Hindawi
Discrete Dynamics in Nature and Society
Volume 2021, Article ID 4878442, 13 pages
https://doi.org/10.1155/2021/4878442

Research Article

Modeling Hospitalization Decision and Utilization for the Elderly in China

Xin Xu and Dongxiao Chu

School of Finance, Capital University of Economics and Business, Beijing 100070, China

Correspondence should be addressed to Dongxiao Chu; chudongxiao@cueb.edu.cn

Received 15 May 2021; Revised 16 June 2021; Accepted 21 June 2021; Published 1 July 2021

Academic Editor: Daqing Gong

Getting medical services has become more difficult and expensive in China, which led to a problem of illness not being treated and a large number of zeros in the statistics of being hospitalized for the elderly. Traditional classic models such as the Poisson model and the negative binomial model cannot fit this kind of data well. One aim of this study was to use zero-inflated and hurdle models to better solve the problem of excess zeros. Another aim was to discover the factors affecting the decision-making behavior of the elderly being hospitalized and hospitalization service utilization. Therefore, the XGBoost model was firstly introduced to rank the importance of influencing factors in this paper. It was found that the zero-inflated negative binomial model performed best. The results showed that the elderly who had enjoyed NRCM or ERBMI/URBMI were more likely to have a higher number of hospitalizations. This indicated that the high cost of hospitalization had prevented the willingness of the elderly being hospitalized, but the basic medical insurance had increased the times of their repeated hospital readmissions. Policy efforts should be made to improve the level of basic medical insurance.

1. Introduction

Population aging is the trend of economic and social development in China. There were approximately 176 million Chinese people aged 65 and above in 2019, accounting for 12.6% of the total population. According to the United Nations' standard of whether the proportion of 65-year-old and upper-aged population in the total population is more than 14%, China has officially entered the "aging society" and begun the countdown stage. Compared with other people, the elderly have the characteristics of a high prevalence rate, more chronic and serious diseases, the heavy burden of medical expenses, and so on, which might lead to more medical needs and more hospitalization services [1, 2].

Previous studies mainly focused on the demands and utilization of health services [3], the equality of health services, fall-risk-increasing drugs and falls by the elderly [4], the impacts of health insurances on outpatient visits of older adults, and so on [5–10]. Older adults are more likely to suffer from diseases and need hospitalizations more urgently [3], resulting in higher direct and indirect medical care costs. Government healthcare finance policy has affected the utilization of health services in the United States, Korea, Vietnam, Singapore, and China [11–16]. However, the problem of the affordability of healthcare seems not to be mitigated by the development of social health insurance (SHI), even though such schemes now cover almost the whole population in China. A Chinese survey shows that public complaints about the problems of healthcare reform and affordability in urban areas increased from 21.1% in 2007 to 34.8% in 2009. Therefore, with the exponential increase in medical costs, some elderly do not have effective access to medical services due to economic poverty, lack of medical security, poor medical accessibility, and other factors, and they had to give up hospitalization or treatment due to these reasons, which seriously affects their health. However, few studies explored the hospitalization decision-making needs and influencing factors of the elderly. Therefore, the present study aimed to explore how demographic, socioeconomic, and health insurance factors would impact the whole hospitalization decision-making process of the elderly under the current healthcare system.

The hospitalization decision of the elderly consists of two parts: whether to be hospitalized or not and how many times to be hospitalized. There were many statistical models for the count data to analyze the factors of hospitalization, such as Poisson and negative binomial models [17]. However, modeling the hospitalization decision-making is more challenging because of healthcare data. This type of data commonly presents the problems of overdispersed and excessive zeros [17, 18]. The counts of hospitalizations for the elderly in China had suffered a serious zero-inflated (Figure 1) problem because the elderly gave up hospitalizations, which led to the observed sample variance larger than the sample mean. The standard Poisson model that the mean and variance of a count response variable are equal is not fit. Ignoring the overdispersion in count data would result in an underestimation of the standard errors, an overestimation of parameter significance level, and a biased hypothesis testing [19, 20]. The negative binomial model characterized as equal mean and variance could deal with the overdispersion caused by heterogeneity in the count data, but it cannot effectively solve the zero-inflated problem [19, 20]. Two-part models such as hurdle models had been widely used for handling counts data with excessive zeros, which allowed a logistic or probit regression modeling the probability that a count is zero or positive integer value and a generalized linear regression for the observer healthcare utilization [21, 22]. Another way to deal with excessive zeros is zero-inflated models. A zero-inflated model is a mixture of regular count models such as Poisson or negative binomial model and a component that accommodates the excessive zeros [19, 20, 23]. These models had been applied, modified, and extended as very popular models for healthcare data [24–29].

Previous studies paid more attention to the utilization of medical services by certain factors such as new rural cooperative medical insurance and urban workers' medical insurance [11, 13], ignoring the analysis of the needs of medical services at the individual level, especially for the hospitalization needs of the elderly. Also, in the research methods, several studies mainly used Poisson, negative binomial, zero expansion, and hurdle models to predict the number of outpatients; few studies used these models to analyze the number of hospitalizations of the elderly. The goal of this study is not only to predict the number of inpatients for the elderly using these models but also systematically to analyze the decision-making and demand factors of hospitalizations of the elderly from the aspects of personal medical needs, economic status, demographic characteristics, medical security, and so on. Many studies have either isolated each influencing factor independently or studied the cross-influencing factors of two factors, lacking a comparison of the importance of multiple variables. Based on the analysis of the factors that affect the number of hospitalizations for the elderly, this paper used the XGBoost (eXtreme gradient boosting) model in machine learning to explore the importance of these factors [30]. XGBoost is an integrated learning model with excellent performance for many classification and regression problems, which is an improvement and expansion of the gradient boosting decision tree (GBDT) model. The advantage of the XGBoost model is that it can prevent overfitting through regularization terms. It uses not only the first derivative but also the second derivative, whose loss function can be customized and the loss accuracy is improved. Although machine learning methods have been applied to many other fields, there are relatively few studies in healthcare, especially the work using the XGBoost method to predict the number of hospitalizations. Due to the available hospitalization count data being strongly right-skewed with excessive zeros, the study attempted to compare different statistical models to obtain an optimal model to predict the number of elderly hospitalizations more accurately.

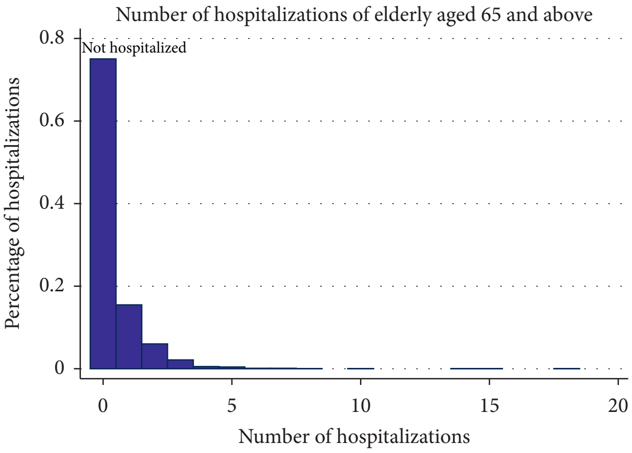

Figure 1: The number of hospitalizations.

2. Materials and Methods

2.1. Data Source and Study Population. The study used the microdata from the Chinese Longitudinal Healthy Longevity Survey (CLHLS). The survey was jointly performed by the Center for Healthy Aging and Development Studies of the National School of Development at Peking University, which is the earliest and longest lasting social science survey nationwide in China. The survey covers 23 provinces, municipalities, and autonomous regions across China. The subjects of the survey are elderly people aged 65 and over. The survey content of the questionnaire for the surviving respondents included the basic status of the elderly and their families, socioeconomic background and family structure, economic sources and economic status, self-evaluation of health and quality of life, cognitive function, personality and psychological characteristics, daily activity ability, lifestyle, life care, disease treatment, and medical expenses.

