_1^+(y_{i1}|y_{i1}>0,\mathbf{x}_{i1})=f_{\mathrm{GB2}}(y_{i1};\mu_{i1},\sigma,\phi_1,\phi_2)$, where the location parameter is further modelled as a linear combination of covariates to control for the observed heterogeneity, $\mu_{i1}=\mathbf{x}_{i1}'\boldsymbol{\beta}^+$. Here $\boldsymbol{\beta}^+$ is the corresponding vector of regression coefficients. From expression (9) in Appendix B, straightforward calculations show that $\beta_k^+=\partial\ln\{E(y_{i1}|y_{i1}>0,\mathbf{x}_{i1})\}/\partial x_{i1,k}$, meaning that, when $x_{i1,k}$ increases by 1 unit, given that other covariates remain the same, the expected value of y_{i1} will increase proportionally by $\exp(\beta_k^+)$. Here subscript k denotes the kth component of the parameter vector. Note that one could permit scale parameter σ to depend on covariate information $\mathbf{x}_{i1}$ to capture potential subject level heteroscedasticity in the data (see Lawless (2003)). Though not reported here, our analysis shows that the extra flexibility does not improve model performance substantially. Thus we choose a parsimonious specification.

The final step towards a full information specification involves the choice of copulas. In this context, we are interested in whether individuals have private information regarding their health risk that affects healthcare utilization. Thus the null hypothesis in our analysis is that medical care utilization and self-assessed health status are independent conditionally on the observable explanatory variables. Because we have no *a priori* knowledge regarding the direction of the association between outcomes, we explore copulas that could accommodate both positive and negative dependence. We consider several choices that permit such flexibility. Their distributions and corresponding partial densities are supplied in Appendix C.

The above framework allows us to examine the semicontinuous expenditure outcome and the discrete health variable jointly. The resulting two-part formulation leads to an interesting interpretation of the modelling results: the hurdle component informs us about the effects of self-assessed health on the probability of utilizing medical care services through the association parameter θ^0 of copula H^0. The dependence parameter θ^+ of copula H^+ in the conditional component reveals the relationship between healthcare expenditure and self-assessed health, given positive consumption of healthcare services. Note that we have the flexibility to specify different copulas for the hurdle and the conditional components. It is straightforward to see that the standard hurdle model is a special case of the bivariate model with a product copula, where the equations of self-perceived health and care utilization are assumed to be independent. Thus,

the superiority of the bivariate approach could be supported by the evidence of simultaneity, which can be evaluated by using the usual test statistics as demonstrated later in Table 7 in Section 4.3.

In this work, we employ the hurdle-type model to handle the semicontinuous component in the bivariate modelling framework. An alternative approach to incorporating a mass probability at zero in an otherwise continuous distribution would be to use a censoring model, where the observed healthcare expenditure is equal to either 0 or a desired healthcare expenditure, whichever is higher. We choose the hurdle-type model over the censoring-type model because, first, the censored regression model usually requires a strong distributional assumption on the continuous component. One typically assumes normality and the resulting formulation is known as the 'tobit' model. In contrast, the two-part framework retains flexibility in the specification of the amount distribution. This property is more suitable for healthcare applications where the amount component often exhibits skewness and long tails. Second, the hurdle model is more in line with the economic theory on healthcare demand, where healthcare utilization involves a two-part decision process. It is an individual's decision to seek treatment whereas the physician mainly determines the intensity of expenditures. Third, the two-part model allows for separate sets of covariates for the hurdle and conditional components. As predicted by economic theory, variables that affect the hurdle part may differ from those that affect the conditional part in the context of care consumption. Indeed, as we show in Section 4, the individual characteristics are less relevant in the conditional component than in the hurdle component.

3.3. Inference

Because of the parametric nature of the model, parameters could be estimated by using likelihood-based methods. Using equations (2) and (3), we can show that the log-likelihood functions for the hurdle and conditional components are

$$\begin{aligned} \mathrm{ll}^0 = & \sum_{\{i:y_{i1}=0\}} \log\left\{ \prod_{j=1}^{5} G^0(y_{i1}=0, y_{i2}=j|\mathbf{x}_i)^{I(y_{i2}=j)} \right\} \\ & + \sum_{\{i:y_{i1}>0\}} \log\left\{ \prod_{j=1}^{5} G^0(y_{i1}>0, y_{i2}=j|\mathbf{x}_i)^{I(y_{i2}=j)} \right\} \end{aligned} \tag{4}$$

and

$$\begin{aligned} \mathrm{ll}^+ = & \sum_{\{i:y_{i1}>0\}} (\log\{g_1^+(y_{i1}|y_{i1}>0, \mathbf{x}_{i1})\} + \log[h_1^+\{G_1^+(y_{i1}|y_{i1}>0, \mathbf{x}_{i1}), G_2^+(y_{i2}|y_{i1}>0, \mathbf{x}_{i2}); \theta^+\} \\ & - h_1^+\{G_1^+(y_{i1}|y_{i1}>0, \mathbf{x}_{i1}), G_2^+(y_{i2}-1|y_{i1}>0, \mathbf{x}_{i2}); \theta^+\}]) \end{aligned} \tag{5}$$

respectively. We employ a full information maximum likelihood approach to estimate the parameters in the marginal regression models and the copula functions simultaneously. Alternative estimation techniques for copula models and their advantages and disadvantages are discussed in Trivedi and Zimmer (2007).

Though the two-part framework examines the effect of self-perceived health on medical care utilization from two perspectives separately, the overall effect could be studied by estimating the quantity

$$\Delta(j, k|\bar{\mathbf{x}}_i) = E(y_{i1}|y_{i2}=k, \bar{\mathbf{x}}_i) - E(y_{i1}|y_{i2}=j, \bar{\mathbf{x}}_i) \tag{6}$$

for an average individual. This quantity is known as the partial effect at average, PEA, where explanatory variables are set to their sample mean, and the term $E(y_{i1}|y_{i2}=j,\bar{\mathbf{x}}_i)$ could be calculated from the two-part framework:

$$E(y_{i1}|y_{i2}=j,\bar{\mathbf{x}}_i)=\text{Prob}(y_{i1}>0|y_{i2}=j,\bar{\mathbf{x}}_i,\theta^0)\,E(y_{i1}|y_{i1}>0,y_{i2}=j,\bar{\mathbf{x}}_i,\theta^+).$$

Specifically, from the estimate of copula $H^0(\cdot;\theta^0)$ in the hurdle component, we can derive the probability of utilizing healthcare conditionally on self-assessed health as follows:

$$\begin{aligned}\text{Prob}(y_{i1}>0|y_{i2}=j,\mathbf{x}_{i1},\mathbf{x}_{i2},\theta^0)&=\frac{\text{Prob}(y_{i1}>0,y_{i2}=j|\mathbf{x}_{i1},\mathbf{x}_{i2},\theta^0)}{\text{Prob}(y_{i2}=j|\mathbf{x}_{i2})}\\&=\frac{G_2^0(j|\mathbf{x}_{i2})-H^0\{G_1^0(0|\mathbf{x}_{i1}),G_2^0(j|\mathbf{x}_{i2});\theta^0\}-G_2^0(j-1|\mathbf{x}_{i2})+H^0\{G_1^0(0|\mathbf{x}_{i1}),G_2^0(j-1|\mathbf{x}_{i2});\theta^0\}}{G_2^0(j|\mathbf{x}_{i2})-G_2^0(j-1|\mathbf{x}_{i2})}\end{aligned}$$

where the numerator could be estimated on the basis of a specific copula function, and the denominator is estimated by an ordered logit model. By substituting the maximum likelihood estimates for the model parameters, we can calculate this conditional probability.

