Professor Xihua LIU, PhD
School of Business,Qingdao University
Email:xihua-liu@163.com
Xuetao ZHANG, PhD
School of Business,Qingdao University
Email: qddxzxt@126.com
Master Xuejing YANG (Corresponding author)
School of Economics, Qingdao University
Email: qdulc2016@163.com

FRAUD RISK MEASUREMENT OF BASIC MEDICAL INSURANCE FOR URBAN AND RURAL RESIDENTS IN CHINA

Abstract. *This paper gathers the loss data due to medical insurance fraud from 2006 to 2018 , in order to measure the fraud risk of basic medical insurance for urban and rural residents in China. The fraud risk measurement problem has two features: "low loss with high frequency" and "high loss with low frequency", which can separately use normal distribution and Generalized Pareto Distribution (GPD). Under the framework of loss distribution method, the TVaR value of fraud risk loss can be computed from the two-stage (PSD-LDA) model, which is based on the Bayesian Markov Chain Monte Carlo (MCMC). Furthermore, the fraud risk reserve required to be accrued for urban and rural residents' medical insurance can be determined. Research findings indicate the fraud of basic medical insurance has the characteristic of Tail-risks for urban and rural residents in China, the maximum fraud loss within one year being 95% likely to be 72.0 million yuan, as well as 99.9% likely to be 168.5 million yuan, and the government requiring 9.47‰ of the risk reserve to withstand the risk of fraud. The result provides decision basis for scientific calculation of financing standards, fraud risk pricing and early warning of fraud risk in China.*

Key words: *Medical insurance fraud, Loss distribution method, PSD-LDA, Bayesian MCMC.*

JEL classification: G32 si I13

1. Introduction

With the rapid development of medical insurance industry, health insurance fraud has caused losses of hundreds of millions of dollars in many countries every year (USGAO, 1992), thus, it poses a great threat to the safety of health insurance funds. Ekin et al. (2018) point out that medical fraud has recently attracted more attention because of the increases in healthcare spending and overpayments. The

DOI: 10.24818/18423264/53.3.19.16

total health care expenditures in the United States have reached $3.2 trillion, which corresponds to $9,990 per person in 2015 CMS (Bauder et al., 2017). In order to calculate financing standards scientifically, to promote the stable development of the health insurance market and ensure the effective implementation of medical insurance policies, fraud risk measurement is becoming more and more important. The first report on the cost of world health care was published in 2009, using 66 successful measurement projects in six countries to calculate fraud losses (the fraud rate is 5.6%). Kong et al. (2012) point out that fraud in the field of medical insurance involves a huge amount of money[1]. After auditing, medical expenses for those who do not meet the basic medical insurance conditions of urban workers are reimbursed for 82.69 million yuan in the whole year in China. In January 2017, the announcement of the Audit Result of Medical Insurance Fund evidence that 140 million yuan of medical insurance personal account funds were withdrawn cash or used to purchase commodities and other expenditures in China. With the continuous increase of medical insurance fraud, providing the decision basis for scientific calculation of financing standards is quite significant. At present, many scholars have analyzed insurance fraud from a theoretical point of view. From the Perspective of the Austrian Insurance Industry, Maximilian et al. (2018) introduce the basic underlying principles of insurance and the underlying reasons for the prevalence of insurance fraud. They illustrate the types of insurance fraud committed by the offenders and the dimensions of insurance fraud in Austria. Tseng et al. (2019) point out the relationships among fraud types, moral intensity, demographic variables, customers' ethical attitudes, and intentions towards insurance frauds. Understanding the relationships among these variables could provide implications for those involved in the practice of anti-fraud programs, such as significant impacts on the respondents' perceptions of moral intensity and fairness. Looking at the development process of the insurance industry, we have the same experience in Europe and America. The national health strategies adopted by governments should be constantly updated according to the system needs, ensuring the implementation of an anti-fraud body. Accordingly, Lobont et al. (2019) proved that the quality of the public system in European countries is greatly affected by a reduced level of government expenditures on health, facilitating various forms of system sabotages. Jim (2012) illustrates that a survey of 33 organizations in six European countries finds out that the loss rate of medical fraud is 3.29%-10.0%, averaging 5.59%, the incidence of medical fraud was 0.47%-7.1%, averaging 4.23%. Ghcan (2016) points out that medical fraud amounts to 180 billion euros per year, equivalent to 6% of global health expenditure and the total annual GDP of Finland or Malaysia. Pei (2009) finds that 302 insurance companies in the United States failed between 1969 and 1990, from which 30% failed because of inadequate prevention and control of insurance fraud. Relevant data point out that 42 states in the United States have established insurance fraud prevention and control agencies, whose input-output ratio is 1:27,

[1] Data from the National Audit Result of Social Security Funds, 2010.

DOI: 10.24818/18423264/53.3.19.16

that is, every dollar invested in the prevention and control of insurance fraud crime can recover the loss of $27. To sum up, the governance of insurance fraud is one of the problems that need to be solved urgently. At the same time, strengthening the research on insurance fraud can provide certain theoretical and practical significance for anti-fraud.

