A Probabilistic Method for Computing Quantitative Risk Indexes from Medical Injuries Compensation Claims

S. Dalle Carbonare1; F. Folli2; E. Patrini3; P. Giudici4; R. Bellazzi1

1Dipartimento di Ingegneria Industriale e dell’Informazione, Università di Pavia, Pavia, Italy; 2Risk Manager, Azienda Ospedaliera di Lodi, Lodi, Italy; 3MARSH S.p.A., Milano, Italy; 4Dipartimento di Scienze Economiche ed Aziendali, Università di Pavia, Pavia, Italy

Keywords
Probabilistic models, Monte Carlo methods, Loss Distribution Approach, risk management

Summary
Background: The increasing demand of health care services and the complexity of health care delivery require Health Care Organizations (HCOs) to approach clinical risk management through proper methods and tools. An important aspect of risk management is to exploit the analysis of medical injuries compensation claims in order to reduce adverse events and, at the same time, to optimize the costs of health insurance policies. Objectives: This work provides a probabilistic method to estimate the risk level of a HCO by computing quantitative risk indexes from medical injury compensation claims. Methods: Our method is based on the estimate of a loss probability distribution from compensation claims data through parametric and non-parametric modeling and Monte Carlo simulations. The loss distribution can be estimated both on the whole dataset and, thanks to the application of a Bayesian hierarchical model, on stratified data. The approach allows to quantitatively assessing the risk structure of the HCO by analyzing the loss distribution and deriving its expected value and percentiles. Results: We applied the proposed method to 206 cases of injuries with compensation requests collected from 1999 to the first semester of 2007 by the HCO of Lodi, in the Northern part of Italy. We computed the risk indexes taking into account the different clinical departments and the different hospitals involved. Conclusions: The approach proved to be useful to understand the HCO risk structure in terms of frequency, severity, expected and unexpected loss related to adverse events.

1. Introduction

Over the last decade, the prevention of adverse events and medical injuries due to malpractice or suboptimal care delivery have become a worldwide concern [1–6] and the use of information technology solutions has been advocated [7] and put in practice by several authors [8, 9]. In recent years, the number of compensation claims for medical injuries have increased due to the greater expectation and awareness of patients. The analysis of medical injuries compensation claims is a very helpful way to estimate risk [10], because these data contain information about the department/unit where the injury happened, the estimated costs of the injury as well as the occurrence of the claims over the years [11–13]. Furthermore, compensation claims are readily available data, unlike data strictly related to clinical errors, which are hardly collected. In order to deal with the compensation claims submitted by patients, Health Care Organizations (HCOs) stipulate insurance policies, whose premiums are constantly growing, whereby for the HCO is very helpful have a precise assessment of its risk to reduce adverse events and better manage claims [14, 15].

Rather interestingly, the area of financial risk management offers several methods and tools that can be properly adapted to perform risk assessment starting from the analysis of compensation claims, with particular reference to Bayesian methods [16–20]. In particular, in this paper we propose a general method that extends the Loss Distribution Approach [21–23], an approach grounded on actuarial science able to quantify the expected loss of an HCO on the basis of an estimate of the frequency and the severity of adverse events. By the analysis of the loss distribution we compute a set of indexes, which can quantify the risk level of an HCO both at the whole organization level, and at each department/unit level. Such analysis can be useful to identify the more appropriate risk management actions. Our approach is aimed at providing a general framework...
that can be applied in different practical situations. For this reason, it combines a variety of methods through a procedure that allows the data analyst choosing the best strategy to deal with the data at hand. As an example, we herein show the results obtained by applying the new method to analyze the claims collected by the Lodi HCO, in the Northern part of Italy, which is in charge of managing four hospitals and several outpatients’ services.

In Italy, the National Healthcare System covers the healthcare needs for all eligible residents. In this system, the healthcare resources are distributed to each Italian region, that redistributes them to the Local Health Care Agencies, which supply healthcare services through a number of HCOs, such as in- and out-patients hospitals, or aggregations of such hospitals. Therefore, HCOs are the main institutions in charge to manage the clinical risks related to health service delivery.