Similarly, we can derive the conditional density of healthcare expenditures given both self-perceived health and positive care consumption from copula $H^+(\cdot;\theta^+)$ in the conditional component. The conditional density has the following copula representation:

$$\begin{aligned}f(y_{i1}|y_{i2},y_{i1}>0,\mathbf{x}_{i1},\mathbf{x}_{i2},\theta^+)&=\frac{f(y_{i1},y_{i2}|y_{i1}>0,\mathbf{x}_{i1},\mathbf{x}_{i2},\theta^+)}{f(y_{i2}|y_{i1}>0,\mathbf{x}_{i2})}\\&=g_1^+(y_{i1}|y_{i1}>0,\mathbf{x}_{i1})\left[\frac{h_1^+\{G_1^+(y_{i1}|y_{i1}>0,\mathbf{x}_{i1}),G_2^+(y_{i2}|y_{i1}>0,\mathbf{x}_{i2});\theta^+\}}{G_2^+(j|\mathbf{x}_{i2})-G_2^+(j-1|\mathbf{x}_{i2})}\right.\\&\quad\left.-\frac{h_1^+\{G_1^+(y_{i1}|y_{i1}>0,\mathbf{x}_{i1}),G_2^+(y_{i2}-1|y_{i1}>0,\mathbf{x}_{i2});\theta^+\}}{G_2^+(j|\mathbf{x}_{i2})-G_2^+(j-1|\mathbf{x}_{i2})}\right]\end{aligned}$$

where the copula function is from the conditional model, g_1^+ and G_1^+ correspond to the GB2 distribution and G_2^+ denotes the conditional cumulative logit model. From the above density function, we can further calculate the expected expenditure given consumption of medical care services:

$$E(y_{i1}|y_{i2},y_{i1}>0,\mathbf{x}_{i1},\mathbf{x}_{i2},\theta^+)=\sum_{y_{i1}>0}y_{i1}\,f(y_{i1}|y_{i2},y_{i1}>0,\mathbf{x}_{i1},\mathbf{x}_{i2},\theta^+).$$

4. Results

This section presents the estimation results for the bivariate copula-based hurdle model by using the MEPS data. We discuss the estimate for the hurdle part and the conditional part separately. Furthermore, we discuss the evidence of private information regarding health risk and quantify its effect on medical care utilization.

4.1. Hurdle model

The hurdle model (equations (2) and (4)) is estimated for four measures of utilization (office based, hospital, ER and total visits) by using the Frank copula. As we would expect, given the structure of the joint model of medical care utilization and self-assessed health, parameter estimates from these models share some similarity. Consequently, we report the estimation results selectively.

Table 3 shows the marginal effects for self-perceived health status, estimated from the joint model with office-based expenditures as the utilization measure. Except for age and residence, all explanatory variables are significant determinants of self-assessed health. Recall that health is ordered from good risk to bad risk; thus a positive or negative coefficient corresponds respectively to a lower or higher likelihood of a perception of better health. For example, wealthier and well-educated individuals are more likely to perceive themselves to be healthy, whereas female and married individuals are more likely to perceive themselves as being unhealthy. It is important to note that all three health condition variables show a statistically significant effect, i.e. the worse the health condition variables are, the more likely is the individual's self-perception of bad health.

To quantify the goodness of fit for the marginal of self-perceived health, we compare the actual and fitted frequencies of health status as shown in Table 4. To derive the fitted frequency, we calculate the probability of being in the jth ($j = 1, \ldots, 5$) health category for each individual in the sample; then we average the probabilities across all observations. For example, we observe that among the sample 30.93% rate their health as excellent and 1.42% as poor. Our copula model predicts that 30.88% of individuals are of excellent health and 1.46% are of poor health. The close match between the actual and fitted frequencies suggests a satisfactory fit of the copula model. We also perform a formal χ^2-test. The small statistic reported at the bottom of Table 4 is consistent with the hypothesis of no difference between the predicted and sample distribution of health.

The second marginal in the hurdle copula model concerns healthcare utilization. Table 5 summarizes the estimate for the use of office-based medical care, hospital and ER as well as total care utilization. In the hurdle component, the marginal informs about the probability of using a certain type of care. Most socio-economic characteristics significantly affect the likelihood of visiting an office or hospital, whereas only a few of them are statistically significant in the equation for ER visits. Among health condition variables, physical limitation and chronic conditions certainly increase the chance of care utilization. Across different types of care, some

Table 3. Estimation of self-assessed health status

Parameter	*Estimate*	*Standard error*	*p-value*
intercept1	−1.158	0.162	<0.001
intercept2	0.551	0.162	0.001
intercept3	2.565	0.164	<0.001
intercept4	4.689	0.183	<0.001
age	−0.002	0.002	0.260
female	0.136	0.038	<0.001
married	0.120	0.046	0.009
hispanic	0.522	0.056	<0.001
black	0.302	0.057	<0.001
asian	0.539	0.074	<0.001
edu	−0.080	0.008	<0.001
familysize	0.045	0.015	0.002
income	−3.437	0.388	<0.001
msa	0.007	0.057	0.904
northeast	−0.094	0.062	0.129
midwest	0.009	0.058	0.878
southern	−0.060	0.052	0.248
limitation	0.710	0.057	<0.001
chronic	0.491	0.015	<0.001
smoke	0.412	0.052	<0.001

Table 4. Goodness of fit for health

Health	*Observed*	*Fitted*
1	30.93	30.88
2	34.85	34.74
3	25.47	25.51
4	7.33	7.42
5	1.42	1.46
χ^2-statistic		0.273

Table 5. Estimation of healthcare utilization: hurdle component

Parameter	*Results for the following types of service:*							
	Office based		*Hospital*		*ER*		*Total*	
	Estimate	*Standard error*	*Estimate*	*Standard error*	*Estimate*	*Standard error*	*Estimate*	*Standard error*
intercept	−1.353†	0.207	−3.566†	0.247	−1.528†	0.287	−1.154†	0.212
age	0.000	0.002	0.008†	0.003	−0.017†	0.003	−0.001	0.003
female	1.024†	0.052	0.811†	0.058	0.151‡	0.069	1.014†	0.053
married	0.360†	0.061	0.344†	0.070	−0.058	0.082	0.346†	0.062
hispanic	−0.309†	0.072	−0.324†	0.089	−0.060	0.105	−0.283†	0.073
black	−0.531†	0.074	−0.172‡	0.084	0.303†	0.095	−0.397†	0.076
asian	−0.559†	0.093	−0.336†	0.119	−0.472†	0.168	−0.563†	0.094
edu	0.081†	0.011	0.042†	0.013	−0.029§	0.015	0.079†	0.011
familysize	−0.118†	0.019	−0.025	0.022	−0.016	0.027	−0.116†	0.019
income	2.939†	0.531	0.087	0.546	−0.934	0.736	2.849†	0.545
msa	0.249†	0.076	−0.153§	0.080	−0.015	0.101	0.247†	0.078
northeast	0.259†	0.084	0.394†	0.092	0.311†	0.109	0.200‡	0.086
midwest	0.172‡	0.077	0.738†	0.083	0.137	0.104	0.187‡	0.080
southern	0.024	0.068	0.147§	0.081	−0.074	0.098	−0.028	0.070
limitation	0.690†	0.095	0.650†	0.071	0.701†	0.086	0.812†	0.104
chronic	0.552†	0.027	0.259†	0.019	0.175†	0.023	0.584†	0.028
smoke	−0.306†	0.068	−0.229†	0.079	0.279†	0.085	−0.206†	0.070

†Significant at the 1% level.
‡Significant at the 5% level.
§Significant at the 10% level.

covariates affect the likelihood of consumption in the same direction, and others in opposite ways. For example, female and married individuals have a higher chance of using all types of care. However, better education increases the probability of visiting offices and hospitals but decreases the probability of visiting an ER, which supports the hypothesis that better educated people invest more in preventative care (see, for example, Kenkel (1994)). This observation suggests both a substitution effect and a complementary effect among different types of medical care. Cares are complements if, when an individual has a higher chance of using a certain type of care, he or she has a higher chance of using all types of care. In the meantime, types of care may be substitutes because, if one type of care could improve health, the likelihood of using other types of care will be reduced.

