A Methodology for the Extraction of Quantitative Risk Indexes from Medical Injuries Compensation Claims

Simona DALLE CARBONARE\textsuperscript{a,b,1}, Fulvia FOLLI\textsuperscript{c}, Emanuele PATRINI\textsuperscript{c,d}, Riccardo BELLAZZI\textsuperscript{a}

\textsuperscript{a} Dipartimento di Informatica e Sistemistica, University of Pavia, Italy
\textsuperscript{b} Institute for Advanced Studies (IUSS), Pavia, Italy
\textsuperscript{c} Unità di Risk Management, Azienda Ospedaliera of Lodi, Italy
\textsuperscript{d} Healthcare Risk Management, MARSH S.p.A., Milano, Italy

Abstract. The prevention of adverse events and medical injuries due to malpractice or suboptimal delivery of health care services is one of the major concerns of citizens and Health Care Organizations. One way to understand adverse events is to analyze the compensation requests for medical injuries that are claimed to hospital or health care services. In this paper we describe the results obtained by applying a probabilistic model, called the actuarial model, to analyze 317 cases of injuries with compensation requests collected from 1999 to the first semester of 2007 by the Azienda Ospedaliera (A.O.) of Lodi, in the Northern part of Italy. The approach, adapted from operational and financial risk management, proved to be useful to understand the risk structure in terms of frequency, severity, expected and unexpected loss related to adverse events.

Keywords. risk management, medical injuries, probabilistic models, Monte Carlo simulations, actuarial model

1. Introduction

The prevention of adverse events and medical injuries due to malpractice or suboptimal delivery of health care services is one of the major concerns of citizens and Health Care Organizations (HCO). The main goal of risk management in health care is to lower the frequency and the impact of adverse events by monitoring them, understanding their reasons and implementing suitable strategies to avoid them. One way to understand adverse events is to analyze the compensation requests for medical injuries that are claimed to hospital or health care services. Mining and understanding such data is a way to learn what are the main reasons for the adverse events, to plan actions to mitigate risk, and to properly negotiate the contract with insurance companies. In this paper we describe the results obtained by applying the so-called “actuarial model” [1–4] to analyze the claims collected by the Azienda Ospedaliera (A.O.) of Lodi, in the Northern part of Italy. The A.O. of Lodi is a HCO which handles four hospitals and

\textsuperscript{1} Corresponding Author: Simona Dalle Carbonare, Dipartimento di Informatica e Sistemistica, Università degli Studi di Pavia, Via Ferrata 1, 27100 Pavia, Italy; E-mail: simona.dallecarbonare@gmail.com.
several outpatients’ services. The actuarial model allowed us to estimate the probability
distribution of the frequency and severity of adverse events, and to compute some
useful indexes that can be applied to compare hospitals and to better understand the
structure of the HCO risk.

2. Materials and Methods

2.1. Data

We analyzed the data coming from a database of medical injury compensation claims
collected by the A.O. of Lodi, Italy. The dataset contains information about 317 cases
of injuries with compensation requests collected from 1999 to the first semester of
2007. The tuples have 33 features which can be divided into different categories:

- variables related to the claim like Paid Amount, Request Status, Request Date,
  Claim Closing Date, etc;
- variables related to the injury, such as Risk Area, Location Code, Injury Date,
  Injury Description, Consequences, etc;
- administrative codes like Insurance Contract, Injury Code, etc.

The feature Request Status can have the value Open, in case the claim did not (yet)
result in a compensation or Closed if the request had been already closed and
compensation had been decided. In the dataset there are 110 open and 207 closed
claims of which only one has a missing value for Paid Amount. The minimum value of
Paid Amount is 0 (in case of a closed claim with no compensation), while the
maximum value is 315,000 €; the mean value is 13,026 €; the 5th and 95th percentile are
respectively 0 and 73,165 €.

When dealing with hospitals it is very interesting to stratify the data with respect to
some key features, to extract risk indices useful in clinical governance and in risk
management, thus allowing to intervene to prevent further injuries or to better
determine the contract with insurance companies. In this paper we selected the
Location Code, which identifies one of the four structures managed by the Lodi A.O.,
and Risk Area, which represents the department involved in the injury.

2.2. Methods

In this paper we show the application of a method to extract quantitative risk indices
from claims and compensation requests based on the so-called actuarial model [1–4].

The actuarial model is a well known approach to estimate marketing risk in
banking and financial activities. The main quantity of interest computed by this model
is called Value at Risk (Var) [5, 6]. Var represents the maximum expected loss given a
confidence level, a time interval and a currency. The actuarial model derives Var as a
percentile, generally the 99th, of the probability distribution of yearly losses (L). L may
be obtained by convolution of the probability distribution of the frequency of an injury
(F) and the probability distribution of the severity of the injury (S). Given the
complexity of the analytical solution of the convolution integral, an approach widely
used to estimate the losses distribution is to resort to Monte Carlo simulation [7–9]. By
applying a Monte Carlo simulation, \( L \) is estimated on the basis of simulated scenarios obtained by sampling values extracted from \( F \) and \( S \). In the actuarial model, the \( F \) and \( S \) distributions are usually estimated from data with standard parametric fitting. Once \( L \) is obtained, it is possible to estimate its sample mean, median and the so-called “Unexpected Losses” (UL) defined as the Var minus the mean. Being a loss a monetary value, the UL is of particular interest for determining the amount of the insurance premium.

In our work, the severity has been represented by the paid amount of a compensation claim and the frequency has been computed as the number of claims to the hospital in a certain interval of time. Therefore, the severity is a continuous variable, whose probability distribution can be modeled with a log-normal density, while the frequency is a discrete variable, which probability distribution can be modeled with a Poisson density.