2. Methods

As previously mentioned, this paper grounds on the so-called Loss Distribution Approach (LDA) [21–23], which allows describing the loss of an organization, in our case an HCO, through its probability distribution. Once the loss distribution is estimated, it is possible to compute quantitative risk indexes to identify the HCO risk level on the basis of the compensation claims.

The LDA is a well-known approach to estimate operational risk in banking and financial activities. This method estimates the loss probability distribution as the convolution of two distributions: i) the probability distribution, \( F \), of the injuries frequency and ii) the probability distribution, \( S \), of the injuries severity\(^a\). The loss distribution fully characterizes the losses of an HCO and can be used to derive summary statistics, such as the expected value, the median and the percentiles; these statistics can be considered as quantitative risk indexes of the HCO. One of the main quantities of interest computed in LDA analysis is the so-called “Value at Risk” (VaR) [25], which represents the worst expected loss at a certain confidence level. Such confidence level depends on the selection of i) the percentile of the loss distribution (usually the 99th), ii) the time interval considered for computing the loss probability distribution (usually one year), iii) the currency (in our case Euro).

The application of the LDA requires the definition of frequency and severity. In our case, a straightforward choice is to represent the severity as the amount of money that has been paid for each claim. The frequency values can be computed on the data dividing the total time span of interest, i.e. covered by the dataset, into equally sized intervals and by counting the number of requests claimed in each time interval.

The main steps of our approach consist therefore in:

1. estimating the frequency distribution \( F \) and the severity distribution \( S \);
2. computing the yearly distribution of loss \( L \);
3. deriving a set of risk indexes.

The next paragraphs will describe these steps with particular emphasis on the extensions of the method that we have developed to cope with step 1.

2.1 Estimating the Frequency and Severity Probability Distributions

The estimate of \( F \) and \( S \) can be performed in different ways; a proper strategy should take into account prior knowledge and the number and quality of available data. To this end, we have implemented a general approach which may adapt to different cases. Such strategy is based on the following steps:

1. parametric modeling:
   - for both frequency and severity two appropriate probability distributions are fitted from the available data by applying the maximum likelihood estimation of their parameters (parametric fitting);
   - the model goodness of fit is evaluated;
   - in case the parametric fitting is satisfactory, the distribution with the best adaptation to the data is chosen;
2. non-parametric modeling: it is fitted in case parametric modeling does not satisfy quality criteria.

In more detail, as regards frequency, we estimate two well known discrete distributions from the empirical frequency values: the Poisson and the negative binomial distribution. We apply the \( \chi^2 \) statistical test to evaluate their goodness of the fit. If the parametric fitting is satisfactory (i.e. the null hypothesis of perfect fit is not rejected for at least one distribution) we choose the distribution with the best fit, otherwise we exploit a semi-parametric fitting procedure based on multinomial distributions. According to this technique, the probability of each distinct frequency value \( f_k \) with \( k = 1, 2, \ldots, K \) is estimated as \( F(f_k) = n_k/n \), where \( n_k \) is the number of frequency values equal to \( f_k \) and \( n \) is the total number of frequency available. For example, if the whole dataset covers a time span of two years and we consider quarterly time intervals, we have \( n = 8 \) frequency values. Supposing that the available frequency values are \( \{0, 0, 1, 2, 2, 3, 4\} \), we have \( K = 5 \) distinct frequency values and \( f_k = \{0, 1, 2, 3, 4\} \). By fitting a multinomial distribution, the probabilities related to the frequency values are therefore \( f_k, F(f_k) = \{(0, 2/8), (1, 1/8), (2, 3/8), (3, 1/8), (4, 1/8)\} \).