4.2. Conditional model

The conditional component of the copula model examines the expenditure of various types of medical care given that there is care utilization. This subsection discusses the results of the marginal models. To be consistent with the hurdle component, we estimate the model (equations (3) and (5)) by using the Frank copula. Table 6 displays the marginal effects for the expenditures of office-based, hospital, ER and total care utilizations. Many covariates are statistically significant for office-based expenditures but much less so for hospital and ER expenditures. Similarly to the hurdle part, the effects of the same explanatory variable on various types of expenditure might not agree with each other, presumably because of the mixed effects of substitution and complementarity. Very intuitively, a worse health status indicated by physical limitation or chronic conditions leads to higher healthcare expenditures regardless of the type of care. A perhaps surprising observation in Table 6 is that income is not significant in explaining health expenditures. This is, however, consistent with the health economics literature, which found that, whereas national income is an important factor in determining national health expenditures, individual income is not significant in explaining individual health expenditures under insurance (see Getzen (2000)).

To examine the goodness of fit for the marginals, we look into the residuals from the GB2 regression. A residual is defined as $\varepsilon_{i1} = \{\ln(y_{i1}) - \hat{\mu_{i1}}\}/\hat{\sigma}$, where $\hat{\mu}_{i1} = \mathbf{x}'_{i1}\hat{\boldsymbol{\beta}}^{+}$, and $\hat{\boldsymbol{\beta}}^{+}$ and $\hat{\sigma}$ are maximum likelihood estimates of the parameters. If the marginal model is appropriate, then the residuals are approximately independent and identically distributed for some large sample

Table 6. Estimation of health care expenditures: conditional component

Parameter	*Results for the following types of service:*							
	Office based		*Hospital*		*ER*		*Total*	
	Estimate	*Standard error*	*Estimate*	*Standard error*	*Estimate*	*Standard error*	*Estimate*	*Standard error*
intercept	4.041†	0.248	26.740†	4.746	7.138†	0.491	−8.442†	1.276
age	0.007†	0.002	−0.016†	0.005	0.001	0.004	0.005†	0.002
female	0.504†	0.031	−0.074	0.094	0.193‡	0.074	0.525†	0.036
married	0.168†	0.038	0.238‡	0.109	0.119	0.088	0.201†	0.045
hispanic	−0.114‡	0.047	0.003	0.148	−0.031	0.110	−0.088	0.054
black	−0.250†	0.048	−0.034	0.132	−0.168§	0.101	−0.182†	0.055
asian	−0.297†	0.062	−0.172	0.198	−0.406‡	0.190	−0.280†	0.072
edu	0.053‡	0.007	0.019	0.019	0.015	0.016	0.045†	0.008
familysize	−0.062†	0.012	0.114§	0.036	0.011	0.029	−0.049†	0.014
income	0.271	0.297	−0.572	0.924	0.463	0.765	0.049	0.344
msa	0.153†	0.048	0.067	0.120	0.104	0.108	0.090	0.056
northeast	0.062	0.049	−0.163	0.147	−0.187	0.119	0.187	0.058
midwest	−0.063	0.047	−0.262‡	0.131	−0.062	0.113	0.105§	0.054
southern	−0.076§	0.042	0.012	0.131	−0.016	0.108	−0.031	0.049
limitation	0.430†	0.043	0.432†	0.102	0.061	0.087	0.532†	0.050
chronic	0.164†	0.011	0.109†	0.027	0.046§	0.024	0.209†	0.013
smoke	−0.162†	0.044	−0.035	0.122	0.024	0.090	−0.127‡	0.050
SIGMA	1.920†	0.300	6.904‡	2.847	0.957†	0.270	4.900†	0.489
ALPHA1	5.486†	1.616	15.340	12.777	1.264†	0.483	162.053†	31.816
ALPHA2	4.369†	1.232	256.400†	88.810	2.856§	1.591	10.994†	2.161

†Significant at the 1% level.
‡Significant at the 5% level.
§Significant at the 10% level.

size. We use a quantile–quantile (*qq*-) plot of residuals to assess model fit. Fig. 2 exhibits the GB2 *qq*-plots for the positive expenditures of office-based, hospital, ER and total visits. The alignment along the 45° line suggests that the GB2 distribution provides a reasonable fit for all four types of expenditure. Plots of residuals against covariates can also be used to check the heterogeneity in the scale parameter. Our analysis does not detect a pattern in these plots, which confirms the assumption of constant scale.

The marginal model of self-perceived health status examines the factors that affect individuals' judgement of their own health among those who have consumed at least some medical care services during the year. We choose not to report the estimates of the conditional marginal effects because the analysis does not add new insights into the determination of self-assessed health risk when compared with the hurdle component. The goodness of fit is examined in the same manner as in Table 4 and statistical tests show a reasonable fit of the conditional ordered logit regressions.