China issued the Opinions on Integrating the Basic Medical Insurance System for Urban and Rural Residents in 2016, which integrates the basic medical insurance for urban residents and the New Rural Cooperative Medical System (NCMS) to establish a unified basic medical insurance system for urban and rural residents. The system mainly serves to assist the urban and rural residents who become poor due to serious illness. Xiang et al. (2014) illustrate that China has achieved full coverage of urban and rural areas by the end of 2008， since the establishment of medical aid system in rural and urban areas in 2003 and 2005. This policy has benefited nearly 100 million people in need. The medical insurance system for urban and rural residents has alleviated difficult and expensive medical treatment and reduced the burden of medical treatment for urban and rural residents. With the continuous implementation of the system of medical insurance for urban and rural residents, one after another fraud of the medical insurance funds for urban and rural residents has occurred. The number of medical insurance fraud cases has increased continuously, and the fraud cases become increasingly serious, complicated, gang-forming and concealed. For example, the medical insurance fraud cases discovered in Shanghai in 2010 amounted to more than 5 million yuan; the medical insurance fund fraud cases discovered in Liuzhou, Guangxi in 2011 amounted to more than 3.5 million yuan, and medical insurance fraud cases discovered in Hengshan County in 2013 amounted to more than 3 million yuan. However, there is no authoritative organization to release statistical data on the fraud of medical insurance for urban and rural residents in China. According to statistics from medical insurance settlement centers around the country, unreasonable hospitalization expenses and the amount involved in fraud are also increasing year by year. Wei (2014) reveals that insiders estimate that the fund losses caused by medical insurance fraud for urban and rural residents in China account for about 20% to 30% of the total amount of reimbursement. Insurance fraud also exists in developed countries. Li and Liu (2010) point out that medical insurance fraud has attracted great attention of governments in various countries. Although the perpetrators of these medical insurance fraud cases have been sanctioned by relevant laws, they still bring irreparable losses to the medical insurance fund and infringe the interests of other honest insured persons. Seen from the greater consequences, the principle of utmost good faith in insurance operation has been seriously damaged by medical insurance fraud, and the contractual basis of insurance operation has also been violated. However, the relevant responsible departments have not implemented effective anti-fraud measures, and anti-fraud is not yet included in the central tasks in China. The reason is the externality of high costs and benefits brought by fraud identification. Therefore, an in-depth study on the identification and measurement of medical

DOI: 10.24818/18423264/53.3.19.16

insurance fraud from a quantitative perspective are significant. It can not only reveal the internal characteristics and laws of medical insurance fraud to some extent but also has far-reaching reference value for perfecting our insurance credit system, improving the efficiency of our medical insurance policy implementation and promoting anti-fraud research.

For this reason, this paper intends to study three aspects. First, we extend the research sample from the NCMS to medical insurance for urban and rural residents on the basis of literature, to further expand the scope of sample data. Second, we calculate the risk value of fraud losses caused by each subdivided fraud perpetrator, so as to provide more sufficient decision-making basis for the effective management and supervision of the medical insurance agencies and regulatory agencies. Third, we use Bayesian MCMC method based on Gibbs sampling to estimate the parameters, so as to solve the problem that the estimation error may increase due to Maximum Likelihood Estimation (MLE) used when the sample data is insufficient.

The remainder of our study is organized as follows: Section 2 briefly reviews related literature. Section 3 discusses our empirical methodology. Section 4 provides an overview of our dataset. Section 5 presents the empirical results. Section 6 concludes our paper.

2. Literature review

Previously, the empirical analysis methods of medical insurance fraud adopted in academic circles are mainly statistics-based, namely establishing statistical regression models based on claim cases. Among them, LOGIT, PROBIT and other regression models are widely used. Lious et al. (2008) identify fraud of medical service providers in Taiwan's health insurance system and find that the logical regression method has the best identification ability. Brockett and Derrig (2012), Li (2015), Jane (2018) study fraud probability by establishing a PRIDIT model and the improved PRIDIT model. Suliyanto (2019) uses LOGIT and PROBIT regression methods to design risk models for diabetes mellitus and hypertension and to identify important factors affecting diabetes mellitus and hypertension. Xu (2019) uses binary basic model, multi-LOGIT model, and bivariate PROBIT model to discuss the possibility of occurrence of the risk of the credit platform. In addition to discrete models, Hubick (1992), Viaene (2004), Ortega et al. (2006) apply an artificial neural network to fraud identification of health service providers. Marisa et al. (1996) propose the use of data mining techniques to identify medical insurance fraud, which is adopted by American health care financial management authority. Wan et al. (2006) use data mining method to recognize the fraud of suppliers. At the same time, artificial intelligence fraud identification technology is also widely used in the field of health insurance. According to behavior rules of heuristic learning and machine learning, the technology can be used to identify frauds in health insurance claims, such as Genetic algorithm (Hong, 2000), Bayesian network (Viaene, 2004), Decision tree (Amira et al., 2016), Random forest (Li et al., 2018). The estimated results are in good agreement with the basic

DOI: 10.24818/18423264/53.3.19.16

nature of the data set. It is a useful supplement to the relevant domestic research and has important guiding value in anti-fraud practice. However, a statistical regression model can only deal with limited data sets and require high data integrity.

At present, for empirical study, the lack of data is a major research obstacle (the data is not shared and is only owned by medical insurance agencies) and the complexity of medical insurance fraud (fraud can be original and speculative, can be single and collusive). Fraud indicators may be or may not be related to the individual characteristics of the actor. In practice, there may be numerous limitations for a single identification method. Two-Stage Loss Distribution Method (PSD-LDA) is popular in measuring the risk of commercial banks. LDA is first used to extract the text features hiding in the text descriptions of the accidents appearing in the claims, and deep neural networks then are trained on the data, which includes the text features and traditional numeric features for detecting fraudulent claims. The results reveal that their proposed framework that combines deep neural networks and LDA is a suitable potential tool for automobile insurance fraud detection. Li and Lin (2011) use aggregate risk model to measure the risk of medical insurance fraud according to the principle of loss distribution method and give the amount of risk reserve against fraud. Lin and Li (2014) take part in the fraud loss data of the National Center for Manufacturing Sciences (NCMS) from 2004 to 2012, as samples. Adopting a two-stage loss distribution method (PSD-LDA) they measure the Tail VaR value and the pure premium of the fraud risk loss of the NCMS. However, the studies do not further calculate the risk value of fraud losses of various fraud perpetrators, and estimation of the model parameters of the extreme part by using maximum likelihood estimation method would make the result unreliable due to insufficient samples. Liu and Zhu (2018) further reduce the interference of weak factors on neural network identification through combining logistic regression analysis method. Antonio et al. (2018) introduce a 3-parameter compound model to account for the unimodal hump-shaped, right skewed and heavy tails distribution of insurance losses. The model is applied to three famous insurance loss datasets and is a commonly used risk measure. Sadgali et al. (2019) indicate the fact that financial fraud presents more and more threat which has serious consequences in the financial sector. They propose a state of art on various fraud techniques such as classification, clustering, and regression.