As regards severity, we exploit the exponential as well as the lognormal distributions (the latter computed only for the positive samples). In analogy with the procedure described above, we perform the Kolmogorov-Smirnov statistical test to evaluate the goodness of fit. If at least one of the parametric model is acceptable (i.e. the null hypothesis is not rejected for at least one model) we choose the model with the best fit, otherwise we apply a non-parametric strategy based on the kernel density estimation [26, 27]. According to this approach, the severity probability distribution can be estimated from the available severity values \( s_i \) with \( i = 1, \ldots, n \) as:

\[ s_i = \frac{1}{n} \sum_{j=1}^{n} \delta(s_j - s_i) \]

\( \delta \) is the Dirac delta function.
$S(s) = \frac{n_1(s)}{n\lambda}$

Where $n_1(s)$ is the number of points within a $\lambda$-sized neighborhood of $s$, normalized by the total number of severity data $n$ multiplied by the neighborhood size $\lambda$. In order to obtain a smooth distribution, a Gaussian kernel function can be used, so that the observations around $s$ are decreasingly weighted with the increase of the distance between the observation and $s$. By applying a Gaussian kernel $\gamma_j(\cdot)$ with mean and standard deviation equal to 0 and $\lambda$, respectively, the severity density probability estimation becomes:

$S(s) = \frac{1}{n} \sum_j \gamma_j(s - s_j)$

The approach for the estimation of frequency and severity distributions described above can lead to poor estimates when considering very small datasets. When performing a stratified analysis, the dataset is divided in several groups, so the size of the data for the single groups may strongly decrease, especially for the frequency values. In order to obtain meaningful results even in this case, in stratified analyses we provide the possibility to apply the previous approach or to resort to a robust strategy based on Bayesian hierarchical modeling [17, 28]. These models are based on the idea that the strata are not completely independent, but, on the contrary, they are part of a wider population with which they share some informative content. The probability distributions for each data subset are hence computed by considering both the data of the single groups and the dataset. To obtain the severity probability distribution related to each group we estimate the parameter $\theta^j$ as:

$\hat{\theta}^j = \frac{1}{\left(\sigma^j\right)^2 + \tau^2} \bar{s}^j + \frac{1}{\left(\sigma^j\right)^2} \bar{\tau}^j$

where $\mu$ and $\tau^2$ are respectively the mean and the variance of $\bar{\theta}^j$, while $\bar{s}^j = \frac{1}{n^j} \sum s^j$ and $(\sigma^j)^2 = \frac{1}{n^j} \sum (s^j - \bar{s}^j)^2$ are the mean and variance of the $j$-th group of data. The parameters $\mu$ and $\tau^2$ can be estimated as:

$\hat{\mu} = \frac{\sum_j \frac{1}{\left(\sigma^j\right)^2 + \tau^2} \bar{s}^j}{\sum_j \frac{1}{\left(\sigma^j\right)^2 + \tau^2}}$

and $\hat{\tau}^2 = \frac{\sum_j (\bar{s}^j - \bar{s})^2}{J - 1}$

The probability distribution of the severity related to each group $j$ is thus estimated as a lognormal distribution, with an associated normal distribution characterized by mean and variance equal to $\bar{s}^j$ and $\sigma^2$, respectively, computed taking into account both the data of each group and the dataset as a whole.
2.2 Estimating the Loss Probability Distribution

As previously reported, the loss distribution $L^y$ is obtained as the convolution of the two probability distributions $F$ and $S$. Given that the convolution integral cannot be analytically solved, an approach widely used to estimate the loss distribution is to resort to Monte Carlo simulations [29, 30]. Under the assumptions that $F$ and $S$ are known, the Monte Carlo approach estimates $L^y$ on the basis of simulated scenarios obtained by sampling values extracted from $F$ and $S$.

The $u$-th scenario is drawn following these steps:

1. one frequency value, say $f^y$, is sampled from the estimated frequency distribution $F$;
2. $f^y$ severity values are sampled from the estimated severity distribution $S$: $\tilde{s}_{uv} = \{\tilde{s}_{u1}, \tilde{s}_{u2}, \ldots, \tilde{s}_{uj}\}$
3. the sum of the $\tilde{s}_{uv}$ is the loss related to the $u$-th scenario: $l^y_u = \sum_{v=1}^{n} \tilde{s}_{uv}$.

