Digital Personalized Health and Medicine
L.B. Pape-Haugaard et al. (Eds.)

doi:10.3233/SHTI200208

Using Big Data Analytics to Identify Dentists with Frequent Future Malpractice Claims

Wanting CUI[1a], Joseph FINKELSTEIN[a]
[a]*Icahn School of Medicine at Mount Sinai, New York, NY, USA*

Abstract. Healthcare spending has been growing at an increasing rate in the US, due in part to medical malpractice costs. Dental malpractice is an area that has not been studied in depth. Using National Practitioner Data Bank (NPDB), we explored the extent of dental malpractice claims and sought to construct a predictive model that can help us identify dental practitioners at risk of performing medical malpractice. Over 1,500 dental malpractice claims were reported annually, and over $1.7 billion being paid out by medical malpractice insurers over the past 15 years. Majority of claims resulted in minor injuries, and the number of major injury claims increased over years. In prediction, we randomly split the data into train (75%) and test (25%) datasets. We trained and tuned models using 5-fold cross validation on the training set. Then, we fitted the model on the test data for performance measures. We used Logistic Regression, Random Forest (RF) and XGBoost and tuned the hypermeters of models accordingly through grid search and cross validation. XGBoost was the best machine learning model to predict the risk of dentists having several malpractice reports. The best performing model had an accuracy of 72.8% with 30.6% F1 score. The NPDB database is a valuable dataset to study dental malpractice claims. Further analysis of information extracted from this dataset is warranted.

Keywords. Data Science, Big Data Analytics, Machine Learning, Predictive Model

1. Introduction

The United States currently ranks highest in health care spending among the developed nations of the world. In 2017, health care spending was $3.5 trillion in total, which accounted for a 17.9% share of the nation's GDP. It has also increased over 50% in the past 10 years, compared to $2.3 trillion in 2007 [1]. One of the reasons for the increasing cost of health care is defensive medicine. Defensive medicine refers to physicians prescribe diagnostic test or medical treatment that depart from normal medical practice as a safeguard from litigation [2]. According to a recent study, there is a statistically significant correlation between specialists concerns regarding potential medico-legal disputes and the choice of defensive medical procedures [3].

Although the number of malpractice claims has decreased since 2001, they are still very common in the United States. According to a 2016 Benchmark Survey, a third of

[1] Wanting Cui, Icahn School of Medicine at Mount Sinai, 1770 Madison Ave, New York, NY, USA, 10035, E-mail: wanting.cui@mssm.edu

physicians have been sued at least once in their career [4]. In addition, the average costs of malpractice cases have increased year over year. The malpractice insurance premiums have also increased accordingly, due to the high cost of malpractice claims. As a result, physicians' incomes may be affected by the rising cost of doing business as well.

While physicians' (MD, DO) malpractice cases have been frequently studied in recent years, there is little study on dental malpractice disputes and predictive models on future prevention are rarely constructed [5, 6]. In this article, using data from the National Practitioner Data Bank, we reported findings concerning the distribution of dental malpractices and the extent of dental malpractice risk. By studying malpractice, we hope to mitigate the high insurance premiums caused by the prevalence of malpractice claims. Our objectives were to review recent trends in dental malpractices, based on cost, number of claims, severity of injuries, physicians' and plaintiffs' information; and to create machine learning models predicting dentists with higher risk of frequent malpractices based on their first malpractice claim.

2. Method

2.1. Dataset

The dataset comes from National Practitioner Data Bank's (NPDB) Public Used Data File [7]. National Practitioner Data Bank, started in 1990, is a database that contains information related to the professional competence and conduct of physicians, dentists, and other health care practitioners. We accessed the data in 2019, so it included reports received from September 1, 1990 through December 31, 2018.

This analysis studied dental malpractice reports aiming to find important factors and to identify possible practitioners who could have repeat acts of malpractice in the future based on characteristics of their first offence. Due to several new variables added since January 31, 2004, and an average 4-year-gap between an incident and its payment, a 10-year-period (2005 – 2014), based on the report filing year (ORIGYEAR) was used in this study. In addition, only dentists (LICNFELD: 30) and dental residents (LICNFELD: 35) were included in this study. Records with missing data were deleted. Furthermore, only the first malpractice report of each practitioner was kept.

2.2. Variable Selection

Columns that are irrelevant to malpractice, such as Adverse Action Classification, Basis for Action, etc., were deleted from the dataset. After adjusting for inflation based on CPI-U for the U.S. City Average for All Items from the Bureau of Labor Statistics, we replaced Payment (PAYMENT) and Total Payment (TOTALPMT) by PAYMENT_ADJ and TOTALPMT_ADJ correspondingly. According to guidelines, we created a new variable 'STATE', which equals Work State if a work state value was reported and Home State if no work state was reported [7]. Since, for most practitioners, their License States (LICNSTAT) and their work States (STATE) are identical, we created a new binary variable State Difference (STATEDIFF), which is 1 if license state is different from state, and 0 other wise. We counted the occurrences of each LICNSTAT, weighed against population of that state and rescaled those values to between 0.0001 and 0.9999. This information was stored in variable

'STATWEGHT_ADJ'. Similar to States, we created a new variable Payment Difference (PAYDIFF) to track the difference between payment and total payment. In addition, we created a variable Years of Experience (YREXP), for years between Graduation (GRAD) and Year of Act or Omission 1 (MALYEAR1). We also added variable Duration for years between MALYEAR1 and Year of Original Report processed (ORIGYEAR). A binary prediction variable 'RISK' was created based on the number of malpractice reports a practitioner had. The risk is 0 for a practitioner with only one malpractice report on record and the risk is 1 if a practitioner has more than one malpractice reports.