Generally speaking, the research on medical insurance fraud at home and abroad focuses on qualitative analysis, which may be due to the difficulty in data acquisition or the complexity and relative invisibility of medical insurance fraud. The empirical studies are a few quantitative studies on the identification and measurement of medical insurance fraud. This paper is based on the relevant data, such as the fraud frequency of the medical insurance for urban and rural residents, the amount involved and the amount of losses from 2006 to 2018. According to the principle of loss distribution method and *VaR* method, the modeling is conducted on the frequency and severity of fraud losses respectively to fit their optimal

DOI: 10.24818/18423264/53.3.19.16

distribution function. Then, the optimal distribution function of total losses is fit by the Monte Carlo simulation method. After that, the risk loss reserve that should be accrued for fraud is measured to provide decision-making reference for calculating the financing standards of the medical insurance funds for urban and rural residents.

3. Methodology

Value at Risk (*VaR*) model is a way to manage the risk of financial derivatives, which is used to predict the maximum loss in financial transactions. In order to measure the fraud risk of the new rural cooperative medical system in an all-round way, we consider that Tail VaR (*TVaR*) meets the consistency principle of risk measurement. Lin *et al.* (2014) point out that the PSD-LDA method to measure the *TVaR* value of the fraud risk in the new rural cooperative medical system is credible. Since the frauds of medical insurance for urban and rural residents are characterized by "high frequency with low loss"(LLHF) and "low frequency with high loss"(HLLF), the relevant data is divided into two stages.

Based on the principle of loss distribution method, Wang *et al.* (2012) propose that the fraud risk of basic medical insurance for urban and rural residents is measured by using a two-stage model. First, under a certain confidence level, the models of fraud loss frequency and fraud loss severity are respectively constructed according to the principle of the *VaR* method, and their optimal distribution functions are fitted. Second, the total loss distribution is fitted based on the Monte Carlo simulation method, and the fraud risk reserve of the insurance for urban and rural residents is provided. The steps are as follows:

Step 1: Estimating the probability distribution of fraud loss frequency,

Step 2: Estimating the probability distribution of fraud loss severity,

Step 3: Monte Carlo simulation of the total loss distribution,

Step 4: Provision for fraud risk reserve.

Among them, it is assumed that fraud events are independent of each other, and fraud frequency and fraud severity are also independent of each other.

3.1 Threshold selection

The selection of a threshold is very important. Firstly, the data is divided into two stages, namely LLHF and HLLF, by threshold. Then, ξ and β in the generalized Pareto distribution (GPD) model are estimated. Generally, the hill graph and excess mean function graph are used to determine the threshold.

The samples are arranged in ascending order to obtain a fraud loss amount sequence $x = \{x_1, x_2, \cdots, x_n\}$. The threshold u is any uniform measurement between x_1 and x_n . The definition of the sample excess mean function is:

$$e(u) = \frac{\sum_{i=k}^{n}(x_i - u)}{N_u} \quad k = min\{i \mid x_i > u\} \tag{1}$$

Where N_u indicates the fraud frequency values exceeding the threshold. When the

DOI: 10.24818/18423264/53.3.19.16

threshold μ is sufficiently large, namely $x > u$, the curve $e(u)$ formed is approximately linear. If the slope $e(u)$ is positive, it indicates that the collected sample data conforms to the GPD distribution with positive shape parameters. Therefore, a suitable threshold u is selected to make the $e(u)$ graph approximate to a linear function with a positive slope when $x > u$, which indicates that the fraud data follows a thick-tailed GPD distribution.

3.2 Loss frequency distribution

Loss frequency N refers to the number of loss events in a certain period. It is defined as the fraud frequency of medical insurance for urban and rural residents each year. Poisson distribution, binomial distribution, and negative binomial distribution are usually adopted for frequency distribution. It is difficult to estimate the frequency of the low frequency part that exceeds the threshold. Peaks-over-threshold model provides a good method. According to the analysis and the proof of Leadbetter *et al.* (2008), the random sequence exceeding the threshold u gradually converges to a Poisson distribution with severity λ ($\lambda > 0$). Therefore, for the high frequency part, it is assumed that the occurrence frequency of fraud risk loss events follows the Poisson distribution.

3.3 Loss severity distribution

For the "high frequency with low loss" and "low frequency with high loss" of fraud risks of medical insurance for urban and rural residents, a two-stage model is used to define and fit the loss severity distribution function (Mo, 2017). Assuming that the loss distribution function is:

$$F(x) = \begin{cases} Ln(\mu, \sigma^2) & 0 < x \le u \\ GPD(\xi, \beta) & x > u \end{cases} \tag{2}$$

Among it, the high-frequency low-loss sequence follows a lognormal distribution.

Namely, when $0 < x \le u$, $x \sim Ln(\mu, \sigma^2)$, and the density function is:

$$f(x) = \begin{cases} \dfrac{1}{\sigma x \sqrt{2\pi}} \exp(-\dfrac{(Lnx - u)^2}{2\sigma^2}) & 0 < x \le u \\ 0 & x > u \end{cases} \tag{3}$$

where μ is mean value, σ is standard deviation, x is a random variable. The parameter estimation is realized by Bayesian simulation. The low frequency with high loss sequence follows GPD distribution. When $x > u$, $x \sim GPD$（ξ, β） and the density function is:

DOI: 10.24818/18423264/53.3.19.16

$$g_{\xi,\beta}(y)=\begin{cases}\frac{1}{\beta}\exp(\frac{\text{-y}}{\beta}) & \xi=0\\ \frac{1}{\beta}(1+\frac{\xi}{\beta}y)^{-1-\frac{1}{\xi}} & \xi\neq 0\end{cases} \tag{4}$$

The distribution function is:

$$G_{\xi,\beta}(y)=\begin{cases}1-\exp(\frac{\text{-y}}{\beta}) & \xi=0\\ 1-(1+\frac{\xi}{\beta}y)^{-\frac{1}{\xi}} & \xi\neq 0\end{cases} \tag{5}$$

Here, $y=x-u$, $\beta>0$. And when $\xi\geq 0$, $y\geq 0$. When $\xi<0$, $0\leq y\leq -{}^{\beta}\!/_{\xi}$. When $\xi>0$, thick-tailed GPD is matched. When $\xi=0$, the exponential distribution prevails. When $\xi<0$, Pareto distribution is thin-tailed. Wherein β represents a scale parameter and ξ represents a shape parameter.