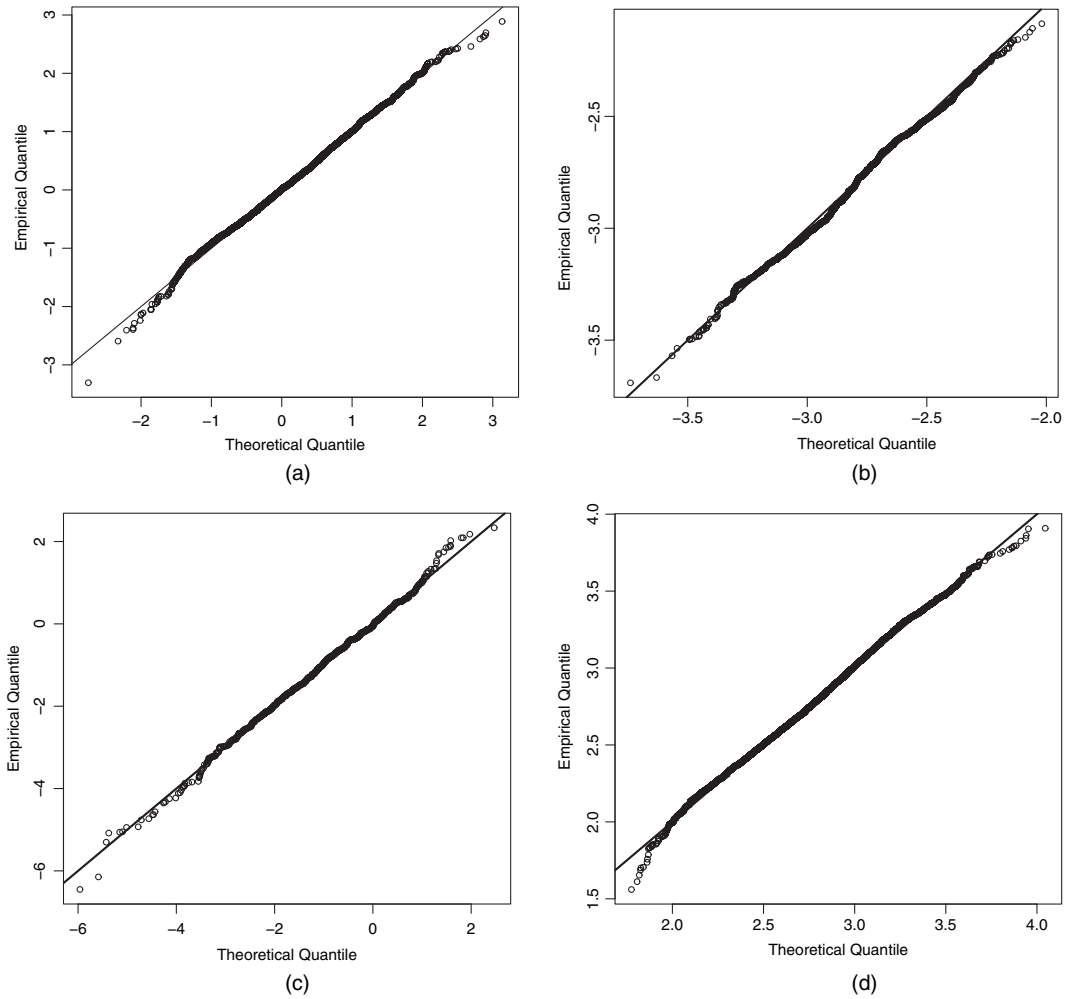

Fig. 2. GB2 distribution *qq*-plots for positive healthcare expenditures: (a) office based; (b) hospital; (c) ER; (d) total

4.3. Dependence and private information

This section discusses the dependence between healthcare utilization and self-assessed health status. One property of the copula framework proposed is that such dependence is estimated in both hurdle and conditional components, where separate copulas are allowed for each component. The dependence is captured by the association parameter in the corresponding parametric copula, where the former measures the association between the probability of using medical care and self-perceived health and the latter examines the relationship between the amount of expenditures and self-perceived health given that there is medical care consumption.

Four types of medical care expenditures, corresponding to office-based visits, hospital visits, ER visits and total care usage, are considered in this study. Table 7 summarizes the estimated dependence parameters in the Frank copula model with the associated estimation errors. We observe a positive relationship between healthcare utilization and self-perceived health in both the hurdle and the conditional components. We also note the similarity of the dependence parameters in both parts for each type of care. In fact, the two-part framework allows us to test this equality. For example, for the hypothesis that the hurdle and conditional dependence parameters are the same for office-based healthcare expenditures, we can consider the test statistic

$$\hat{z} = \frac{0.533 - 0.505}{\sqrt{(0.094^2 + 0.076^2)}} = 0.233,$$

which indicates that the difference in parameters is not significant at the conventional level.

The most important implication drawn from Table 7 is the presence of private information in the consumption of medical care services. In our application, the dependence parameter θ in the copula captures the residual correlation between the two response variables after purging off the effects of explanatory variables. Thus, the positive association that is represented by θ indicates that, given the observed information from covariates, individuals still hold additional information regarding their inherent health risk which significantly affects their consumption of medical care services. More interestingly, our formulation demonstrates the effect of such private information from two perspectives: the hurdle component suggests that an individual of higher health risk has a higher chance of using healthcare services, and the conditional component implies that a bad health risk is also associated with high expenditures given the utilization of healthcare. Such a relationship is found for all three types of care. To show the statistical significance of the positive association, we perform a Wald test and a likelihood ratio test. The results for the two statistical tests are also reported in Table 7. The small p-values

Table 7. Dependence of the Frank copula in the hurdle and conditional components

Type of service	θ	*Standard error*	*Wald*	*p-value*	*Likelihood ratio*	*p-value*
Hurdle component						
Office	0.533	0.094	32.167	<0.001	32.346	<0.001
Hospital	0.509	0.103	24.486	<0.001	24.702	<0.001
ER	0.613	0.125	23.896	<0.001	24.152	<0.001
Total	0.573	0.097	35.227	<0.001	35.472	<0.001
Conditional component						
Office	0.505	0.076	44.482	<0.001	44.654	<0.001
Hospital	0.407	0.145	7.895	<0.005	7.956	<0.005
ER	0.656	0.205	10.214	<0.005	10.274	<0.005
Total	0.572	0.075	58.578	<0.001	58.860	<0.001

demonstrate the significant effect of private health information on medical care utilization and thus give evidence of adverse selection in the private healthcare market. It is worth mentioning that, though the ER is often seen as a way of receiving medical care for the uninsured population, our analysis excludes individuals without any private insurance. Hence it should not be a major concern from the perspective of testing adverse selection.

In addition, two types of robustness test are performed to validate the above relationship. The first test is with regard to the functional forms of the copula. In addition to the Frank copula, we also estimate the model with two other representative copulas for each component: the Farlie–Gumbel–Morgenstern copula and the Gaussian copula. The Gaussian copula belongs to the family of elliptical copulas. The Farlie–Gumbel–Morgenstern copula is more restrictive in dependence modelling. We report the estimates of association parameters in Table 8. The dependences that are captured by different copulas are consistent for both the hurdle and the conditional components, suggesting that the results are robust to various parametric distributional forms of the copula function.

The second test concerns the evidence of private information. It is arguable that the conditional dependence between self-perceived health and care utilization could be caused by any exogenous health shock to an individual after the selection of health insurance, i.e. for instance a detrimental health shock may cause one's self-assessed health to deteriorate and care utilization to increase simultaneously, thus resulting in the positive relationship between the two. If this happens after

Table 8. Robustness test on the dependence parameters in the hurdle and conditional components

	θ	*Standard error*	*p-value*	θ	*Standard error*	*p-value*
	Office-based healthcare			*Hospital healthcare*		
Hurdle component						
Farlie–Gumbel–Morgenstern copula	0.264	0.046	<0.001	0.254	0.051	<0.001
Gaussian copula	0.094	0.016	<0.001	0.092	0.017	<0.001
Invariant health group	0.639	0.132	<0.001	0.687	0.146	<0.001
Varying health group	0.357	0.135	0.008	0.320	0.146	0.028
	ER healthcare			*Total*		
Farlie–Gumbel–Morgenstern copula	0.297	0.061	<0.001	0.283	0.047	<0.001
Gaussian copula	0.098	0.019	<0.001	0.102	0.016	<0.001
Invariant health group	0.805	0.186	<0.001	0.643	0.135	<0.001
Varying health group	0.414	0.170	0.015	0.449	0.139	0.001
	Office-based healthcare			*Hospital healthcare*		
Conditional component						
Farlie–Gumbel–Morgenstern copula	0.253	0.038	<0.001	0.208	0.073	<0.005
Gaussian copula	0.096	0.013	<0.001	0.080	0.025	<0.005
Invariant health group	0.702	0.107	<0.001	0.642	0.202	0.002
Varying health group	0.292	0.108	0.007	0.193	0.208	0.356
	ER healthcare			*Total*		
Farlie–Gumbel–Morgenstern copula	0.302	0.097	<0.005	0.284	0.037	<0.001
Gaussian copula	0.107	0.034	<0.005	0.106	0.013	<0.001
Invariant health group	0.756	0.306	0.014	0.754	0.105	<0.001
Varying health group	0.501	0.274	0.068	0.366	0.106	0.001