The CLHLS conducted a baseline survey in 1998, followed by seven-wave surveys in 2000, 2002, 2005, 2008–2009, 2011–2012, 2014, and 2017–2018 in randomly selected about half of the counties and city districts in 23 Chinese provinces. The CLHLS aimed to understand the demographic characteristics, lifestyles, health services, behavioral, economic status, and so on, among the Chinese elderly including ages 65 years old and over. Detailed information about the survey design and assessment of data quality has been reported in previous studies [31–33]. The participants of the CLHLS baseline survey were older adults aged 80 years and over, and the age range was adjusted to 65 years and over after 2002.

We use CLHLS data from the latest follow-up cross-sectional survey (2017–2018) for the surviving, including 10 participants since 1998, 30 participants since 2000, 1,330 participants since 2000, 2,440 participants since 2008 and 2009, 2,884 participants since 2011/2012, 3,463 participants in 2014, and 12,411 participants for the first time in 2018. After filtering for missing and invalid values and outliers, 5,287 participants with complete information on healthcare utilization, sociodemographic, and economic characteristics were included in the analyses of impact factors on hospitalization.

2.2. Covariate Selection. The Andersen Behavioral Model of Health Service Use provides a framework for the study of hospitalization that outlines the three determinants: predisposing, enabling, and need factors [34]. In light of this, we evaluated the effects of health status and functional disabilities, as need factors and associated sociodemographic factors, as predisposing and enabling factors, on hospitalization utilization (Table 1).

2.2.1. Need Factors. This study integrated the health status and functional disabilities as need factors [35]. The health status was evaluated by self-rated health that was a multicategorized variable in order in which “1” to “5,” represented “very bad” to “very well.” The functional disabilities were measured by activities of daily living (ADLs). ADLs requiring any assistance were defined as “with difficulties.” The measure of ADLs was assessed as “no difficulty” or “with difficulties.”

2.2.2. Predisposing and Enabling Factors. According to Andersen’s behavior model, we evaluated sociodemographic characteristics associated with predisposing and enabling factors in this study. The predisposing factors included age, gender (male = 1 and female = 0), marital status (not in marriage = 0 and in marriage = 1), years of education, smoking (smoked in the past = 1 and not smoking = 0), and alcohol (drunk in the past = 1 and not drinking = 0). The enabling factors included total income of individual’s household last year, hospitalization expenditure last year, out-of-pocket expenses for hospitalization, and medical security plans: Urban Employee Basic Medical Insurance (UEBMI), Urban Resident Basic Medical Insurance (URBMI), and New Rural Cooperative Medical Scheme (NRCMS), free medical treatment, and others [34, 36]. Although the data about family income, hospitalization expenses, and out-of-pocket expenses in the survey were right-censored, it did not affect the conclusion because this kind of expenses exceeding 100,000 accounted for a relatively small proportion.

2.2.3. Description of Covariates. One of the outcomes of this study was the number of hospitalizations among the older patients. The CLHLS asked the participants how many times they suffered from a serious illness that required hospitalization or were bedridden at home in the past two years. We took the number of hospitalizations of the sick elderly in the past two years as a dependent variable and regarded the response too seriously ill but “bedridden at home” in the past two years in the questionnaire as not be hospitalized. The response variable was a type of discrete integer numerical variable.

Descriptive analyses were conducted to examine the outcome, demographic, and socioeconomic characteristics of the 5,287 participants (Table 1). Figure 1 shows that the distribution of the number of hospitalizations is heavily right-skewed and the variance is greater than the mean, which could be deemed as overdispersed data. Meanwhile, there are a large number of zero counts presented in Figure 1.

2.3. Statistical Models. The number of hospitalizations was assumed as an overdispersed and zero-inflated count variable; either Poisson or negative model might fit the data well. To accommodate the excess zeros, we utilized hurdle and zero-inflated models to fit appropriately. The difference between the zero-inflated and the hurdle models is that zero observations come from “structure” and “sampling” in zero-inflated models, nevertheless zeros were from one “sampling” source in hurdle models. Given the characters of our data, we applied and compared the classical count regression models such as Poisson and negative binomial (NB), zero-inflated Poisson (ZIP), zero-inflated negative binomial (ZINB), hurdle Poisson (HP), and hurdle negative binomial (HNB).

2.3.1. Poisson Model. Poisson regression is usually a benchmark model to fit count. Within the study, we assume that the number of hospitalizations of the elderly (Y_i) obey the Poisson distribution with parameter (λ_i), and the probability function is defined as [19, 20]

$$P(Y_i = y_i|X) = \frac{\exp(-\lambda_i)\lambda_i^{y_i}}{\Gamma(1+y_i)}, \quad y_i = 0, 1, 2, \ldots, \tag{1}$$

where X is the factors, $E(Y_i|X) = \lambda_i = \exp(X_i\beta)$, and β are the estimated parameters. It was assumed that the mean and the variance for Poisson are equal (e.g.,$\mathrm{Var}(Y_i|X) = E(Y_i|X)$).

2.3.2. Negative Binomial Model. The negative binomial distribution is another method alternative to the Poisson model when the data are heterogeneous [19, 20]. The Poisson distribution assumes that the older patients are homogeneous, that is, the mean can be regarded as a fixed value, which does not match the fact. So if the average number of hospitalizations in Poisson is regarded as a random variable distributed to a gamma distribution, we can find the negative binomial model as follows:

Table 1: Description of variables.

	Variable and type	Definition	Mean	SD	Skewness	Kurtosis
Response	Number of hospitalizations (count variable)	Number of hospitalizations in the past two years	0.4135651	0.9690706	5.341749	59.89688
Need factors	Self-rated health (category variable)	Very bad or bad = 0 (reference); general = 1; good or very good = 2	1.467221	0.5201101	−0.0986902	1.542248
	ADL difficulties (category variable)	Unrestricted = 0 (reference); restricted = 1; very limited = 2	0.4175326	0.6554114	1.298326	3.421162
Predisposing factors	Age (continuous variable)	65–117 years old	84.08502	11.69399	0.1188886	1.886834
	Gender (category variable)	Male = 1; female = 0 (reference)	0.4445494	0.4969627	0.2242063	1.050268
	Marital status (category variable)	In marriage = 1; not in marriage = 0 (reference)	0.4549405	0.4980125	0.1811815	1.032827
	Years of education (continuous variable)	0–20 years	3.612507	4.378844	1.116702	3.493054
	Smoking or not (category variable)	Smoked in the past = 1; not smoking = 0 (reference)	0.3121103	0.4633984	0.8111184	1.657913
	Drinking or not (category variable)	Drunk in the past = 1; not drinking = 0 (reference)	0.2597771	0.4385536	1.097245	2.203946
Enabling factors	Household income (continuous variable)	0–99 (thousands of yuan)	43.40865	36.74898	0.4600004	1.676671
	Hospitalization medical expenditure (continuous variable)	0–99 (thousands of yuan)	4.456139	13.51289	4.847745	29.40199
	Out-of-pocket expenses (continuous variable)	0–99 (thousands of yuan)	2.347067	8.9899	7.401468	68.09451
	Medical insurance (category variable)	NRCMS = 1; UEBMI/URBMI = 2; free medical treatment = 3; others = 0 (reference)	1.218591	0.6759684	0.4205635	3.35609

Note. In the questionnaire, the family income, hospitalization expenses, and out-of-pocket expenses of hospitalization exceeding 100,000 yuan are recorded as 99,998. For the convenience of calculation, this study takes 1,000 yuan as the unit.

$$P[Y = y_i|X] = \frac{\Gamma(\sigma^{-2} + y_i)}{\Gamma(\sigma^{-2})\Gamma(1 + y_i)}\left(\frac{\sigma^{-2}}{\lambda_i + \sigma^{-2}}\right)^{\sigma^{-2}}\left(\frac{\lambda_i}{\lambda_i + \sigma^{-2}}\right)^{y_i}, \quad y_i = 0, 1, 2, \ldots; \sigma^2 > 0, \tag{2}$$

with the mean $E[Y_i|X] = \lambda_i$ and variance function $\mathrm{Var}[Y_i|X] = \lambda_i + \sigma^2\lambda_i^2$, where σ^2 is known as the dispersion parameter. We can find that the variance is a quadratic function of the mean and is greater than the mean, which may solve the problem of heterogeneity.