The steps for estimating parameters by maximum likelihood estimation method are as follows:

If the sample is $\{x_1, x_2,\dots,x_n\}$ and the threshold is u, set $y_i=x_i-u$ and set the number of samples larger than the threshold u as N_u, and $N_u=\varphi$ can be obtained. Then, the log-likelihood function of the GPD distribution is:

$$L(\xi,\beta|Y)=-N_u Ln\beta-(1+\frac{1}{\xi})\sum_{i=1}^{N_u}Ln(1+\frac{\xi}{\beta}Y_i) \quad \xi\neq 0 \tag{6}$$

When the threshold u is determined, the log-likelihood equation can be obtained from the above equation:

$$\frac{\partial L}{\partial\beta}=-\frac{N_u}{\beta}+(1+\xi)\sum_{i=1}^{N_u}\frac{Y_i}{\beta(\beta+\xi Y_i)} \tag{7}$$

$$\frac{\partial L}{\partial\xi}=\frac{1}{\xi^2}\sum_{i=1}^{N_u}Ln(1+\frac{\xi}{\beta}Y_i)-(1+\frac{1}{\xi})\sum_{i=1}^{N_u}\frac{Y_i}{\beta+\xi Y_i} \tag{8}$$

The estimated parameter $\hat{\xi}$ and $\hat{\beta}$ can be obtained by solving the above equation.

3.4 Bayesian MCMC estimation

After the threshold is determined, if the excess quantity is small, there will be a big error in estimating the parameters by using the maximum likelihood estimation method. Therefore, this paper adopts a more effective Bayesian MCMC method to estimate the model parameters based on characteristics of the fraud data, including fewer samples and high loss severity (Bermúdeza *et al.*, 2008). In this paper, a wider prior distribution form is chosen for GPD parameters. In other words,

ξ (ξ ~Gamma(a1,b1)) follows GPD with the parameter a_1 , b_1 and β (β~Gamma(a2, b2)) follows Gamma distribution with parameter a_2 and b_2 . According to Bayesian basic principle, the joint distribution of parameter ξ and β is:

$$f(\xi,\beta|x)=\frac{L(x|\xi,\beta)f(\xi)f(\beta)}{\int\int L(x|\xi,\beta)f(\xi)f(\beta)d\xi d\beta} \tag{9}$$

Wherein $L(x|\xi,\beta)$ is a likelihood function, and sample information enters the estimation process through the likelihood function. Equation (9) can be expressed as:

$$f(\xi,\beta|x)\propto L(x|\xi,\beta)f(\xi)f(\beta) \tag{10}$$

$\propto$ indicates being proportional. The posterior distribution of parameters is:

$$f(\xi,\ \beta|x)\propto\ \xi^{-n+a_1-1}\beta^{-n+a_2-1}\exp\left[-b_1\xi-b_2\beta+(\frac{1}{\xi}-1)\sum_{i=1}^{n}\ln(1-\frac{x_i}{\beta})\right] \tag{11}$$

3. 5 Monte Carlo simulation of total loss distribution

Monte Carlo simulation method is used to study the total distribution of fraud losses. The frequency distribution and severity distribution of fraud losses are combined to study the total loss distribution of fraud of medical insurance for urban and rural residents (Ralf *et al.*, 2010, Wei *et al.*, 2018). Considering that the calculation of analytical expressions is tedious, therefore, the Monte Carlo simulation method is adopted here for research. First, Simulating the distribution function of fraud frequency for n times. Generating n random numbers and recording these numbers as $m_1,m_2,...,m_n$. Then, simulating the fraud amount losses for m_1 times and a total of m_1 fraud losses, namely $L_1,L_2,...,L_{m_1}$, can be obtained, totaling m_1 fraud losses. The obtained result is the simulation result S of the total fraud losses within a certain period of time. Repeating these steps on m_2 to m_n and n simulated losses S can be obtained. The distribution of fraud losses can be obtained according to the amount of n fraud losses. Then, based on the *VaR* model and fraud loss distribution function, the fraud losses at different confidence intervals can be obtained.

4. Data

The fraud loss data of basic medical insurance for urban and rural residents covering the period from 2006 to 2018 in China was selected[2]. The Ministry of Labor and Social Security issued the Notice on Special Expansion of Medical Insurance for Migrant Workers in China in 2006. Opinions put forward to promote

[2] Since the NCMS and medical insurance for urban residents began to integrate in 2016, the NCMS data and data of medical insurance for urban and rural residents from 2006 to 2015 were integrated to maintain the consistency of the sample data.

DOI: 10.24818/18423264/53.3.19.16

migrant workers to participate in medical insurance work in an all-round way. The implementation of this policy has led to a rapid increase in the number of insured persons and a marked increase in insurance fraud cases in the same year. In this study, all the data came from public media reports, and fraud cases and fraud loss data are collected through public information channels. The contents collected in each fraud case mainly include the time of fraud committing, the amount involved, the amount of losses, the frequency of fraud, the fraud perpetrator, etc. 347 pieces of news about the fraud of basic medical insurance for urban and rural residents are collected. To unify the data content and format, the fraud cases reported by various media are sorted out before the data are analyzed.

There are mainly four types of fraud perpetrators in practice: the insured of medical insurance for urban and rural residents, management personnel of medical insurance for urban and rural residents, medical personnel of designated medical institutions, and specialized fraud gangs. The fraud perpetrators in the statistical analysis are classified a bit differently as follows: First, the fraud committed through the use of a friend’s medical insurance card is regarded as fraud by the insured of medical insurance for urban and rural residents. Second, their use of medical insurance cards to commit fraud is regarded as fraud by specialized fraud gangs. Third, the fraud perpetrators are not the insured of medical insurance for urban and rural residents, and the fraudulent use of other people's medical insurance cards is regarded as fraud by specialized fraud gangs. Last for those not specified by news reports, they are recognized as specialized fraud gangs.