Table 9. Effects of self-assessed health on health care utilization

Health score	*Results for the following types of services:*					
	Office based		*Hospital*		*ER*	
	Hurdle probability	*Conditional mean*	*Hurdle probability*	*Conditional mean*	*Hurdle probability*	*Conditional mean*
1	0.752	877	0.134	5493	0.075	1235
2	0.784	981	0.155	5928	0.090	1331
3	0.813	1101	0.179	6657	0.109	1511
4	0.825	1163	0.191	7167	0.118	1671
5	0.828	1177	0.194	7315	0.120	1732

the individual's selection of insurance, it does not constitute information asymmetry between the individual and the insurer. To test this hypothesis, we divide individuals into two groups: one group with invariant self-perceived health and one group with varying self-perceived health. We can do so because, in addition to the beginning of the year self-assessed health that is used in the model estimate, the survey data also contain a self-assessed health at the end of the year. Presumably each individual has the option to choose an insurance plan once during the year. Thus, if the residual association is even partially attributed to the detrimental health shock, we expect to observe a higher positive dependence in the varying health group than in the invariant health group. We recalibrate the Frank copula model and display the estimates of dependence parameters in the hurdle and conditional parts in Table 8 as well. For both groups we observe significant positive association for all types of care usage. When comparing the size, the dependence in the invariant health group in some cases is even slightly higher. These results suggest that exogenous health shocks, if they exist, are not the main cause for the residual relationship between self-perceived health and care utilization.

Though self-perceived health is not included as an explanatory variable in the equation of care utilization, its direct effect on utilization could be derived from the copula framework proposed. We examine the effect of self-perceived health from two perspectives: the effect on the probability of using medical care and the effect on the expenditures of medical care when it is consumed.

To illustrate the effect of self-perceived health risk on the likelihood of using medical care and subsequently on the expenditures given healthcare consumption, we present results in Fig. 3 and Table 9 for a hypothetical individual of average characteristics. The hypothetical individual is assumed to have the average characteristics of the entire sample. Fig. 3 shows the conditional density $f(y_{i1}|y_{i2}, y_{i1}>0, \bar{\mathbf{x}}_{i1}, \bar{\mathbf{x}}_{i2}, \hat{\theta}^{+})$ with $\bar{\mathbf{x}}_{i1}$ and $\bar{\mathbf{x}}_{i2}$ set to the value of the sample average and $\hat{\theta}^{+}$ estimated from the Frank copula. Figs 3(a), 3(b) and 3(c) correspond to expenditures related to office-based, hospital and ER visits respectively. Within a single panel, each curve corresponds to the density conditionally on a specific scale of self-assessed health. The conditional distribution of expenditures varies substantively across health status, especially for office and ER visits. It is important to note that, for all three types of medical care, the expenditure distribution for worse self-perceived health appears to have longer tails, which suggests a higher likelihood of incurring extreme expenses and thus a larger expected healthcare expenditure.

Table 9 displays the hurdle probability $\text{Prob}(y_{i1}>0|y_{i2}=j, \bar{\mathbf{x}}_{i1}, \bar{\mathbf{x}}_{i2}, \hat{\theta}^{0})$ and the conditional

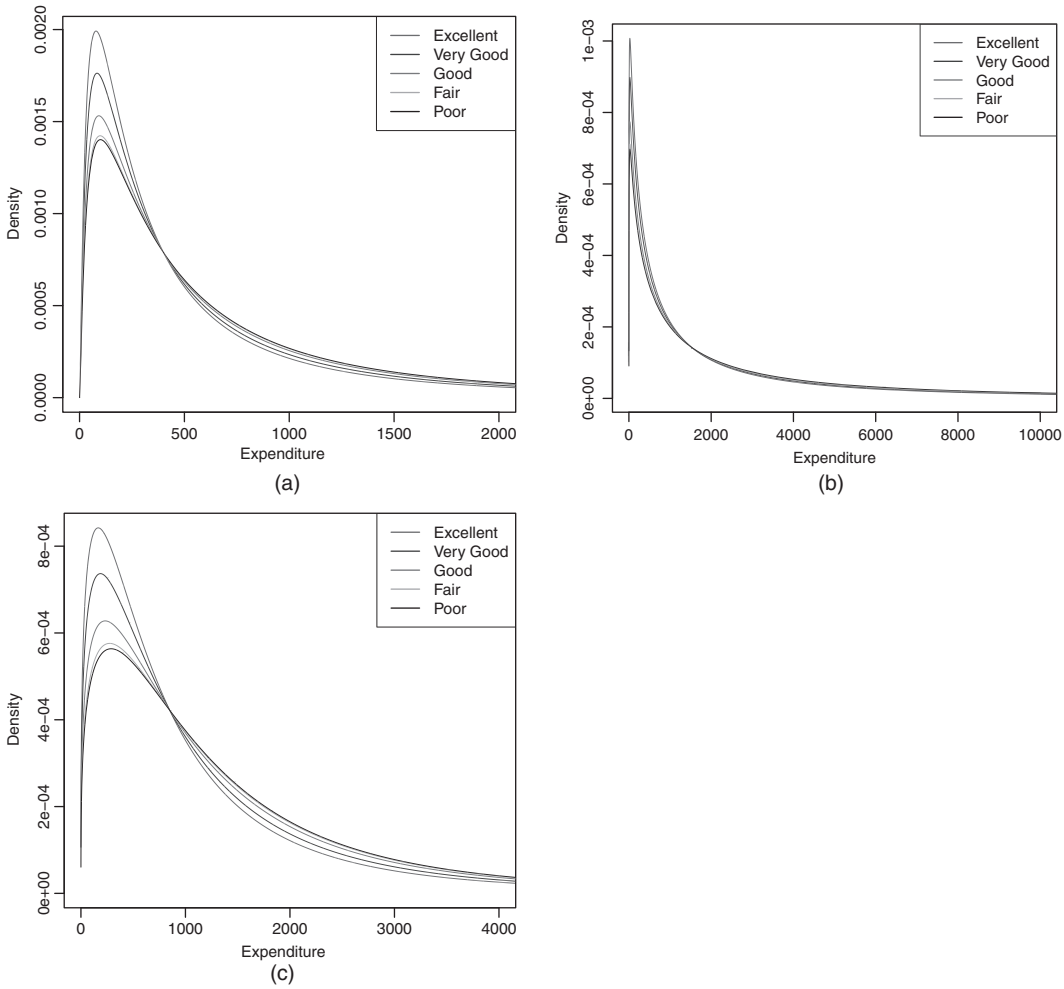

Fig. 3. Distribution of conditional healthcare expenditures by self-assessed health status: (a) office based; (b) hospital; (c) ER

mean $E(y_{i1}|y_{i2}=j, y_{i1}>0, \bar{\mathbf{x}}_{i1}, \bar{\mathbf{x}}_{i2}, \hat{\theta}^{+})$ regarding office-based, hospital and ER visits respectively. These quantities are calculated for an average individual for $j=1,\ldots,5$, i.e. various self-perceived health. Consistent with the positive dependence in both the hurdle and the conditional components, an individual with worse health status is more likely to obtain care and also is expected to have higher expenditures given the consumption of medical care. Such a positive relationship is found for all types of medical care service. Note that the analysis in Fig. 3 and Table 9 could be performed for other hypothetical individuals, such as a median individual. Though not reported here, we find a similar relationship between self-perceived health and care utilization drawn from other individuals and the results are robust to the specification of the copula.