2.3.3. Zero-Inflated Models. Figure 1 shows that there is a disproportionately large frequency of zeros in the number of inpatients that leads to poor performances of Poisson and negative binomial models. We promise a method to overcome the problem using zero-inflated models. The zeros in zero-inflated models came from two components: one part arising from a parent distribution and the other corresponds to the excessive zeros that could not be accounted for by the distribution [37]. The zero-inflated model is described as follows [19, 38]:

$$P[Y = y_i|X] = \begin{cases} \pi + (1 - \pi) \times f(y_i|x_i), & \text{if } y_i = 0, \\ (1 - \pi) \times f(y_i|x_i), & \text{if } y_i > 0, \end{cases} \tag{3}$$

where $f(\cdot)$ is a general count model. The mean and variance of the zero-inflated model are defined as follows:

$$\begin{aligned} E[Y_i|X] &= (1 - \pi) \times E[\cdot]_f, \\ \mathrm{Var}[Y_i|X] &= (1 - \pi)\left(\mathrm{Var}[\cdot]_f^2 + \pi E[\cdot]_f^2\right), \end{aligned} \tag{4}$$

where $E[\cdot]_f$ and $\mathrm{Var}[\cdot]_f$ are the mean and variance, respectively, for the count model. When the distribution of f is a Poisson, we can define the zero-inflated Poisson (ZIP) model:

$$\Pr[Y = y_i|X] = \begin{cases} \pi + (1 - \pi) \times \exp(-\lambda_i), & y_i = 0, \\ (1 - \pi) \times \dfrac{\exp(-\lambda_i)\lambda_i^{y_i}}{\Gamma(1 + y_i)}, & y_i > 0, \end{cases} \tag{5}$$

and what's more, the zero-inflated negative binomial (ZINB) model is described as follows:

$$P[Y = y_i|X] = \begin{cases} \pi + (1-\pi) \times \left(\dfrac{\sigma^{-2}}{\lambda_i + \sigma^{-2}}\right)^{\sigma^{-2}}, & y_i = 0, \\ (1-\pi) \times \dfrac{\Gamma(\sigma^{-2} + y_i)}{\Gamma(\sigma^{-2})\Gamma(1+y_i)} \left(\dfrac{\sigma^{-2}}{\lambda_i + \sigma^{-2}}\right)^{\sigma^{-2}} \left(\dfrac{\lambda_i}{\lambda_i + \sigma^{-2}}\right)^{y_i}, & y_i > 0. \end{cases} \quad (6)$$

2.3.4. Hurdle Model. In addition, except for the aforementioned zero-inflated modes, the hurdle model is a widely used alternative for the count data with excessive zeros. The hurdle reflects a two-part decision-making process. The elder patient decides whether to be hospitalized or not firstly and then makes a second decision about how many times for inpatients. Therefore, the zeros are determined by one density $f_1(\cdot)$, so that $P[Y_i = 0|x_i] = f_1(0)$, and while the positive counts are from another density $f_2(\cdot)$. This leads to the hurdle model

$$P[Y_i = k|X] = \begin{cases} f_1(0) & \text{if} \quad k = 0, \\ \dfrac{1 - f_1(0)}{1 - f_2(0)} \times f_2(\cdot) & \text{if} \quad k > 0. \end{cases} \quad (7)$$

The model collapses to the standard count model only if $f_1(0) = f_2(0)$. The density $f_2(\cdot)$ is a count density such as Poisson or negative binomial model, whereas $f_1(\cdot)$ could also be a count data density, or more simply, the probabilities $f_1(0)$ and $1 - f_1(0)$ may be estimated using a logit or probit model [19, 20]. Although there is much literature using the probit model in the first part of the hurdle model, there is no obvious evidence that the choice of probit model or logit model has a serious impact on the results, and there is much work preferring to choosing the logit model mainly because its explanation and calculation are more convenient [21, 25, 26, 28, 37, 38]. In fact, there is also no evidence showing that the conclusions of the logit model are better or more reliable than the probit model in our study. The mean of the hurdle model is determined by the probability of crossing the threshold and by the moments of the zero-truncated density as follows:

$$E[Y_i|X] = \frac{1 - f_1(0|X)}{1 - f_2(0|X)} \times \mu_2(X), \quad (8)$$

where $\mu_2(X)$ is the untruncated mean in density $f_2(y|x)$. The hurdle model variance is shown in the following equation:

$$\mathrm{Var}[Y_i|X] = \frac{1 - f_1(0|X)}{1 - f_2(0|X)} \times \sigma_2^2 + \frac{(1 - f_1(0|X)) \cdot (f_1(0|X) - f_2(0|X))}{(1 - f_2(0|X))^2} \times (\mu_2(X))^2, \quad (9)$$

where $\sigma_2^2(X)$ is the untruncated variance in density $f_2(y|X)$.

2.4. XGBoost (eXtreme Gradient Boosting) Model. The XGBoost model is a collection of a series of decision trees [30]. It is a type of boosted tree model (boosted trees). The decision tree in the model does not independently make predictions on the input samples but based on the prediction results of the previous round of the model. Learning the error of the prediction improves the prediction accuracy of the model. Let $\widehat{y}_i^{(t)}$ denote the preresult of the model for the i-th sample after the t-th iteration and $f_t(x_i)$ denote the prediction score of the t-th decision tree for the i-th sample, then the expression of $\widehat{y}_i^{(t)}$ is as follows:

$$\widehat{y}_i^{(t)} = \sum_{k=1}^{t} f_k(x_i) = \widehat{y}_i^{(t-1)} + f_t(x_i). \quad (10)$$

At the t-th iteration, the prediction result $\widehat{y}_i^{(t-1)}$ of the previous $(t-1)$ times of the model has been given, so the training goal of the model is to select the appropriate prediction function f_t as the minimized objective function. The meanings of $\widehat{y}_i^{(t)}$ and x_i are the same as above, which represent the number of hospitalizations and influencing factors, respectively. Assuming that $L(y_i, \widehat{y}_i)$ is the loss function that measured the degree of deviation between the predicted category of the sample and the true category, the objective function of the XGBoost model training at the t-th iteration has the following expression:

$$O^{(t)} = \sum_{i=1}^{n} L\left(y_i, \widehat{y}_i^{(t-1)} + f_t(x_i)\right) + \Omega(f_t) + \text{constant}. \quad (11)$$

The first part of the objective function measures the loss of the model's prediction errors for all samples, and the second term (Ω) is a regularization term, which is a penalty for the complexity of the newly added model. Adding a regularization term to the objective function is helpful to prevent the model from overfitting, and it is also an improvement of XGBoost over the GBDT model. For the decision tree model f_t, its complexity can be measured by the following formula:

$$\Omega\left(f_t\right) = \gamma T + 0.5\lambda \sum_{j=1}^{T} \varpi_j^2, \tag{12}$$

where T represents the number of leaf nodes of the decision tree, ϖ_j represents the prediction score of the corresponding point of each leaf node, and γ and λ are the structure part and leaf weight of the decision tree, respectively. According to formulas (11) and (12), we obtain the penalty coefficient. According to equations (11) and (12), the XGBoost model expands the loss function to the quadratic term by applying the Taylor series at $y^{(t-1)}$ and uses the first and second derivatives of the error function, so it is more accurate than the GBDT model. After unfolding the objective function, through a series of transformations, under the condition of a given decision tree structure, the optimization objective can be transformed into the problem of solving the minimum value of the quadratic function of one variable. Finally, through the greedy algorithm, we continuously try to segment the existing leaf nodes and compare the gain of the objective function before and after the segmentation, until the optimal decision tree model of the t-th iteration is obtained.