We applied the actuarial model both to the entire data set and to the data sets obtained after stratification for Location Code and Risk Area. However, when performing stratified analysis, the number of data for each subset is often very small so it was not possible to perform a standard parametric fitting. Therefore we relied on two different approaches to derive the probability distributions \( F \) and \( S \). Firstly, we checked if parametric fitting was possible by using standard statistical tests for probability density comparison (either the \( \chi^2 \) in case of discrete distributions or Kolmogorov-Smirnov for continuous ones); if not, we resorted to non parametric probability density estimates (such as Kernel Density Estimation [10]). Secondly, we considered the use of Bayesian “population” models to fit the probability distributions. In this case the stratified data are not considered as isolated groups but as members of a wider population, the whole dataset. The parameters of the distribution of each group are therefore estimated taking into account both its data subset and the whole dataset, gaining in results accuracy and significance. In particular, for the severity we applied to a logarithmic transformation of the data a continuous model based on the Gaussian distribution [11], while for the frequency we used a discrete model [12] based on the claims count for each category adjusted by population parameters\(^2\).

3. Results

The developed methodology has been applied both to the whole dataset and to data stratified with respect to Location Code and Risk Area. Our analysis considered only the closed claims, in our case 206.

By literature search and experimental trials and tests, we set for both kinds of analysis 15,000 simulations and a three-month time interval to compute the injuries frequency. Moreover, to obtain more meaningful results in non stratified analysis, we computed the sample mean and standard deviation of the extracted indices on 10 different trials. The results about non stratified analysis are shown in Table 1. In particular, Table 1 shows that the estimate of the expected loss of the A.O. was around 500 k€, while the unexpected loss is three times higher. This result is quite interesting as 500 k€ is what the A.O. may expect to pay every year to compensate claims; a suitable option to save insurance policy costs is to stipulate contracts with insurance

\(^2\) The methods and simulations have been implemented in Matlab.
companies that cover only the UL, dealing with the expected ones with a fund handled by the A.O. administration itself.

<table>
<thead>
<tr>
<th>Mean result among 10 trials</th>
<th>Yearly Var €</th>
<th>Unexpected Losses (UL) €</th>
<th>Losses Mean €</th>
<th>Losses Median €</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean</td>
<td>2,014,070</td>
<td>1,519,810</td>
<td>494,273</td>
<td>389,748</td>
</tr>
<tr>
<td>Standard deviation</td>
<td>136,415</td>
<td>128,418</td>
<td>12,779</td>
<td>5,151</td>
</tr>
</tbody>
</table>

Table 2. Results about Location Code (15,000 simulations, three-month time interval for frequency)

<table>
<thead>
<tr>
<th>Location Code</th>
<th>Yearly Var €</th>
<th>Unexpected Losses (UL) €</th>
<th>Losses Mean €</th>
<th>Losses Median €</th>
<th>Var/Losses Median</th>
<th>UL/number of beds</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hospital 1</td>
<td>1,412,000</td>
<td>1,150,800</td>
<td>261,210</td>
<td>178,240</td>
<td>8</td>
<td>2,615</td>
</tr>
<tr>
<td>Hospital 2</td>
<td>122,860</td>
<td>87,719</td>
<td>35,143</td>
<td>29,746</td>
<td>4</td>
<td>467</td>
</tr>
<tr>
<td>Hospital 3</td>
<td>1,334,300</td>
<td>1,176,600</td>
<td>157,770</td>
<td>52,800</td>
<td>25</td>
<td>12,517</td>
</tr>
<tr>
<td>Hospital 4</td>
<td>441,830</td>
<td>387,550</td>
<td>54,282</td>
<td>25,157</td>
<td>18</td>
<td>1,872</td>
</tr>
<tr>
<td>Others</td>
<td>91,766</td>
<td>82,098</td>
<td>9,669</td>
<td>2,610</td>
<td>35</td>
<td>N/A</td>
</tr>
</tbody>
</table>

In stratified analysis we have chosen the parametric fitting for severity, which in general has enough values, and the hierarchical Bayesian model for frequency, which usually has a small number of values because it is computed on time intervals. Moreover the categories with less than ten cases have been grouped in the Others category to obtain more significant results.
The results for Location Code and Risk Area are respectively shown in Tables 2 and 3. Figures 1 and 2 show the histograms of losses distributions for the hospitals with the lower and the higher ratio between Var and losses median for Location Code.

Looking at Table 2, it is interesting to note that there are great differences in the UL between the different hospitals managed by the A.O. The column that reports the ratio between the UL and the number of beds of the hospital, clearly shows that Hospital 3 has a much higher risk than all the other hospital, being UL around 12 k€ per bed. Also Figure 2 shows that the Hospital 3 has a long tail loss distribution, with non-zero probability that very high losses occur. This is clear also by comparing it with the distribution of Hospital 2 (Figure 1).

Finally, Table 3 reports the results obtained on the data of the different risk areas. The Obstetrics and Gynaecology area is largely the most risky one, being its UL over number of beds ratio more than ten times higher than the other areas.

4. Conclusions

In this paper we have shown the application of a probabilistic approach to perform a risk analysis on the data coming from medical injuries compensation claims. The approach, adapted from actuarial analysis performed in operational and financial risk by banks and insurance companies, proved to be useful and successfully dissected the data coming from 317 cases collected over eight years. We also performed a number of other different analysis, ranging from time series evaluation to survival curves; such analysis are not reported in this short paper, but their results, together with the ones presented in this paper, are now used by the A.O. to identify the major risk sources. This is the first step towards the definition and application of effective actions to prevent adverse events.

References