We further calculate the partial effect $\Delta(j,k|\bar{\mathbf{x}}_i)$ according to equation (6) for every pair of self-perceived health statuses and the results are displayed in Table 10. The quantity is calculated for expenditures related to office-based, hospital and ER visits and the estimates with corresponding 95% confidence intervals are based on the Frank copula. A worse self-assessed health is in general

Table 10. Average treatment effects

Health comparison	*Results for the following types of service:*								
	Office based			*Hospital*			*ER*		
	PEA	*95% confidence interval*	*APE*	*PEA*	*95% confidence interval*	*APE*	*PEA*	*95% confidence interval*	*APE*
Excellent *versus* very good	110	[79,141]	111	184	[141,228]	205	28	[16,39]	34
Excellent *versus* good	235	[203,270]	230	454	[407,500]	455	72	[59,85]	75
Excellent *versus* fair	300	[267,336]	308	633	[586,685]	634	105	[91,119]	104
Excellent *versus* poor	314	[282,349]	334	679	[628,735]	704	116	[101,130]	117
Very good *versus* good	125	[89,164]	120	270	[223,320]	249	44	[30,57]	41
Very good *versus* fair	191	[156,228]	197	448	[399,499]	429	77	[62,92]	71
Very good *versus* poor	204	[170,242]	223	495	[439,549]	499	88	[73,102]	83
Good *versus* fair	66	[28,107]	77	178	[124,237]	179	34	[18,49]	30
Good *versus* poor	79	[39,120]	103	225	[166,290]	249	44	[28,59]	42
Fair *versus* poor	14	[−26,54]	26	46	[−14,111]	70	10	[−7,27]	12

related to a higher expected expenditure and such a relationship is significant for most pairs in our comparison. Note that we do not find a significant treatment effect between fair and poor health status. Another way to calculate the partial effect of self-perceived health is to evaluate it at the level of each individual and then to take the average, i.e.

$$\bar{\Delta}(j,k) = \frac{1}{n}\sum_{i=1}^{n} \Delta(j,k|\mathbf{x}_i).$$

We can think of the above equation as the average partial effect APE rather than a partial effect at the average. The estimate of $\overline{\Delta}(j,k)$ is also reported in Table 10. It is not surprising to observe the similarity between PEA and APE, because both quantities evaluate the change in healthcare utilization for an otherwise identical individual except for self-perceived health switching from j to k. As a result, we conclude that, after controlling for the available information indicated by the observed covariates, we find a strong positive relationship between medical care expenditures and self-perceived health status, suggesting that the private information that an individual holds regarding his or her health risk does play a critical role in healthcare utilization.

It is worth stressing the difference between the partial effect in Table 10 and the information in Section 2.2. For example, Table 1 suggests an average difference of $2595 in office-based care utilization between the poor and excellent health groups. This amount is about 10 times the partial effect that is reported in Table 10. The substantial difference is explained by the following fact: the results in Table 1 represent a simple average by clusters of self-perceived health status. The clustering process and thus the implied differences in care utilization are to a great extent attributed to the effects of observed covariates, which can be explained by the distinct composition of individual characteristics across different groups of self-perceived health status. This is supported by the statistical significance of most explanatory variables in Table 3. In contrast, the partial effect in Table 10 answers the question how much more or less healthcare would an individual have consumed if he or she had assessed his or her health status differently? Essentially, the partial effect measures the difference in care utilization for an otherwise equal individual with the effects of covariates purged off.

5. Summary and concluding remarks

Information asymmetry is an important source of inefficiency in the current healthcare delivery system and hence contributes to the high rising healthcare costs in the global market. This work studied a stylized form of information asymmetry—adverse selection, where patients privately informed about their poor health consume more care without paying actuarially fair premiums for health insurance. Setting us apart from most of the literature testing adverse selection, we examined directly the relationship between individuals' self-perceived health and healthcare expenditures. In doing so, we proposed a copula regression where the association parameter, accommodating the residual dependence between the two responses, reveals whether individuals hold private information regarding their health which affects the utilization of medical care services.

We demonstrated the flexibility of using copulas in multivariate dependence modelling that involves both discrete and mixed outcomes. In fact, self-perceived health status is an ordinal categorical variable, and healthcare expenditure is a semicontinuous variable, characterized by a fraction of 0s and skewed positive values. Moreover, the semicontinuous nature of healthcare expenditures motivated the two-part modelling framework that led to a very interesting interpretation of the results. Specifically, the copula in the hurdle component captures the relationship between self-assessed health and the probability of using care, and the other copula in the conditional component accommodates the effect of self-perceived health on expenditures given the consumption of care services.

When applying the copula-based hurdle model to a sample of the US civilian non-institutionalized population in the MEPS data, we considered expenditures related to three types of care: office based, hospital and ER visits. Regardless of the type of care, we found that, conditionally on the observable covariates, self-perceived health is positively related to both the likelihood of using healthcare and the amount of healthcare expenditures. Furthermore, our formulation allows us to quantify the effect of private health information on the utilization of healthcare services. We illustrated such effects by calculating the probability of using care and expected expenditures conditionally on a specific health status, and then the treatment effect for each pair of health statuses. Note that the standard hurdle model does not offer such easiness in the calculation of the partial effects of asymmetric health information.

Our analysis suggested a 10% positive residual association between self-perceived health and care utilization, which translates to a difference of over US $1000 care consumption per year between the highest and lowest risk categories. Note that this difference is obtained after risk adjustment based on the public information that is available to insurers. Thus, if an average individual is charged US $500 per month for health insurance, the above difference in care consumption amounts to about 20% of the yearly premium, suggesting significant cross-subsidization of high risk individuals by their low risk counterparts.

The existence of private health information that affects medical care utilization has important implications on the optimal payment design in healthcare systems. We showed that private health information is positively related to the utilization of medical care services. However, one limitation of this study is that we did not distinguish whether the positive relationship is caused by the risk effect or a 'health attitude' effect. The former effect refers to the situation where an individual in worse health knows his or her true health status and is expected to consume more medical care services, leading to the subsequent positive relationship between self-assessed health and healthcare expenditures. In the latter case, for instance, an individual could be pessimistic about his or her health and thus tends to seek and receive more care. Regardless of whether an individual's higher medical care consumption is a result of one's

accurate evaluation of his or her health or a biased evaluation, both are information that is held privately by the individual, and thus their effects on care utilization constitute the existence of adverse selection. We attribute our findings to a combined effect of these two mechanisms, but it would be worthwhile to distinguish the two effects because presumably the health attitude effect could be alleviated by consumer education, whereas the risk effect cannot.

Acknowledgements

We thank the Joint Editor and two reviewers for providing detailed suggestions and comments that helped to improve the quality of the paper greatly.

Appendix A: Joint distribution

From the relationship between the distribution function and copula function, we have

$$G^0(y_{i1} \leqslant 0, y_{i2} \leqslant j|\mathbf{x}_i) = G^0(y_{i1} = 0, y_{i2} \leqslant j|\mathbf{x}_i) = H^0\{G_1^0(0|\mathbf{x}_{i1}), G_2^0(j|\mathbf{x}_{i2}); \theta^0\}.$$

Thus, we can write equation (2) as

$$\begin{aligned} G^0(y_{i1} = 0, y_{i2} = j|\mathbf{x}_i) &= G^0(y_{i1} = 0, y_{i2} \leqslant j|\mathbf{x}_i) - G^0(y_{i1} = 0, y_{i2} \leqslant j-1|\mathbf{x}_i) \\ &= H^0\{G_1^0(0|\mathbf{x}_{i1}), G_2^0(j|\mathbf{x}_{i2}); \theta^0\} - H^0\{G_1^0(0|\mathbf{x}_{i1}), G_2^0(j-1|\mathbf{x}_{i2}); \theta^0\} \end{aligned}$$

and

$$\begin{aligned} G^0(y_{i1} > 0, y_{i2} = j|\mathbf{x}_i) &= \Pr(y_{i2} = j) - \Pr(y_{i1} = 0, y_{i2} = j) \\ &= G_2^0(j|\mathbf{x}_{i2}) - H^0\{G_1^0(0|\mathbf{x}_{i1}), G_2^0(j|\mathbf{x}_{i2}); \theta^0\} \\ &\quad - G_2^0(j-1|\mathbf{x}_{i2}) + H^0\{G_1^0(0|\mathbf{x}_{i1}), G_2^0(j-1|\mathbf{x}_{i2}); \theta^0\}. \end{aligned}$$