As for the fraud time, since the specific time of committing fraud in some cases is unknown and the fraud time is controversial, the following adjustments are made: First, if the fraud time is more than one year and the specific time cannot be determined, the mid-term time of the year involved should be taken as fraud time. Second, for fraud cases without reporting time of committing fraud, the fraud time can be inferred from other times reported by the media~~,~~ and shall be subject to the inferred time. If it is impossible to estimate, the fraud time shall be defined as two months before the time of fraud occurrence. Third, for fraud cases where the time of committing fraud and the time of fraud occurrence are unknown, the fraud time is defined as two months before the time of news reporting.

The definition of the frequency of frauds will be different. For research purposes, the following adjustments are made: if a report involves only one fraud case, it will be regarded as a fraud. If a news report covers more than one fraud case, the frauds frequency shall be calculated separately.

This study collects the amount of losses and suspected amounts of fraud cases. Among them, the fraud amount refers to the amount of losses, and refers to the specific losses of the medical insurance fund for urban and rural residents caused by fraud. The specific explanations are as follows: first, the fraud amount in the report is uncertain and is expressed by definite data (for example, the fraud amount loss is more than 200,000 yuan, and 200,000 yuan is taken). Second, if the fraud amount is not stated in the report but there is an amount involved, the amount involved shall be regarded as the fraud amount.

DOI: 10.24818/18423264/53.3.19.16

The frequency of frauds of medical insurance for urban and rural residents from 2006 to 2018, the amount of losses and the amount involved (as shown in Table 1).

Table 1. Data for fraud loss in different years

Years	Frequency of frauds (times)	The amount involved (million)	The amount of losses (million)
2006	9	276.27	224.40
2007	11	1,154.46	978.35
2008	17	596.13	559.72
2009	25	2,761.50	2,640.92
2010	21	2,103.20	1,932.09
2011	23	2,987.16	2,799.48
2012	40	4,401.29	4,224.70
2013	47	9,928.46	9,347.09
2014	49	9,642.39	8,887.10
2015	35	7,102.18	5,990.30
2016	30	5,472.16	4,085.32
2017	27	5,017.58	4,249.46
2018	13	3,691.04	2,259.45

As it can be seen from the table above, there were only 9 frauds in 2006, which were the fewest and accounted for 2.59% of the total frauds. There were 49 frauds in 2014, which were the most and accounted for 14.12% of the total frauds. Seen from the perspective of the amount of losses, the amount of fraud losses in 2006 was 2.224 million yuan, which was the least and accounted for 0.51% of the total amount of losses. The amount of fraud losses in 2013 was 93.471 million yuan, which was the most and accounted for 20.94% of the total amount of losses. The reason for the fewest frauds of insurance for urban and rural residents in 2006 is that the medical insurance system for urban and rural residents had not been popularized nationwide and the occurrence of fraud was not universal. However, the main reason for the serious fraud occurrence in 2013 is that serious cases of defrauding medical insurance funds for urban and rural residents occurred in some areas.

Shown from the statistical analysis of fraud cases of medical insurance for urban and rural residents from 2006 to 2018, the most common frauds are committed by a single type of perpetrator, with 264 cases, accounting for 76.08%. Seen from the amount of fraud, the amount of losses caused by a single type of perpetrator is also the largest, which indicates that the fraud perpetrators of medical insurance for

DOI: 10.24818/18423264/53.3.19.16

urban and rural residents are mainly concentrated in a certain type. Therefore, the following part focuses on studying the frauds with a single type of perpetrator.

For the frequency of frauds committed by the fraud perpetrator, the medical personnel of designated medical institutions committed 94 frauds, accounting for 35.61%. The medical personnel of designated medical institutions mainly defrauded residents' medical fund and hospitalization medical records by falsely listing charging items, exaggerating illness conditions, fabricating false hospitalization data, and borrowing patients' medical insurance cards to reimburse expenses in the name of hospitalization, even though patients were not hospitalized.

In terms of the amount of losses caused by fraud perpetrators, the largest amount of losses of the medical insurance funds was caused by the personnel of designated medical institutions. The amount reached 173.56 million yuan, accounting for 55.09%. The lowest amount of losses, as much as 41.29 million yuan, was caused by the insured, accounting for 13.11%.

In addition, among the cases with two types of fraud perpetrators, the fraud perpetrators causing serious losses of medical insurance funds for urban and rural residents are management personnel of designated medical institutions and medical insurance funds. According to statistics, from 2006 to 2018, there were 45 frauds causing a loss of 86.55 million yuan. It can be further concluded from the fraud loss data that the empirical data is obviously right-biased, with sharp peaks and thick tails. Its skewness coefficient is 8.68 and kurtosis coefficient is 85.50. To make the analysis results more convincing, resampling was further performed for 1000 times based on the Bootstrap method. The loss severity statistics obtained from the mean value of Bootstrap samples confirms that the data is obviously right-biased, with sharp peaks and thick tails.

The q-q figure fitted by fraud loss data also indicates that the loss data has obvious characteristics, including a sharp peak and thick tail. In addition, normal distribution, lognormal distribution, Weibull distribution, and generalized Pareto distribution(GPD) were used to fit the loss severity, and KS test was performed. When the significance level is 5%, only the fitting of the lognormal distribution, GPD, and Weibull distribution passed the test, out of which lognormal distribution was the best (KS=0.0618) and GPD was the second (KS=0.0710). Lognormal distribution has a good effect on estimating the high-frequency part. However, it has the defect of underestimating the probability of the tail extreme value, which may affect the estimation of the overall fraud risk. Whilst, the GPD distribution fits the tail well. Therefore, PSD-LDA model is used to define medical insurance fraud in stages, in which lognormal distribution defines the distribution part of LLHF and GPD defines the distribution part of HLLF in sections.

5. Empirical results

As highlighted in the figure of excess mean function (as shown in Figure 1), the figure tends to incline upward in a straight line starting from $u = 2.8$ million yuan. Seen from the Hill figure shown in Figure 2, the figure of shape parameters

DOI: 10.24818/18423264/53.3.19.16

tends to be stable, starting from the threshold value $u = 2.8$ million yuan. Therefore, the threshold $u = 2.8$ million yuan (corresponding $N_u = 44$) is selected. In theory, a higher threshold can be selected, but the excess number will be fewer, which reduces the accuracy of the parameters of probability distribution obtained by fitting data of low frequency with high loss cases.