These steps are repeated many times and then $L$ is estimated as the sampling distribution obtained considering all the $\{l^y_u\}$ values generated. The loss distribution $L$, computed as above, characterizes the losses in time ranges of width $T$, chosen in order to compute the frequency distribution $F$. The choice of $T$ is a trade-off between the number of frequency samples and their accuracy, and is often a fraction of the year. Therefore, to compute the yearly loss distribution $L^y$, which is usually the most interesting one from the decision-makers viewpoint, we simply draw samples from $L$ the number of times needed to cover the yearly time span. For example, if we use a three-month time interval for the frequency computation in order to obtain a value sampled from the yearly loss distribution $L^y$, we must sum four values drawn from the quarterly loss distribution $L$.

2.3 Computing Risk Indexes

Once obtained the yearly loss distribution $L^y$, we compute a set of quantitative risk indexes such as:

- VaR, i.e. the 99th percentile of $L^y$,
- the expected loss $E[L^y]$, 
- the median of $L^y$,
- the unexpected loss ($UL$), which is equal to the difference between VaR and expected loss.

Moreover in stratified analyses, besides the same indexes calculated above, we also compute two additional indexes. The first is the ratio between the VaR and the median of the loss distribution, i.e. the ratio between the 99th and the 50th percentile, which gives information about the loss distribution shape. The second, which applies to the hospital departments only, is the ratio between unexpected loss and number of available beds for each group; this allows comparing the results of the compensation claims analyses normalized with respect to the amount of activity of the different clinical departments.

3. Data and Settings

The general approach described above has been applied to a data set coming from a database of medical injury compensation claims collected by the Health Care Organization of Lodi, in the Northern part of Italy. The dataset contains 206 cases of injuries with compensation requests collected from 1999 to the first semester of 2007. The claims have 33 features, which can be divided into different categories:
- features related to the claim, like Paid Amount, Claim Status, Claim Date, etc;
- features related to the injury, such as Risk Area, Location Code, Injury Date, Injury Description, Consequences, etc;
- administrative codes, like Insurance Contract, Injury Code, etc.

When applying our method to this dataset, we represent the severity with the Paid Amount, while the frequency has been computed as the number of compensation requests claimed to the HCO within a fixed interval of time, chosen to be equal to three months. The maximum and the average paid amount values were 315 K€ and 13 K€ respectively while the 5th and 95th percentiles were 0 and 73 K€, respectively. As we can deduce from these data, there are several claims (63) whose Paid Amount is 0. These cases are correspondent to claims with no compensation provided by the HCO. We included these claims in the analysis because they characterize the frequency and severity distributions of the compensation requests, representing low severity cases. We have performed both a global analysis considering the whole data set and two stratified analyses based on the attributes Risk Area, which represents the

<table>
<thead>
<tr>
<th>Table 1</th>
<th>Number of compensation claims and paid amount for Risk Area</th>
</tr>
</thead>
<tbody>
<tr>
<td>Risk Area</td>
<td>Number of Claims</td>
</tr>
<tr>
<td>Surgery</td>
<td>78</td>
</tr>
<tr>
<td>Obstetrics and gynecology</td>
<td>21</td>
</tr>
<tr>
<td>Medicine</td>
<td>41</td>
</tr>
<tr>
<td>First aid</td>
<td>38</td>
</tr>
<tr>
<td>Services</td>
<td>14</td>
</tr>
<tr>
<td>Others</td>
<td>14</td>
</tr>
<tr>
<td>Total</td>
<td>206</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Table 2</th>
<th>Number of compensation claims and paid amount for Location Code</th>
</tr>
</thead>
<tbody>
<tr>
<td>Location Code</td>
<td>Number of Claims</td>
</tr>
<tr>
<td>Hospital 1</td>
<td>120</td>
</tr>
<tr>
<td>Hospital 2</td>
<td>21</td>
</tr>
<tr>
<td>Hospital 4</td>
<td>22</td>
</tr>
<tr>
<td>Hospital 2</td>
<td>32</td>
</tr>
<tr>
<td>Others</td>
<td>11</td>
</tr>
<tr>
<td>Total</td>
<td>206</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Table 3</th>
<th>Results of global analysis</th>
</tr>
</thead>
<tbody>
<tr>
<td>Result among 10 trials</td>
<td>VaR (€) Yearly</td>
</tr>
<tr>
<td>Mean</td>
<td>2,014,070</td>
</tr>
<tr>
<td>Standard deviation</td>
<td>136,415</td>
</tr>
</tbody>
</table>
hospital department involved in the injury, and Location Code that identifies which, among the four hospitals and the outpatients’ services managed by the Lodi HCO, was involved in the specific injury.