The importance of explanatory variables in the XGBoost model can be measured in a variety of ways. For example, we can calculate the number of times the explanatory variable is used as a split feature in all decision trees, calculate the Gini coefficient reduction value of all nodes split by this feature, or calculate the sum of information gain. The importance of all explanatory variables is arranged from large to small to get the importance ranking of the explanatory variables in the XGBoost model. In this paper, we calculated the number of explanatory variables used as split features in all decision trees as a reference standard for importance.

2.5. Model Assessment. To compare the nonnested modes based on maximum likelihood, we use the Akaike information criterion (AIC) [39], which can be expressed as follows:

$$\mathrm{AIC} = -2 \ln L + 2k, \tag{13}$$

where $\ln L$ is the maximum log-likelihood and k is the number of parameters in the model. The lower the AIC, the better. Another evaluation method is the Bayesian information criterion, which can be expressed as follows:

$$\mathrm{BIC} = -2 \ln L + (\ln n) \cdot k, \tag{14}$$

with the model with the lowest BIC preferred, which was proposed by Schwarz [40]. We compared the predicted versus the observed number of hospitalizations for the competing model and found a well-fitting regression model that leads to predicted values closer to the observed data value. We also conducted a marginal analysis based on the results of regression to explain the impacts on the elderly hospitalization decision-making. All analyses were performed using Stata (Stata/SE version 15.1 for Windows, StataCorp LLP, College Station, TX, USA) statistical software.

3. Results

3.1. Descriptive Analysis. Table 1 presents the descriptive statistics of the number of elderly hospitalizations, demographic, socioeconomic, and health status. The gender distribution was almost balanced, and the average age of the elderly was 84 years. The elderly were more divorced, widowed, or single, and their education years were no more than 4 years. Medical consumption accounts for more than 60% of total household consumption, of which out-of-pocket expenses account for 17.7% of medical consumption expenditure. It can be seen from the box plot (Figure 2) that the elderly with more years of education have relatively better health status. Moreover, as is evident from Figure 3 that the higher the household income, the better the self-rated health status of the elderly.

The number of hospitalizations in our data is overdispersed because of its variance (0.97) greater than the mean (0.41). The t-test and auxiliary regression were applied to further verify whether the number of hospitalizations is overdispersed significantly [41]. We assume that the variance and mean of the number of hospitalizations were satisfied with the following relations, using the statistics to test the parameter was equal to zero or not. The t-test result showed $t = 5.84$, and the value of p was almost equal to zero, which indicated that the data were overdispersed remarkably. The proportion of elderly people who were not hospitalized within two years was as high as 75.06% as shown in Figure 1.

3.2. Model Evaluation and Comparison. In this study, we used the Poisson, negative binomial, zero-inflated Poisson, zero-inflated negative binomial, hurdle Poisson, and hurdle negative binomial models to fit the data, respectively. Table 2 reports the results of model comparison based on the AIC, BIC, and the log-likelihood. The Poisson model had the worst goodness of fit, while the zero-inflated negative was the best model in the six competing models according to the criteria, which was also consistent with the data characteristics above. Therefore, the conclusion of the zero-inflated negative binomial distribution model was more reliable.

3.3. Factors Associated with Hospitalization. Table 3 presents the analysis of the number of hospitalizations for a series of demographic, health status, socioeconomic, medical expenditure, and habits factors, using different regression. The results were made up of two separate parts: one part models the odds of being hospitalized and the other depicts the number of hospitalizations for the elderly who had at least one hospitalization. The logistic component of ZINB showed that the older adults whose daily activities were restricted, respectively, had 0.691 (95% CI: 0.501, 0.952) and 0.573 (95%

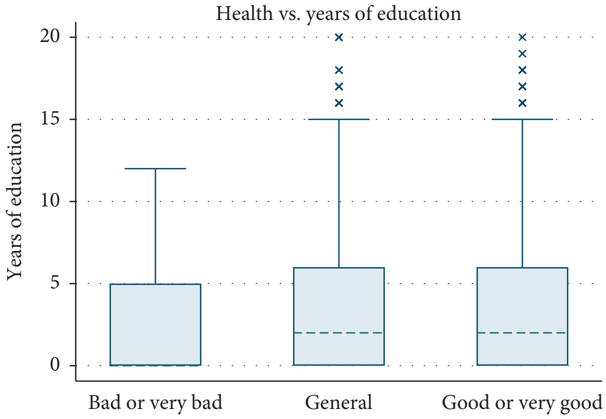

Figure 2: Years of education and self-rated health status.

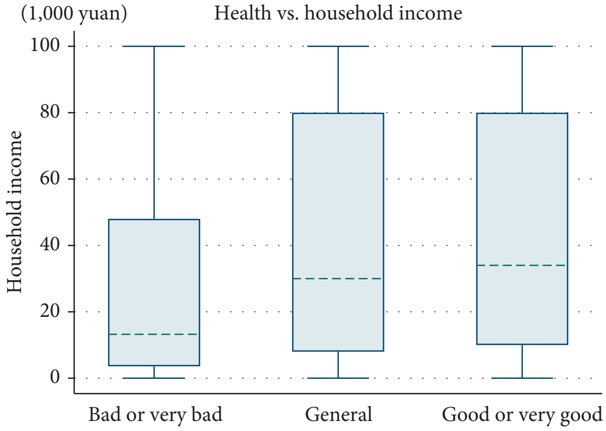

Figure 3: Household income and self-rated health status.

Table 2: Model comparison based on the AIC, BIC, and log-likelihood of six regression models.

Model	Number of parameters	Log-likelihood	AIC	BIC
Poisson	18	4,443.5	8,923	9,041.315
Negative binomial	19	4,089.657	8,217.313	8,342.201
Zero-inflated Poisson	36	3,444.408	6,960.817	7,197.445
Zero-inflated negative binomial	37	3,411.642	6,897.285	7,140.486
Hurdle Poisson	36	4,032.43	8,136.861	8,348.498
Hurdle negative binomial	37	3,942.265	7,958.531	8,175.353

CI: 0.363, 0.906) times lower odds of being hospitalized than those patients who were unrestricted. Males were 0.686 (95% CI: 0.483, 0.976) times less likely to be hospitalized than females. The elderly people with an additional year of education had 1.046 (95% CI: 1.003, 1.090) times of being hospitalized higher than before. For extra 1,000 yuan in hospitalization, the odds of being hospitalized were 0.145 (95% CI: 0.089, 0.234) times lower. Patients enrolled in URCMS were 1.937 (95% CI: 1.258, 2.983) times more likely to be hospitalized than patients enrolled in other medical plans. The results of the count component of the ZINB model indicated that the older adults who were in good health have 0.660 (95% CI: 0.462, 0.943) times lower odd of hospitalizations. There was a higher probability of hospitalization among the participants who needed assistance with ADLs, and the IRR is 1.302 (95% CI: 1.156, 1.466) and 1.471 (95% CI: 1.260, 1.717), respectively. Additional 1,000 yuan for household income and out-of-pocket expenses, the IRR of hospitalizations was, respectively, 0.998 (95% CI: 0.997, 1.000) and 0.994 (95% CI: 0.989, 0.999) times less likely to have multiple hospitalizations. On the contrary, the elderly people with more hospitalization medical expenditure had higher IRR, that is, 1.009 (95% CI: 1.005, 1.013) times more likely to have repeated hospital readmissions. For those who were at risk of using multiple hospitalizations service, NRCMS and UEBMI/URBMI were significantly associated with more use. They had higher IRR, respectively, 1.223 (95% CI: 1.015, 1.473) and 1.318 (95% CI:

Table 3: Regression analysis of factors using the ZINB model.