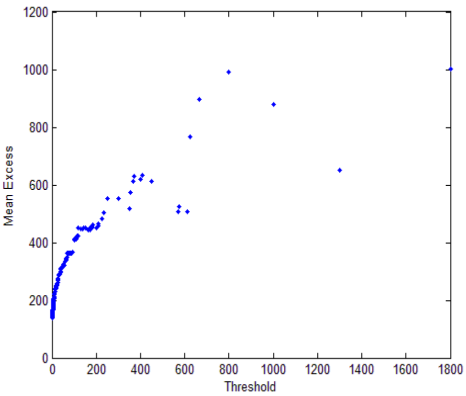

Figure 1. Excess Mean Function figure

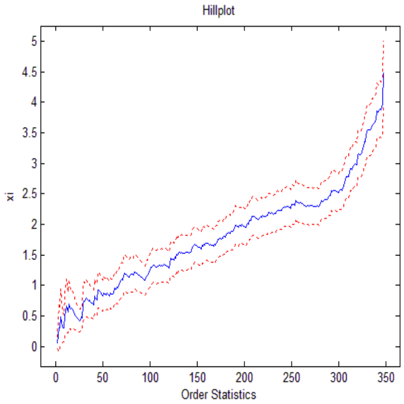

Figure 2. Hill figure

When threshold $u = 2.8$ million yuan, $N_u = 44$. GPD distribution parameters are obtained according to Equation (7). Excess distribution fitting figure (as shown in Figure 3) and q-q figure indicate that GPD has a good fitting effect, and it is appropriate to select the threshold $u = 2.8$ million yuan.

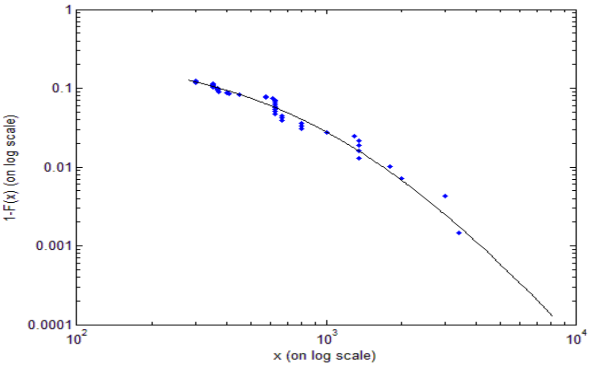

Figure 3. Excess Distribution Fitting figure

When $u = 2.8$ million yuan, put GPD distribution parameter δ=0.2712 and $\beta = 38.6588$ into Equation (5) to obtain the generalized Pareto distribution of the sequence of low frequency with high loss. Since δ=0.2712, GPD is thick-tailed. Its density distribution function is:

DOI: 10.24818/18423264/53.3.19.16

$$G_{\xi,\beta}(x) = 1-(1+\frac{0.2712(x-280)}{38.6588})^{-\frac{1}{0.2712}} \qquad \xi \neq 0 \tag{12}$$

The distribution function of 303 samples below the threshold is analyzed. Previous studies mainly use exponential distribution, normal distribution, lognormal distribution, and Weber distribution to fit the severity of fraud losses. As can be seen from the histogram of fraud losses of medical insurance for urban and rural residents (as shown in Figure 4), the amount of fraud losses from 2006 and 2018 exceeded 20 million yuan, and most of the fraud losses were concentrated within 1 million yuan. Therefore, the probability distribution of the fraud losses is fitted after taking the logarithm of the amount of fraud losses, which can achieve a better fitting effect. It can be seen from the q-q figure (as shown in Figure 5) that the fitting effect of the lognormal distribution is indeed ideal.

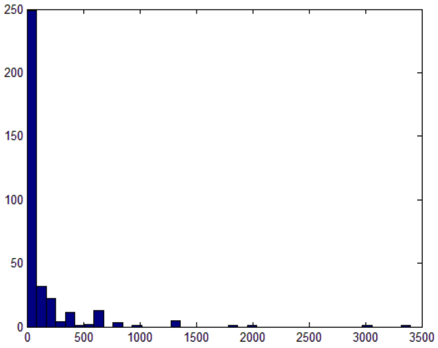

Figure 4. The Histogram of fraud losses of medical insurance

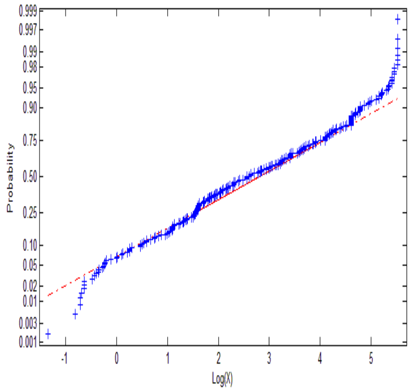

Figure 5. Log(X)~ normal distribution

Therefore, the lognormal distribution is selected to fit the probability distribution function. After further using maximum likelihood estimation, the distribution function of the variable X meeting $\{x \leq 280\}$ can be obtained:

$$x \sim Ln(2.6565, 1.1313^2) \tag{13}$$

The distribution function of fraud loss of medical insurance for urban and rural residents is obtained by combining Equation (12) and Equation (13) as follows:

$$F(x)=\begin{cases}\dfrac{1}{1.6313x\sqrt{2\pi}}\exp(-\dfrac{(Lnx-280)^2}{2\times1.6313^2}) & 0<x\leq 280\\ 1-(1+\dfrac{0.2712(x-280)}{38.6588})^{-\frac{1}{0.2712}} & x>280\end{cases} \quad (14)$$

In addition, Table 2 presents the estimated values of the parameters corresponding to the two-stage loss model when $u=2.8$.