Tables 1 and 2 show the number of compensation claims and the total amount paid for each Risk Area and Location Code, respectively.

Both in global and in stratified analyses we have chosen the same number of simulations to estimate the loss distribution with the Monte Carlo technique and the same time interval for the computation of frequency values. In accordance to the literature [31], we have chosen 15,000 simulations and a time interval of three months. Moreover, in order to obtain more reliable results, in the global analysis the quantitative risk indexes have been computed by averaging the results of 10 Monte Carlo trials.

### 4. Results

In this section, we show how the general method presented in this paper can be applied to a concrete case to gain insight into an HCO risk structure. In the global analysis, the Poisson distribution leads to poor estimation of the frequency model (very small p-value in the \( \chi^2 \) test), while the negative binomial turned out being satisfactory, so it has been preferred over the alternatives. As regards severity, we have chosen the lognormal distribution which shown a good adaptation to the data, while the exponential distribution results in a inadequate fitting (with very small p-value in the Kolmogorov-Smirnov test).

The results of the global analysis, presented in Table 3 show that, by applying our method to estimate the yearly loss distribution of the HCO, the expected loss is about 500 K€, the Value at Risk is about 2 M€, while the unexpected loss is about 1.5 M€. The yearly insurance premium paid by the HCO is about 1.8 M€ and it seems comparable to the computed VaR that represents the worst expected loss. The HCO can use these results to thoroughly understand its risk level and to better manage the compensation claims, for example by allocating an insurance exemption equal to the expected loss and by insuring only the unexpected loss.

In the stratified analyses, as regards the estimation of the frequency distribution, both the negative binomial and the Poisson distributions perform poorly, so we have applied the discrete Bayesian hierarchical model. As concerns severity, the lognormal distribution consistently outperformed the exponential one. For the sake of brevity, we present only a high level view of the results.

A simple strategy to evaluate the risk profile of an institution or of an area is to i) evaluate the VaR, which provides the worst case scenario, and thus highlights the potentially high risk cases, and ii) consider the ratio between the VaR and the median of the distribution, which shows if the distribution is skewed (few events with high loss) or not (several events with medium/high loss).

In Table 4 we observe that the obstetrics and gynecology area has the highest VaR and the highest UL/Number of beds ratio, so it is easy to conclude that this area is the one with the highest associated risk. Moreover, obstetrics and gynecology has also a high VaR/loss median ratio: this means that the 99th percentile is quite far from the 50th percentile, i.e. the loss distribution is characterized by rare but “catastrophic” events. In fact, the histogram of \( L \), in Figure 1, presents the majority of

### Table 4

Results of the stratification by Risk Area

<table>
<thead>
<tr>
<th>Risk Area</th>
<th>Yearly VaR (€)</th>
<th>UL (€)</th>
<th>Loss Mean (€)</th>
<th>Loss Median (€)</th>
<th>VaR/Loss Median (€)</th>
<th>UL/Number of beds (€)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Obstetrics and gynecology</td>
<td>1,691,000</td>
<td>1,541,900</td>
<td>149,100</td>
<td>42,658</td>
<td>40</td>
<td>22,027</td>
</tr>
<tr>
<td>Surgery</td>
<td>1,129,700</td>
<td>927,130</td>
<td>202,570</td>
<td>134,030</td>
<td>8</td>
<td>3,622</td>
</tr>
<tr>
<td>Medicine</td>
<td>689,010</td>
<td>614,170</td>
<td>74,840</td>
<td>32,353</td>
<td>21</td>
<td>1,310</td>
</tr>
<tr>
<td>Services</td>
<td>420,690</td>
<td>385,917</td>
<td>34,773</td>
<td>6,378</td>
<td>66</td>
<td>N/A</td>
</tr>
<tr>
<td>First aid</td>
<td>379,470</td>
<td>317,310</td>
<td>62,160</td>
<td>37,941</td>
<td>10</td>
<td>N/A</td>
</tr>
<tr>
<td>Others</td>
<td>81,471</td>
<td>67,424</td>
<td>14,047</td>
<td>8,327</td>
<td>10</td>
<td>503</td>
</tr>
</tbody>
</table>