Variable	Logit part			Count part		
	OR	95% CI	*P* value	IRR	95% CI	*P* value
Self-rated health: very bad or bad = 0 (reference); general = 1; good or very good = 2	0.947	(0.298, 3.015)	0.927	0.869	(0.614, 1.231)	0.429
	1.080	(0.335, 3.483)	0.898	0.660	(0.462, 0.943)	0.022
ADL difficulties: unrestricted = 0 (reference); restricted = 1; very limited = 2	0.691	(0.501, 0.952)	0.024	1.302	(1.156, 1.466)	≤0.001
	0.573	(0.363, 0.906)	0.017	1.471	(1.260, 1.717)	≤0.001
Age	0.978	(0.816, 1.174)	0.813	1.055	(0.984, 1.132)	0.131
Age square	1.000	(0.999, 1.001)	0.758	1.000	(0.999, 1.000)	0.125
Gender: female = 0 (reference); male = 1	0.686	(0.483, 0.976)	0.036	0.881	(0.766, 1.012)	0.074
Marital status: not in marriage = 0 (reference); in marriage = 1	0.841	(0.601, 1.177)	0.313	0.993	(0.876, 1.127)	0.917
Years of education	1.046	(1.003, 1.090)	0.035	0.998	(0.985, 1.013)	0.885
Smoking or not: not smoking = 0 (reference); smoking = 1	0.820	(0.581, 1.158)	0.260	1.128	(0.988, 1.288)	0.074
Drinking or not: not drinking = 0 (reference); drinking = 1	1.239	(0.884, 1.738)	0.213	1.126	(0.991, 1.278)	0.069
Household income	1.001	(0.997, 1.005)	0.700	0.998	(0.997, 1.000)	0.050
Hospitalization medical expenditure	0.145	(0.089, 0.234)	≤0.001	1.009	(1.005, 1.013)	≤0.001
Out-of-pocket expenses	0.871	(0.626, 1.212)	0.411	0.994	(0.989, 0.999)	0.024
Medical insurance: others = 0 (reference); NRCMS = 1; UEBMI/URBMI = 2; free medical treatment = 3	1.937	(1.258, 2.983)	0.003	1.223	(1.015, 1.473)	0.035
	1.458	(0.901, 2.360)	0.125	1.318	(1.086, 1.599)	0.005
	0.723	(0.340, 1.538)	0.400	1.226	(0.913, 1.645)	0.175

1.086, 1.599) times more likely to be hospitalized compared to those in the other medical plans.

3.4. Importance Ranking of Impact Factors.

The importance of an explanatory variable depends on the sum of the information gain of splitting it divided by the number of times that it is used as a split feature. The larger the value, the greater the importance of the variable. Based on the training results of the XGBoost model, the variables are sorted according to the importance, as shown in Figure 4.

As shown in Figure 4, medical consumption and out-of-pocket expenses were the most important factors for the number of elderly hospitalizations. The main reason was that the current medical burden in China was too heavy [1, 2], which has seriously affected the elderly's decision-making and hospitalization behavior. The importance of medical insurance factors was relatively weak, which was related to the current low level of medical insurance in China. From the perspective of personal characteristics, age factors, education level, health status, and mobility restrictions were more important for the number of hospitalizations for the elderly, while the impact of smoking, drinking, and other lifestyle habits on the number of hospitalizations was not very important. In addition, due to the low importance of age and marital status, they were not shown in Figure 3,

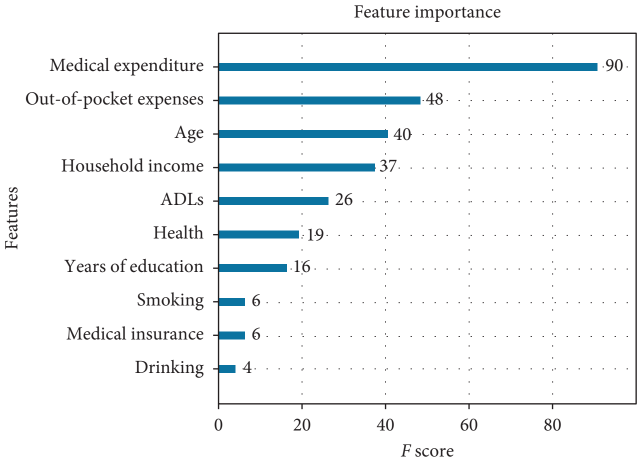

Figure 4: Importance ranking of impact factors.

which also showed that these two types of factors had the weakest influence.

3.5. Prediction of the Number of Hospitalizations. After the model fitting, the best zero-inflated negative binomial model was used to produce the predictions of the expected numbers of the elderly being hospitalized. The average predicted frequencies of the zero-inflated negative binomial model almost tended to the average observed frequencies, which had the smallest deviance from the observed as shown in Figure 5. The results indicated that the zero-inflated negative binomial model handled both issues of overdispersed and excessive zeros effectively and improved both modeling fitting and prediction.

3.6. The Effect of Different Medical Insurance on the Number of Hospitalizations. As outlined in Table 3, different medical insurance has different effects on the numbers of hospitalization of the elderly. Household income, hospitalization medical expenditure, and out-of-pocket medical expense had a significant impact on the numbers of multiple hospitalizations of the elderly. It could be observed from Figure 6 that the elderly with NRCMS, UEBMI/URBMI, and free medical care are more likely to choose hospitalization than those with commercial medical insurance, and the impact of UEBMI/URBMI was more significant. In comparison with the trend of hospitalization expenses, we found another tendency (as shown in Figure 7). Figure 7 reveals that the number of hospitalizations increased first and then decreased with the increase of out-of-pocket expenses. The elderly people with UEBMI had the highest number of hospitalizations. As was evident from Figure 8, the number of hospitalizations of the elderly with high family income decreased on the contrary. The older patients with free medical care had a relatively high number of hospitalizations compared with Figures 6 and 7.