To show equation (3), we can start from the conditional joint distribution:

$$G^+(y_{i1}, y_{i2}|y_{i1} > 0, \mathbf{x}_i) = H^+\{G_1^+(y_{i1}|y_{i1} > 0, \mathbf{x}_{i1}), G_2^+(y_{i2}|y_{i1} > 0, \mathbf{x}_{i2}); \theta^+\}.$$

Then we have

$$\begin{aligned} g^+(y_{i1}, y_{i2}|y_{i1} > 0, \mathbf{x}_i) &= \frac{\partial}{\partial y_{i1}} \{G^+(y_{i1}, y_{i2}|y_{i1} > 0, \mathbf{x}_i) - G^+(y_{i1}, y_{i2} - 1|y_{i1} > 0, \mathbf{x}_i)\} \\ &= g_1^+(y_{i1}|y_{i1} > 0, \mathbf{x}_{i1})[h_1^+\{G_1^+(y_{i1}|y_{i1} > 0, \mathbf{x}_{i1}), G_2^+(y_{i2}|y_{i1} > 0, \mathbf{x}_{i2}); \theta^+\} \\ &\quad - h_1^+\{G_1^+(y_{i1}|y_{i1} > 0, \mathbf{x}_{i1}), G_2^+(y_{i2} - 1|y_{i1} > 0, \mathbf{x}_{i2}); \theta^+\}], \end{aligned}$$

Appendix B: GB2 regression

The generalized beta distribution of the second kind, GB2, can be constructed from gamma distributions. Let $Z_1 \sim \text{gamma}(\phi_1, 1)$ and $Z_2 \sim \text{gamma}(\phi_2, 1)$; then random variable $Y = \beta(Z_1/Z_2)^{1/\alpha}$ is known to follow the GB2 distribution. The GB2 distribution is a four-parameter distributional family with density function

$$f_{\text{GB2}}(y; \alpha, \beta, \phi_1, \phi_2) = \frac{|\alpha| y^{\alpha\phi_1 - 1} \beta^{\alpha\phi_2}}{B(\phi_1, \phi_2)(\beta^\alpha + y^\alpha)^{\phi_1 + \phi_2}}, \qquad y > 0. \tag{7}$$

McDonald and Butler (1990) first employed GB2 regression to investigate the duration of welfare spells. Recently, Sun *et al*. (2008) applied the GB2 distribution in the prediction of nursing home utilization with longitudinal data. Frees and Valdez (2008) used the GB2 distribution to capture the long-tail nature of automobile insurance claims in a hierarchical insurance claims model.

To facilitate regression analysis, we consider the parameterizations $\alpha = \sigma^{-1}$ and $\beta = \exp(\mu)$. Then we have

$$f_{\text{GB2}}(y; \mu, \sigma, \phi_1, \phi_2) = \frac{\exp(\phi_1 z)}{y|\sigma| B(\phi_1, \phi_2)\{1 + \exp(z)\}^{\phi_1 + \phi_2}}, \qquad y > 0, \tag{8}$$

where $z = \{\ln(y) - \mu\}/\sigma$ and $B(\phi_1, \phi_2)$ is the Euler beta function. The GB2 distribution is a member of the log-location–scale family, in which μ is the location parameter, σ is the scale parameter and ϕ_1 and ϕ_2 are shape parameters. It follows that

$$E(Y^m) = \exp(k\mu)\frac{B(\phi_1 + m\sigma, \phi_2 - m\sigma)}{B(\phi_1, \phi_2)}, \qquad -\phi_1 < m\sigma < \phi_2. \tag{9}$$

Thus, if we incorporate covariates in the location $\mu = \mathbf{x}'\boldsymbol{\gamma}_1$, the regression coefficient could be interpreted as the proportional change in the response when the explanatory variable increases by 1 unit. In general, one could also specify $\sigma = \exp(\mathbf{x}'\boldsymbol{\gamma}_2)$ to account for heteroscedasticity, though it is not normally seen in applied studies.

With four parameters, the GB2 distribution provides great flexibility for modelling heavy-tailed and skewed data. Many commonly used skewed distributions are nested as special or limiting cases of the GB2 distribution, such as the standard gamma, Weibull and log-normal distributions (see McDonald and Xu (1995)). One special case that is worth pointing out is the generalized gamma distribution that is derived from the GB2 distribution when the shape parameter ϕ_2 approaches ∞. The generalized gamma distribution has been widely used in the regression context in the health economics literature to describe skewed medical costs; for example, see Manning *et al.* (2005). Another related distribution is the exponential GB2 or EGB2 distribution that is defined on the whole real line. According to McDonald (1984), if $Y \sim \mathrm{GB2}(y; \mu, \sigma, \phi_1, \phi_2)$, then $\ln(Y) \sim \mathrm{EGB2}(y; \mu, \sigma, \phi_1, \phi_2)$.

Appendix C: Copulas and partial density

We employ three copulas in this study: the Farlie–Gumbel–Morgenstern copula, the Frank copula and the Gaussian copula. The functional forms of these copulas are respectively as follows:

$$H(u_1, u_2; \theta) = u_1 u_2\{1 + \theta(1 - u_1)(1 - u_2)\}, \qquad |\theta| \leqslant 1,$$

$$H(u_1, u_2; \theta) = \frac{1}{\theta}\log\left[1 + \frac{\{\exp(-\theta u_1) - 1\}\{\exp(-\theta u_2) - 1\}}{\exp(-\theta) - 1}\right], \qquad -\infty < \theta < \infty,$$

$$H(u_1, u_2; \theta) = \int_{-\infty}^{\Phi^{-1}(u_1)} \int_{-\infty}^{\Phi^{-1}(u_2)} \{2\pi(1 - \theta^2)\}^{-1} \exp\left\{-\frac{s^2 - 2\theta st + t^2}{2(1 - \theta^2)}\right\} \mathrm{d}s\,\mathrm{d}t, \qquad |\theta| < 1.$$

It is easy to show that, when $\theta \to 0$, these copulas yield the product copula $H(u_1, u_2) = u_1 u_2$, indicating the case of independence. Furthermore, a positive θ suggests a positive dependence and vice versa. However, a comparison of θ across copulas is not straightforward owing to the different supports. To address this issue, we can translate the dependence parameter in a copula into some commonly used non-linear association measure, such as Spearman's ρ or Kendall's τ. In addition, the Gaussian copula attains both upper and lower Fréchet–Hoeffding bounds as $\theta \to 1$ and $\theta \to -1$ respectively. Though the dependence is weaker in both tails when compared with the Gaussian copula, the Frank copula also acquires the Fréchet–Hoeffding upper and lower bounds, as $\theta \to \infty$ and $\theta \to -\infty$ respectively. Despite its simplicity, the Farlie–Gumbel–Morgenstern copula is more restrictive in modelling dependence in that neither Fréchet–Hoeffding bound is attained when θ is on the boundaries.