Table 2. GPD-LDA model parameter estimation values

Threshold *u*	Excess number N_u	Low frequency with high loss			High frequency with low loss		
		Poisson	GPD		Poisson	Log norm	
		λ_2	ξ	β	λ_1	μ	σ
280	44	3.0769	0.2712	38.6588	23.6154	2.6565	1.6313

This paper uses the Monte Carlo simulation of the distribution of fraud losses of medical insurance for urban and rural residents from 2006 to 2018. As point out above, the frequency of fraud loss of medical insurance for urban and rural residents follows the Poisson distribution. The severity of fraud loss is defined by a two-stage model, in which the "high frequency with low loss" part follows the lognormal distribution and the "low frequency with high loss" part follows GPD distribution. Therefore, the optimal distribution of total loss is fitted by Monte Carlo simulation, according to the optimal distribution function of fraud loss frequency and fraud loss severity. The steps are as follows:

Step1: The random number m_1 of "low frequency with high loss" part and the random number m_2 of "high frequency with low loss" part which obeys Poisson distribution are respectively generated from the distribution function of fraud loss frequency of medical insurance for urban and rural residents, and are taken as the times of next iteration.

Step2: The distribution functions of "low frequency with high loss" part and "high frequency with low loss" part of frauds of medical insurance for urban and rural residents respectively generate m_1 random numbers and m_2 random numbers, subject to lognormal distribution and GPD distribution. The simulated fraud loss amount is processed and then accumulated and summed up in order to simulate the total fraud loss amount within one year, step 1 and 2 being repeated for 10,000 times to obtain 10,000 fraud loss sequence of observation.

Step3: The fraud loss sequences are sorted from smallest to largest to take different confidence α. 1000(1-α)+1 is taken as the estimated value of *VaR* and the average value of the sequences exceeding *VaR* is taken as the estimated value of *TVaR*.

DOI: 10.24818/18423264/53.3.19.16

According to PSD-LDA method and Monte Carlo simulation, the estimated values of *VaR* and *TVaR* when $u = 2.8$ are obtained.

Table 3. The $TVaR$ and VaR value of fraud risk loss

Method	$1\text{-}\alpha = 0.95$		$1\text{-}\alpha = 0.99$		$1\text{-}\alpha = 0.999$	
	VaR_{95}	$TVaR_{95}$	VaR_{99}	$TVaR_{99}$	$VaR_{99.9}$	$TVaR_{99.9}$
PSD-LDA	0.720	0.929	0.990	1.296	1.685	2.199
POT	0.072	0.122	0.181	0.305	0.532	0.898

Note: This table indicates the estimated value of *VaR* and *TVaR* when *u*=2.8.

As shown in Table 3, according to the calculation results of the two-stage model method, the maximum fraud loss of medical insurance for urban and rural residents within one year is 95% likely to be below 72 million yuan. Then, once this amount is exceeded, the average loss of the excess portion is 95% likely to be below 72 million yuan, and the maximum fraud loss within one year is 99.9% likely to be below 168.50 million yuan, and it is only 0.1% likely to exceed this amount. If it exceeds this amount, the average loss of the excess portion is 99.9% likely to be below 219.90 million yuan.

With the continuous implementation of the medical insurance system for urban and rural residents, the fraud risk is also increasing. Therefore, to ensure the effective operation of medical insurance system for urban and rural residents and to avoid problems such as fund turnover difficulties caused by unexpected excess of expenditures of medical insurance funds for urban and rural residents, part of medical insurance funds for urban and rural residents must be withdrawn as special reserve funds to deal with fraud risk losses. According to measurement results obtained by PSD-LDA method, there is a huge amount of risk fund for fraud. Therefore, *TVaR* value with a confidence level of 99.9% and a one-year period can be used as the fraud risk reserve of medical insurance for urban and rural residents. Since the current *TVaR* is 219.9 million yuan, the fraud risk reserve of the entire medical insurance for urban and rural residents is 219.9 million yuan.

To further study the fraud losses caused by each subdivided fraud perpetrator, the risk values of fraud losses caused by each fraud perpetrator under the confidence of 95% are calculated by the above-mentioned empirical research ideas, as shown in Table 4.

Table 4. Risk values of fraud losses caused by each fraud perpetrator

Type	The insured	Designated medical institutions	Insurance management personnel	Fraud gangs
Value	2579.7	12645.3	4487.9	1713.7

Note: Reliability $1\text{-}\alpha = 0.95$.

Among it, the loss risk value from the fraud of the insured is 25.80 million yuan.

DOI: 10.24818/18423264/53.3.19.16

The loss risk value from the fraud of the personnel of designated medical institutions is 126.45 million yuan. The loss risk value from the fraud of the medical insurance management personnel is 44.88 million yuan, and the loss risk value from the fraud of specialized fraud gangs is 17.14 million yuan. It reflects the great loss of the medical insurance fund caused by insurance fraud. It can be seen that the personnel of designated medical institutions are the main risk source of frauds of medical insurance for urban and rural residents.

6. Conclusions

Focused on the "low frequency with high loss" and "high frequency with low loss" of fraud risks of medical insurance for urban and rural residents, this paper uses a two-stage model based on Monte Carlo simulation to measure *TVaR* value of fraud risk losses of the medical insurance for urban and rural residents. Firstly, the severity and frequency of fraud losses of medical insurance for urban and rural residents from 2006 to 2018 are modeled respectively, and their optimal distribution functions are fitted. Then, according to the optimal distribution function of loss frequency and fraud loss severity, the optimal distribution of total losses is fitted by the Monte Carlo simulation method. After that, the annual provision for fraud risk losses of medical insurance for urban and rural residents is measured. The empirical results evidence that: First, the value of *TVaR* is significantly higher than the value of *VaR* in the same metric method and the same confidence levels, the value of *TVaR* and *VaR* in a high confidence level (99.9%) is higher than the low confidence level (95%), these phenomena demonstrating that the fraud of basic medical insurance has the characteristic of Tail-risks for urban and rural areas in China. Second, the maximum fraud loss of medical insurance for rural and urban residents within one year is 99.9% likely to be below 168.5 million yuan~~,~~ and is only 0.1% likely to exceed this amount. If it exceeds this amount, the average loss of the excess portion is 99.9% likely to be below 219.9 million yuan. Third, from 2006 to 2011, the national average fund-raising of the new rural cooperative medical system was 232.207 billion yuan[3], and 9.47‰ of the risk reserve was required to withstand the risk of fraud. Obviously, this analysis result could be applicable to fraud risk management of medical insurance for urban and rural residents: First, the result provides decision-making reference for scientific calculation of financing standards. Second, relevant statistical analysis highlights that the personnel of designated medical institutions cause the largest fraud loss, therefore, further improvement of the relevant medical reimbursement system is an effective means to avoid the fraud risks of medical insurance for urban and rural residents. Third, the calculation results can be considered for fraud risk pricing and fraud risk early warning.