4. Discussion

In this study, we compared various counts models such as Poisson, negative binomial, zero-inflated Poisson, zero-inflated negative binomial, hurdle Poisson, and hurdle negative binomial to analyze the influencing factors of the elderly people hospitalizations. It appeared to suggest that the zero-inflated negative binomial model was the best model among the six models applied in this study. Generally speaking, the Poisson was not to fit the overdispersed count data because of its characteristics of equal mean and variance. In this case, the negative binomial model was a better candidate model due to its variance greater than the mean. But when the count data were overdispersed and zero-inflated, neither the Poisson nor negative binomial model was suitable. There was a clear phenomenon that the number of elderly hospitalizations suffered from a serious zero-inflated problem. In this situation, the two-part models, such as zero-inflated and hurdle models, would be considered. The difference between zero-inflated and hurdle models lied in their way of interpreting and analyzing zero counts according to their generating data process [19, 20]. In most cases, zero observations in the zero-inflated models had two different sources, namely, structural and sample zeros, but there was only one source in the hurdle models [23]. In the present study, all the study elderly participants had no hospitalizations, and therefore, the excessive zeros either came from the structural zeros, which was true that there

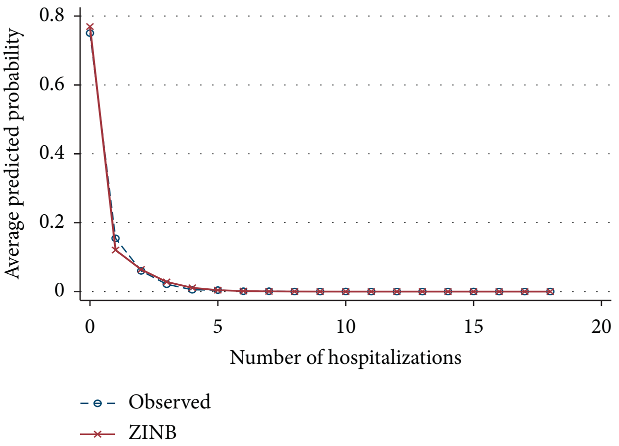

Figure 5: Average frequency of ZINB against the observed.

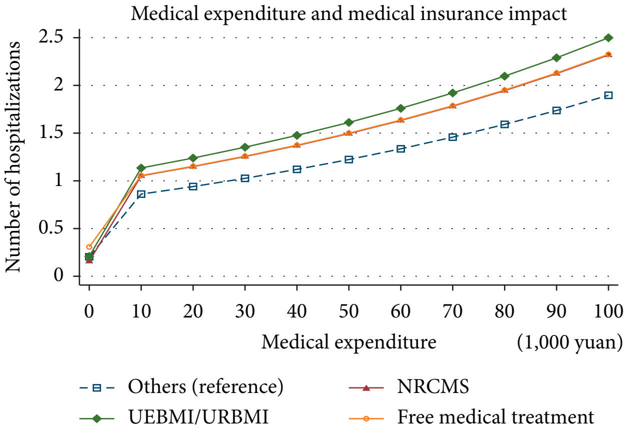

Figure 6: Medical expenditure and medical insurance impact.

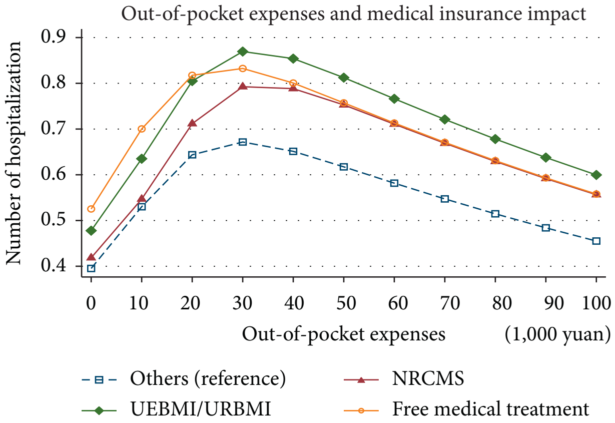

Figure 7: Out-of-pocket expenses and medical insurance impact.

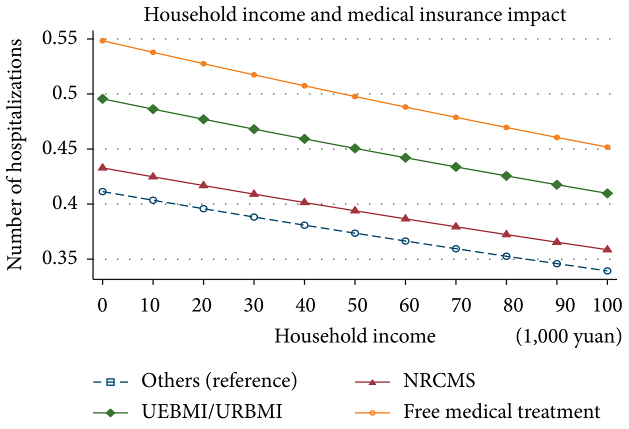

FIGURE 8: Household income and medical insurance impact.

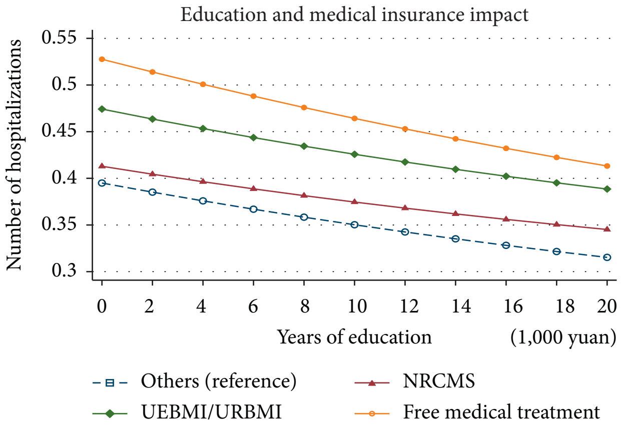

FIGURE 9: Education and medical insurance impact.

was no need for some elderly to be hospitalized at all, or from the sample zeros, which might be that the elderly had a need for hospitalization, but a zero outcome was produced due to sampling variability. Thus, this finding suggested that the zero-inflated negative binomial model would be more appreciate to handle the elderly hospitalization data.

This study led to a better understanding of factors contributing to hospitalization decision-making behavior and hospitalization needs of the elderly. These results suggested that the impacting factors on the odds of having multiple hospitalizations versus no hospitalization for the elderly were significantly different. To investigate whether there was a “U” relationship between the age and the number of hospitalizations, we need to introduce the age square term. As outlined in Table 3, there is no probability of a “U” relationship because of the square coefficient not being significant. In general, after the elderly choose to be hospitalized, the number of hospitalizations would gradually increase with age, which could be indicated from the odds of age and age square less than one.

Findings from this study revealed a clear tendency: the number of hospitalizations of the elderly had decreased with the increase of years of education and household income as detailed in Figures 8 and 9. This point was perhaps due to the impact of health. The higher level of education and the more household income tended to be in better health, which resulted in fewer hospitalizations [42]. In addition, the elderly with high household income might focus more on the quality of hospitalization service rather than the number of times, and they could choose a better quality of hospitalization service than the times.

For the impact of hospitalization medical expenditure and out-of-pocket expenses on the number of hospitalizations, our results indicated a clear distinction. We noticed that high out-of-pocket expenses might prevent the elderly from being hospitalized, but the increasing hospitalization expenditure brought the increasing number of hospitalizations (as shown in Table 3) on the contrary. Although the SHI system in China had covered the proportion of the population from a moderate rate of 50% to a near-universal rate of 95% during 2005–2011 [42, 43], the out-of-pocket rate associated with SHI, which ranges from 40% to 70%, remains a major financial challenge to patients with severe illnesses [43, 44]. It is seen from Figure 7 that the elderly would be more willing to choose hospitalization when the out-of-pocket expenses were within their affordability (about 30,000 yuan); however, most elderly people might give up hospitalization once this amount was exceeded. We also found that the increase in hospitalization medical expenditure did not reduce the repeated hospital readmissions for the elderly. More elderly people would choose to be hospitalized even if the hospitalization medical expenditure was relatively low (about lower 10,000 yuan). But the elderly will be less willing to choose to be hospitalized when the hospitalization medical expenditure exceeds 10,000 yuan, which could be observed in Figure 6.