### Figure 1

Loss distribution for the obstetrics and gynecology Risk Area
values concentrated on the left of the graph, but it has a long right tail related to few events with very high compensations. The loss histogram of risk areas with a low ratio between VaR and median, like surgery, in Figure 2, presents instead more uniform compensations, which are, on average, high. Although obstetrics and gynecology is more risky than surgery, intervening on the surgery healthcare process has the potential of bringing a greater reduction in the number of injuries, since the higher incidence of injuries, although with relatively low loss, could be due to suboptimal care during the day-by-day activities.

As regards the Location Code attribute, we can see in Table 5 that hospital 3 has a slightly lower VaR than hospital 1 but it has the highest UL/Number of beds ratio and a high ratio between VaR and loss median. Looking at the loss histograms, we can see that hospital 3, in Figure 3, is characterized by a small number of injuries with very high compensations, while, in the case of hospital 1, the high VaR and a low VaR to median ratio is due to a large number of quite high compensations, as shown in Figure 4. Therefore, we can conclude that the two hospitals need different types of intervention: hospital 3 needs to improve the management of rare, but complex, events, while hospital 1 needs to identify and reduce systematic procedural errors.

5. Discussion

Risk management techniques are steadily spreading in healthcare activities, with the aim to improve the quality and safety of the services supplied to patients and to reduce adverse events and their related costs. An efficient activity of risk management is made of different phases [1]: the knowledge and the analysis of errors, the detection and the correction of their causes, the application and monitoring of planned actions.

For each phase there are several methods that can be applied. Errors can be identified with tools like incident reporting, case history or patient record revision, analysis of complaints and claims or administrative databases [32]; process analysis and action planning can instead use
root cause analysis (RCA) [33], failure mode and effects analysis (FMEA) [34], failure modes, effects and criticality analysis (FMCA) [35].

The detection of medical errors is often difficult because it is still related to the fear of punishment or legal actions [36]. Incident reporting systems are useful to collect reports of errors or potential errors, to improve the quality and safety of clinical procedures, but they also need a lot of time and resources to be implemented and examined. On the contrary, compensation claims data are routinely collected by HCOs or by insurance companies. These data cannot be used to cover all aspects of risk analysis, since they are collected for reimbursement purposes and therefore they may report predominantly the most serious cases or those resolved through litigation. However, these data can be used to estimate the HCO economic loss due to adverse events and to assess the HCO risk level. LDA analysis allows deriving indexes that can be properly exploited to better manage compensation claims and negotiate lower insurance premiums. Moreover, the indexes computed in stratified analyses can be used to identify the departments with high risk and to better understand their risk structure, thus assessing the priority and type of intervention for risk mitigation.

6. Conclusions

In this paper we show the application of a probabilistic approach, called Loss Distribution Approach (LDA) to estimate the loss distribution and a set of risk indexes for an HCO by data coming from medical injuries compensation claims sent to it. The method, adapted from operational and financial risk management, has been enriched by a new strategy to estimate the probability distribution needed and in particular by the use of Bayesian hierarchical models to achieve better results on small data subsets.

Our approach proved to be useful to analyze the data coming from 206 cases collected over eight years. The analyses have been performed both on the whole dataset and on subgroups of data obtained by stratification over a set of interesting variables.

The method proposed can be used by an HCO to estimate, based on a statistical and probabilistic approach, its risk levels considering both the HCO as a whole and specific departments or areas of interest. Moreover, the knowledge of the risk structure is the first step for an HCO to improve the management of its compensation claims, in a context of ever-growing health insurance premiums that need new methods for the management of clinical risk.

Acknowledgments

This work is based upon the short paper [37] presented at the MIE 2009 conference.

References