[3] Data from the Social Security Center of the Ministry of Human Resources and Social Security, 2012.

Acknowledgments

We would like to thank an anonymous referee for many helpful comments. Thanks for the Project supported by the National Social Sciences Founding of China (No.18BGL200), the National Nature Science Foundation of China (No.71273148), the Key Research Projects on the Application Finance of Social Science Federation in Shandong Province (2017-JRZZ-04).

REFERENCES

[1] **Bodaghi A. & Teimourpour B. (2018),** ***Automobile Insurance Fraud Detection Using Social Network Analysis.*** *Application of Data Management and Analysis.* 5, 11-16;

[2] **Bermúdez, Pérez, Ayuso, Gómez, & Vázquez (2008),** ***A Bayesian Dichotomous Model with Asymmetric Link for Fraud in Insurance.*** *Insurance: Mathematics and Economics.* 2(42), 779-786;

[3] **Bauder, Khoshgoftaar & Seliya. (2017),** ***A Survey on the State of Healthcare up Coding Fraud Analysis and Detection.*** *Health Serv. Outcomes Res. Method.* 1, 31-55;

[4] **Brockett, Patrick & Derrig. (2012),** ***Fraud Classification Using Principal Component Analysis of RIDITs.*** *Journal of Risk and Insurance.* 69(3), 37-44;

[5] **Edelbacher M. & Theil M. (2016),** ***Fraud and Corruption in the Insurance Industry: An Austrian Perspective.*** *Fraud and Corruption.* 9, 161-179;

[6] **Saroliya A. & Kumar R. (2017),** ***Analyses and Detection of Health Insurance Fraud Using Data Mining and Predictive Modeling Techniques.*** *Soft Computing: Theories and Applications.* 11, 41-49;

[7] **Ekin T, Ieva F, Ruggeri F. & Soyer R. (2018),** ***Statistical Medical Fraud Assessment: Exposition to an Emerging Field.*** *International Statistical Review.* 86, 379-402;

[8] **Ghcan (2016),** ***The Global Challenge of Health Care Fraud.*** *http: //www. ghcan. org/challenge. Html,* 10-15*;*

[9] **Hassan A. & Abraham A. (2016),** ***Modeling Insurance Fraud Detection Using Imbalanced Data Classification.*** *Advances in Nature and Biologically Inspired Computing.* 11, 117-127;

[10] **Hubick (1992),** ***Artificial Neural Networks in Australia Department of Industry.*** *Technology and commerce.* 3, 56-89*;*

[11] **Jim (2012),** ***The Financial Cost of Healthcare Fraud.*** *Wfpha,*105-108*;*

DOI: 10.24818/18423264/53.3.19.16

[12] **Kong X., Yang Y., Gong F. & Zhao M. (2012). *Problems and the Potential Direction of Reforms for the Current Individual Medical Savings Accounts in the Chinese Health Care System.*** *Journal of Medicine & Philosophy.* 37(6), 556-576;

[13] **Lious & Riedinger (2014), *EFD: A Hybrid Knowledge Statistical Based System for the Detection of Fraud.*** *Journal of Risk and Insurance*. 69(3), 309-324;

[14] **Li Y, Yan C. & Wei L. (2018), *A Principle Component Analysis-based Random Forest with the Potential Nearest Neighbor Method for Automobile Insurance Fraud Identification*.** *Applied Soft Computing*. 1000-1009;

[15] **Lobont, O.R., Vatavu, S., Brindescu Olariu, D., Pelin, A. & Chis, C. (2019), *E-Health Adoption Gaps in the Decision-Making Process*.** *Revista de Cercetare si Interventie Sociala*. 65, 389-403;

[16] **Marisa S. & Nearhos J P. (1996), *Applying Data Mining Techniques to a Health Insurance Information System.*** *International Conference on Very Large Data Bases. Morgan Kaufmann Publishers Inc.* 286-294;

[17] **Punzo A., Bagnato L. & Maruotti A. (2018), *Compound Unimodal Distributions for Insurance Losses.*** *Insurance: Mathematics and Economics.* 81, 95-107;

[18] **Ralf K., Elke K. & Gerald K. (2010), *Monte Carlo Methods and Models in Finance and Insurance.*** Boca Raton. 1st Edition;

[19] **Sokol, Garcia B. & West M. (2001), *Precursory Steps to Mining HCFA Health Care Claims*.** *Hawaii International Conference on System Sciences*. IEEE;

[20] **Sadgali N. & Sael F. (2019), *Performance of Machine Learning Techniques in the Detection of Financial Frauds*.***Procedia Computer Science.*148,45-54;

[21] **Suliyanto, & Rifada. (2019), *Modeling of Risk for Diabetes Mellitus and Hypertension Using Bi-response Probit Regression*.** *Proceedings of the Third International Conference on Computing, Mathematics and Statistics (iCMS2017)* . 383-389;

[22] **Sheshasaayee A. & Thomas S.S. (2018), *Usage of R Programming in Data Analytics with Implications on Insurance Fraud Detection*.** *International Conference on Intelligent Data Communication Technologies and Internet of Things (ICICI).* 12, 416-421;

[23] **Tseng & Lu-Ming (2019), *Customer Insurance Frauds: The Influence of***

DOI: 10.24818/18423264/53.3.19.16

***Fraud Type, Moral Intensity and Fairness Perception*.** *Managerial Finance*. 45, 452-467;

[24] **Wenzel S., Nakajima S., Cunningham J., CLippert C. & Kloft M. (2017), *Sparse Probit Linear Mixed Model*.** *Machine Learning*. 106, 1621-1642.

[25] **Wang Y. & Xu E. (2018), *Leveraging Deep Learning with LDA-based Text Analytics to Detect Automobile Insurance Fraud*.** *Decision Support Systems*. 105, 87-95.

DOI: 10.24818/18423264/53.3.19.16