Different medical insurance had different effects on the number of hospitalizations for the elderly people enrolled in different SHI from Figures 6–9. The effect of free medical care on the number of hospitalizations of elderly people was not significant. The reason was that the elderly who enjoyed free medical care were generally in good health (among the 195 elderly people with free medical care, only 3 were in poor health). On the other hand, it might be because there is a small proportion (about 3.69%) of such people in the sample data. We noticed that the enrollment in UEBMI/URBMI indicated no significant probability of whether or not being hospitalized, while they were more likely to use more hospitalizations after being hospitalized. The elderly people enrolled in NRCMS were more likely to incur multiple hospitalizations, which showed that this protective effective effect was significant. We observed from Figures 6 and 7 that the elderly with UEBMI/URBMI were hospitalized more often than the elderly with NRCMS. The reason perhaps was that the three schemes differed in the scopes of covered service and conditions, but also the financial generosity of the schemes varied substantially across regions [43]. The reimbursement scope and upper limits of NCRMS were more limited than those of UEBMI/URBMI because of lower funding levels [45]. As a result of NCRMS reimbursement policy’s upper limits for inpatient medical services being usually quite low and failing to compensate adequately for the hospitalization medical expenses, the elderly people had to abandon being hospitalized or reduce the numbers [42].

In this study, we selected the XGBoost model for the first time to rank the importance of the factors that affected the elderly's hospitalization decision and found that there was a slight difference from the significant influencing factors in the zero-inflated negative binomial model, but the difference was not significant. At the same time, the reliability of the influencing factors selected in the zero-inflated negative binomial regression model was also verified. This is also a contribution of this paper. Another contribution is the applications of models. Since there were a large number of zeros in the number of elderly hospitalizations, continuing to use the traditional counting models would underestimate the standard deviation and overestimate the significance level, so we consider using the zero-inflated negative binomial model to solve this problem.

4.1. Limitations. Some limitations of this study should be noted. Firstly, the CLHLS study sample included missing data, and we did not have a conservative evaluation of hospitalization, which might result in some biases. Besides, we could not consider the impact of UEBMI and URBMI on the number of hospitalizations separately, since the information was not captured in the questionnaire. Future research incorporating such information is needed for understanding the influence of different medical insurance including commercial medical insurance and basic medical insurance on the number of hospitalization of the elderly. As mentioned above, only 3.6% of the elderly in the sample have access to free medical care, so findings from this study might underestimate its significance. Further studies based on a large sample are warranted in order to investigate the free medical care effect more properly. Machine learning has been applied in many fields [46, 47]. As a relatively new algorithm in the field of machine learning, the XGBoost method is less used in the field of healthcare. We only applied it to select the variables and find their importance. There is a need to further research in the future. In addition, the impact of major health and public health events such as COVID-19, SARS, and so on will also affect the elderly's decision-making in hospitalization. These issues can be discussed in future studies. Lastly, the data were pooled cross-section data from different years; we could not infer a causal relationship hardly.

5. Conclusions

This study reported that the ZINB model was the best fit among other models account for the number of hospitalizations of the elderly. Our findings indicated that the higher hospitalization medical expenditure inhabited the willingness of the elderly to be hospitalized, and the male elderly were less willing to be hospitalized. Difficulties with upper restricted functioning were even less likely to choose hospitalizations. The NRCMS had significantly increased the willingness of the elderly to be hospitalized. Our findings also revealed that there was no "U" relationship between age and the number of hospitalizations. The elderly with NRCMS or UEBMI/URBMI were more likely to have repeated hospital recommissions, but the free medical treatment had little impact on the hospitalization service. Increasing out-of-pocket expenses for hospitalization would reduce the number of hospitalizations for the elderly, but the impact on their hospitalization needs was not significant. The economic burden of hospitalization for the elderly should be further reduced through various safeguard measures such as major illness medical insurance, medical assistance, and special assistance. Attention should be paid to the medical service needs of low-income elderly; more financial subsidies should be provided to their medical service utilization; and the inequality should be reduced in the utilization of medical services for elderly people with different incomes. Our study led to a better understanding of factors contributing to increased inpatient hospitalizations among the elderly, which might help reduce the financial burden of hospitalization for the elderly and truly achieve "affordable hospitalizations."

Data Availability

Data are available in a public, open-access repository. The data were collected by the project entitled "Chinese Longitudinal Healthy Longevity Survey" (CLHLS) jointly implemented by the Center for Healthy Aging and Development Studies of Peking University. CLHLS is supported by funds from China Natural Science Foundation, China Social Science Foundation, China 211 projects of the Ministry of Education, China 973 projects of the Ministry of Science and Technology, China National Science and Technology Support Plan, NIA/NIH, and UNFPA (http://chads.nsd.pku.edu.cn/xwdt/512297.htm).

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

This study was supported by the Planning Project of Beijing Social Science Fund (Grant no. 20SRB010).

References

[1] X. Liu, H. Wong, K. Liu et al., "Outcome-based health equity across different social health insurance schemes for the elderly in China," *BMC Health Services Research*, vol. 16, no. 9, 2016.

[2] X. Lin, M. Cai, H. Tao et al., "Insurance status, inhospital mortality and length of stay in hospitalised patients in Shanxi, China: a cross-sectional study," *BMJ Open*, vol. 7, no. 7, Article ID e015884, 2017.

[3] L. F. Liu, W. H. Tian, and H. P. Yao, "Utilization of health care services by elderly people with National Health Insurance in Taiwan: the heterogeneous health profile approach," *Health Policy (Amsterdam, Netherlands)*, vol. 108, no. 2-3, pp. 246–255, 2012.

[4] V. Milos, Å Bondesson, M Magnusson, U Jakobsson, T Westerlund, and P Midlöv, "Fall risk-increasing drugs and falls: a cross-sectional study among elderly patients in primary care," *BMC Geriatrics*, vol. 14, p. 40, 2014.

[5] A. Finkelstein and R. Mcknight, “What did Medicare do? The initial impact of Medicare on mortality and out of pocket medical spending,” *Journal of Public Economics*, vol. 92, no. 7, pp. 1644–1668, 2008.

[6] X. Li and W. Zhang, “The impacts of health insurance on health care utilization among the older people in China,” *Social Science & Medicine*, vol. 85, pp. 59–65, 2013.

[7] X. Zhang, M. E. Dupre, L. Qiu, W. Zhou, Y. Zhao, and D. Gu, “Urban-rural differences in the association between access to healthcare and health outcomes among older adults in China,” *BMC Geriatrics*, vol. 17, no. 1, p. 151, 2017.

[8] Z. Zhou, Y. Fang, Z. Zhou et al., “Assessing income-related health inequality and horizontal inequity in China,” *Social Indicators Research*, vol. 132, no. 1, pp. 241–256, 2017.

[9] M. Zhao, B. Liu, L. Shan et al., “Can integration reduce inequity in healthcare utilization? Evidence and hurdles in China,” *BMC Health Services Research*, vol. 19, no. 1, 2019.

[10] J. Z. Ayanian, “Looking back to improve access to health care moving forward,” *JAMA Internal Medicine*, vol. 180, no. 3, pp. 448-449, 2020.

[11] J. Gao, J. H. Raven, and S. Tang, “Hospitalisation among the elderly in urban China,” *Health Policy (Amsterdam, Netherlands)*, vol. 84, no. 2-3, pp. 210–219, 2007.

[12] H. R. Waters, G. F. Anderson, and J. Mays, “Measuring financial protection in health in the United States,” *Health Policy*, vol. 69, no. 3, pp. 339–349, 2007.