Since our application involves joint modelling of continuous and discrete outcomes, the complete specification of the above model also requires the corresponding partial copulas, i.e. $h_1(u_1, u_2; \theta)$. Simple calculations show that the Farlie–Gumbel–Morgenstern, Frank and Gaussian copulas are respectively

$$h_1(u_1, u_2; \theta) = u_2\{1 + \theta(1 - u_1)(1 - u_2)\} - \theta u_1 u_2(1 - u_2), \qquad |\theta| \leqslant 1,$$

$$h_1(u_1, u_2; \theta) = \frac{\exp(-\theta u_1)\{\exp(-\theta u_2) - 1\}}{\exp(-\theta) - 1 + \{\exp(-\theta u_1) - 1\}\{\exp(-\theta u_2) - 1\}}, \qquad -\infty < \theta < \infty,$$

$$h_1(u_1, u_2; \theta) = \Phi\left\{\frac{\Phi^{-1}(u_2) - \theta\Phi^{-1}(u_1)}{\sqrt{(1 - \theta^2)}}\right\}, \qquad |\theta| < 1.$$

References

Akerlof, G. (1970) The market for "lemons": quality uncertainty and the market mechanism. *Q. J. Econ.*, **84**, 488–500.

Chen, Y.-H. (2010) Semiparametric marginal regression analysis for dependent competing risks under an assumed copula. *J. R. Statist. Soc.* B, **72**, 235–251.

Cummins, D., Dionne, G., McDonald, J. and Pritchett, M. (1990) Applications of the GB2 family of distributions in the modeling insurance loss processes. *Insur. Math. Econ.*, **9**, 257–272.

Cutler, D. and Zeckhauser, R. (2000) The anatomy of health insurance. In *Handbook of Health Economics*, vol. 1 (eds. A. Culyer and J. Newhouse), pp. 563–643. New York: Elsevier.

Deb, P., Li, C., Trivedi, P. and Zimmer, D. (2006) The effect of managed care on use of health care services: results from two contemporaneous household surveys. *Hlth Econ.*, **15**, 743–760.

Dionne, G., Michaud, P.-C. and Dahchour, M. (2013) Separating moral hazard from adverse selection and learning in automobile insurance: longitudinal evidence from France. *J. Eur. Econ. Ass.*, **11**, 897–917.

Dionne, G., Pinquet, J., Maurice, M. and Vanasse, C. (2011) Incentive mechanisms for safe driving: a comparative analysis with dynamic data. *Rev. Econ. Statist.*, **93**, 218–227.

Doiron, D., Jones, G. and Savage, E. (2008) Healthy, wealthy and insured?: the role of self-assessed health in the demand for private health insurance. *Hlth Econ.*, **17**, 317–334.

Dowd, B., Feldman, R., Cassou, S. and Finch, M. (1991) Health plan choice and the utilization of health care services. *Rev. Econ. Statist.*, **73**, 85–93.

Frees, E. and Valdez, E. (2008) Hierarchical insurance claims modeling. *J. Am. Statist. Ass.*, **103**, 1457–1469.

Getzen, T. E. (2000) Health care is an individual necessity and a national luxury: applying multilevel decision models to the analysis of health care expenditures. *J. Hlth Econ.*, **19**, 259–270.

Goldman, D., Hosek, S., Dixon, L. and Sloss, E. (1995) The effects of benefit design and managed care on health care costs. *J. Hlth Econ.*, **14**, 401–418.

Holly, A., Gardiol, L., Domenighetti, G. and Bisig, B. (1998) An econometric model of health care utilization and health insurance in Switzerland. *Eur. Econ. Rev.*, **42**, 513–522.

Joe, H. (1997) *Multivariate Models and Dependence Concepts*. New York: Chapman and Hall.

Jürges, H. (2007) True health vs response styles: exploring cross-country differences in self-reported health. *Hlth Econ.*, **16**, 163–178.

Kenkel, D. S. (1994) The demand for preventive medical care. *Appl. Econ.*, **26**, 313–325.

Lawless, J. (2003) *Statistical Models and Methods for Lifetime Data*. New York: Wiley-Interscience.

Liu, L., Strawderman, R., Cowen, M. and Shih, Y. (2010) A flexible two-part random effects model for correlated medical costs. *J. Hlth Econ.*, **29**, 110–123.

Lo, S. M. S. and Wilke, R. A. (2010) A copula model for dependent competing risks. *Appl. Statist.*, **59**, 359–376.

Madsen, L. and Fang, Y. (2011) Joint regression analysis for discrete longitudinal data. *Biometrics*, **67**, 1171–1175.

Manning, W., Basu, A. and Mullahy, J. (2005) Generalized modeling approaches to risk adjustment of skewed outcomes data. *J. Hlth Econ.*, **24**, 465–488.

Manning, W., Newhouse, J. and Ware, J. (1982) The status of health in demand estimation; or, beyond excellent, good, fair, and poor. In *Economic Aspects of Health* (ed. V. Fuchs), pp. 143–184. Chicago: University of Chicago Press.

McDonald, J. (1984) Some generalized functions for the size distribution of income. *Econometrica*, **52**, 647–663.

McDonald, J. (1987) Model selection: some generalized distributions. *Communs Statist. Theor. Meth.*, **16**, 1049–1074.

McDonald, J. and Butler, R. (1987) Some generalized mixture distributions with an application to unemployment duration. *Rev. Econ. Statist.*, **69**, 232–240.

McDonald, J. and Butler, R. (1990) Regression models for positive random variables. *J. Econmetr.*, **43**, 227–251.

McDonald, J. and Xu, Y. (1995) A generalization of the beta distribution with applications. *J. Econmetr.*, **66**, 133–152.

Nelsen, R. (2006) *An Introduction to Copulas*. New York: Springer.

Riphahn, R., Wambach, A. and Million, A. (2003) Incentive effects in the demand for health care: a bivariate panel count data estimation. *J. Appl. Econmetr.*, **18**, 387–405.

Rothschild, M. and Stiglitz, J. (1976) Equilibrium in competitive insurance markets: an essay on the economics of imperfect information. *Q. J. Econ.*, **90**, 629–649.

Shi, P. (2012) Multivariate longitudinal modeling of insurance company expenses. *Insur. Math. Econ.*, **51**, 204–215.

Shi, P., Zhang, W. and Valdez, E. (2012) Testing adverse selection with two-dimensional information: evidence from the Singapore auto insurance market. *J. Risk Insur.*, **79**, 1077–1114.

Smith, M. (2003) Modelling sample selection using archimedean copulas. *Econmetr. J.*, **6**, 99–123.

Song, P. X.-K., Li, M. and Yuan, Y. (2009) Joint regression analysis of correlated data using gaussian copulas. *Biometrics*, **65**, 60–68.

Sun, J., Frees, E. and Rosenberg, M. (2008) Heavy-tailed longitudinal data modeling using copulas. *Insur. Math. Econ.*, **42**, 817–830.

Trivedi, P. K. and Zimmer, D. M. (2007) *Copula Modeling: an Introduction for Practitioners*. Boston: Now.

Van Ourti, T. (2004) Measuring horizontal inequity in Belgian health care using a Gaussian random effects two part count data model. *Hlth Econ.*, **13**, 705–724.
Vera-Hernández, A. (1999) Duplicate coverage and demand for health care: the case of Catalonia. *Hlth Econ.*, **8**, 579–598.
Windmeijer, F. and Santos Silva, J. (1997) Endogeneity in count data models: an application to demand for health care. *J. Appl. Econmetr.*, **12**, 281–294.
Zimmer, D. and Trivedi, P. (2006) Using trivariate copulas to model sample selection and treatment effects: application to family health care demand. *J. Bus. Econ. Statist.*, **24**, 63–76.