[13] A. Wagstaff and M. Lindelow, “Can insurance increase financial risk?” *Journal of Health Economics*, vol. 27, no. 4, pp. 990–1005, 2008.

[14] P. George, B. Heng, J. De Castro Molina, L. Wong, N. Wei Lin, and J. T. Cheah, “Self-reported chronic diseases and health status and health service utilization—results from a community health survey in Singapore,” *International Journal for Equity in Health*, vol. 11, no. 1, p. 44, 2012.

[15] X. Liu, S. Tang, B. Yu et al., “Can rural health insurance improve equity in health care utilization? a comparison between China and Vietnam,” *International Journal for Equity in Health*, vol. 11, 2012.

[16] S. Kim and S. Kwon, “Has the National Health Insurance improved the inequality in the use of tertiary-care hospitals in Korea?” *Health Policy*, vol. 118, no. 3, pp. 377–385, 2014.

[17] R. Winkelmann, *Econometric Analysis of Count Data*, Springer, Berlin, Germany, 2013.

[18] J. M. Woolridge, *Econometric Analysis of Cross Section and Panel Data*, MIT Press Books, vol. 1, no. 2, , pp. 206–209, Cambridge, MA, USA, 2011.

[19] C. E. McCulloch, “Generalized, linear, and mixed models,” *Journal of the Royal Statistical Society*, vol. 52, no. 2, pp. 242-243, 2003.

[20] J. A. Nelder and R. W. M. Wedderburn, “Generalized linear models,” *Journal of the Royal Statistical Society. Series A (General)*, vol. 135, no. 3, pp. 370–384, 1972.

[21] D. Lambert, “Zero-inflated Poisson with an regression, in manufacturing to defects application,” *Technometrics*, vol. 34, no. 1, p. 14, 1992.

[22] W. H. Greene, “Accounting for excess zeros and sample selection in Poisson and negative binomial regression models,” in *Working Papers*New York University, New York, NY, USA, 2008.

[23] M.-C. Hu, M. Pavlicova, and E. V. Nunes, “Zero-inflated and hurdle models of count data with extra zeros: examples from an HIV-risk reduction intervention trial,” *The American Journal of Drug and Alcohol Abuse*, vol. 37, no. 5, pp. 367–375, 2011.

[24] K. Hur, D. Hedeker, W. Henderson, S. Khuri, and J. Daley, “Modeling clustered count data with excess zeros in health care outcomes research,” *Health Services & Outcomes Research Methodology*, vol. 3, no. 1, pp. 5–20, 2002.

[25] R. Winkelmann, “Health care reform and the number of doctor visits-an econometric analysis,” *Journal of Applied Econometrics*, vol. 19, no. 4, pp. 455–472, 2004.

[26] W. Greene, “Models for count data with endogenous participation,” *Empirical Economics*, vol. 36, no. 1, pp. 133–173, 2009.

[27] B. Mihaylova, A. Briggs, A. O’Hagan, and S. G. Thompson, “Review of statistical methods for analysing healthcare resources and costs,” *Health Economics*, vol. 20, no. 8, pp. 897–916, 2011.

[28] X. Liu, B. Zhang, L. Tang et al., “Are marginalized two-part models superior to non-marginalized two-part models for count data with excess zeroes? Estimation of marginal effects, model misspecification, and model selection,” *Health Services and Outcomes Research Methodology*, vol. 18, no. 3, pp. 175–214, 2018.

[29] P. Deb and E. C. Norton, “Modeling health care expenditures and use,” *Annual Review of Public Health*, vol. 39, no. 1, pp. 489–505, 2018.

[30] T. Chen and C. Guestrin, “XGBoost: a scalable tree boosting system,” in *Proceedings of the 22nd ACM SIGKDD International Conference*, San Francisco, CA, USA, August 2016.

[31] Y. Zeng, D. L. Poston, D. A. Vlosky, and D. Gu, “Healthy longevity in China: demographic, socioeconomic, and psychological dimensions,” *Springer Ebooks*, vol. 63, no. 3, pp. 312-313, 2008.

[32] H. Zhang, Y. Wang, D. Wu, and J. Chen, “Evolutionary path of factors influencing life satisfaction among Chinese elderly: a perspective of data visualization,” *Data*, vol. 3, no. 3, p. 35, 2018.

[33] Y. Zeng, Q. Feng, T. Hesketh, K. Christensen, and J. W. Vaupel, “Survival, disabilities in activities of daily living, and physical and cognitive functioning among the oldest-old in China: a cohort study,” *The Lancet*, vol. 389, no. 10079, pp. 1619–1629, 2017.

[34] R. M. Andersen, “Revisiting the behavioral model and access to medical care: does it matter?” *Journal of Health and Social Behavior*, vol. 36, no. 1, pp. 1–10, 1995.

[35] K. N. Ukwaja, I Alobu, S Abimbola, and P. C Hopewell, “Household catastrophic payments for tuberculosis care in Nigeria: incidence, determinants, and policy implications for universal health coverage,” *Infectious Diseases of Poverty*, vol. 2, no. 1, p. 21, 2013.

[36] Y. Pan, S. Chen, M. Chen et al., “Disparity in reimbursement for tuberculosis care among different health insurance schemes: evidence from three counties in central China,” *Infectious Diseases of Poverty*, 2016.

[37] Y. B. Cheung, “Zero-inflated models for regression analysis of count data: a study of growth and development,” *Statistics in Medicine*, vol. 21, no. 10, pp. 1461–1469, 2002.

[38] A. F. Zuur, E. N. Ieno, N. J. Walker, A. A. Saveliev, and G. M. Smith, *Zero-Truncated and Zero-Inflated Models for Count Data*, Springer, New York, NY, USA, 2009.

[39] H. Akaike, B. N. Petrov, and F. Czaki, “Information theory and an extension of the maximum likelihood principle,” in *Proceedings of the 2nd International Symposium on Information Theory*, Amsterdam, Netherland, May 1973.

[40] G. E. Schwarz, “Estimating the dimension of a model,” *The Annals of Statistics*, vol. 6, no. 2, 1978.

[41] A. C. Cameron and P. K. Trivedi, *Microeconometrics: Methods and Applications*, China Machine Press, Beijing, China, 2008.

[42] Y. Li, V. Malik, and F. B. Hu, "Health insurance in China: after declining in the 1990s, coverage rates rebounded to near-universal levels by 2011," *Health Affairs*, vol. 36, no. 8, pp. 1452–1460, 2017.

[43] H. Yu, "Universal health insurance coverage for 1.3 billion people: what accounts for China's success?" *Health Policy*, vol. 119, no. 9, pp. 1145–1152, 2015.

[44] C. Zhang, X. Lei, J. Strauss, and Y. Zhao, "Health insurance and health care among the mid-aged and older Chinese: evidence from the national baseline survey of CHARLS," *Health Economics*, vol. 26, no. 4, pp. 431–449, 2017.

[45] X. Zhang, Q. Wu, Y. Shao, W. Fu, G. Liu, and P. C. Coyte, "Socioeconomic inequities in health care utilization in China," *Asia Pacific Journal of Public Health*, vol. 27, no. 4, pp. 429–438, 2015.

[46] Y. Z. D. Sun, "Machine learning techniques for screening and diagnosis of diabetes: a survey," *Tehnicki Vjesnik-Technical Gazette*, vol. 26, no. 3, pp. 872–880, 2019.

[47] J. Li, S. Pan, L. Huang, and X. Zhu, "A machine learning based method for customer behavior prediction," *Tehnicki Vjesnik-Technical Gazette*, vol. 26, no. 6, pp. 1670–1676, 2019.