2.3. Statistical Methods

In variable analysis, we calculated summary statistics, such as mean, median and quantiles of continuous variables using standard methods. For those variables with skewed distribution, the Mann-Whitney test was performed to compare the difference between two risk levels. For variables with normal distribution, a two sample t-test was used. In addition, Pearson's Chi-Square test was used to identify differences between some categorical variables against the two risk levels.

In prediction, we randomly split the data into train (75%) and test (25%) datasets. We trained and tuned models using 5-fold cross validation on the training set. Then, we fitted the model on the test data for performance measures. We used Logistic Regression, Random Forest (RF) and XGBoost and tuned the hypermeters of models accordingly through grid search and cross validation. Due to imbalanced data, we experimented with up sampling, down sampling and Synthetic Minority Over-sampling Technique (SMOTE).

All analyses were performed in RStudio (R version 3.5.3). All tests were two-sided, with $p < 0.05$ being considered statistically significant.

3. Results

3.1. Variable Analysis

Between 2005 and 2014, there were over 9,000 new dental practitioners who had their first malpractice cases reported. Of those practitioners, 7,494 (83.22%) of them were identified as risk free for multiple malpractices in the near future (Risk 0), and 1,511 (16.78%) of them were identified as at risk of multiple malpractices in the near future (Risk 1). Improper performance was the primary type of complaint.

Despite some fluctuations, the total number of malpractice reports generally decreased from 2005 to 2014. There were over 1,000 cases in 2005; this decreased to 857 cases in 2014 (Figure 1A). Although the number of cases decreased throughout the ten years being studied, mean payment amount after adjusting for inflation (PAYMENT_ADJ) increased for both risk free and at risk practitioners. This suggests that the cost of malpractice increased every year (Figure 1B). In addition, mean payment amount (Risk 0 = $70,360; Risk 1 = $105,660) and median payment amount (Risk 0 = $22,436; Risk 1 = $29,570) for at risk cases were higher than those for risk free cases (p value < 0.001). Average years of experience (YREXP) for the risk free group (mean = 24.17) was also significantly higher ($P < 0.001$) than that of the at risk group (mean = 22.96). Most practitioners have their first claim in their 40s.

Furthermore, there was a significant difference in the duration (DURATION) of each dispute between the two groups (p value < 0.001). The average duration of the risk free group was 3.27 years, in contrast to 3.93 years of the at risk group.

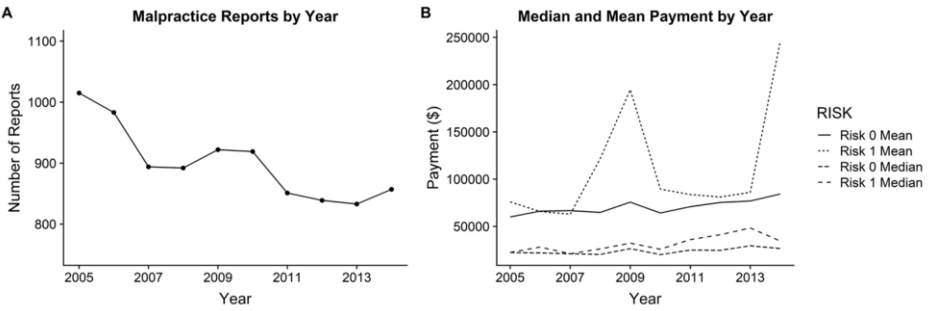

Figure 1. Malpractice reports trend.

There was no significant difference for different levels of outcome (OUTCOME) between the two groups (p value = 0.1159). While the amount of minor injuries decreased gradually from 2005 to 2014, the amount of major injuries increased. The amount of emotional injury and death remained constant. Moreover, there was no significant difference for gender of patients (PTGENDER) between groups (p value = 0.4). However, around two-third of the reports involved female patients and the proportion stayed consistent over the years.

3.2. Predictive Models

We used Logistic Regression as a baseline model. Prior to resampling the data, the logistic model categorized a significant amount of data to the risk free group, due to the imbalanced dataset. Thus, we experimented with up sampling, down sampling and SMOTE and compared the results. The result using Logistic Regression yielded a higher F1 score after resampling, increasing from 2.1% to 30.6%. However, the accuracy of the model decreased from over 80% to 60%. Thus, we would like to find a model that produced a high F1 score without the cost of accuracy. Both Logistic Regression and XGBoost produced relatively high F1 score. Up and down sampling methods also yielded better results than the SMOTE method (Table 1).

Considering XGBoost has a faster model training time and is a robust model which is less subject to overfitting, we decided to use XGBoost as our primary model. We fine-tuned hyper parameters of XGBoost using grid search. We also combined both up sampling and down sampling techniques using the ROSE package in an attempt to find balance between these two methods. The final model accuracy was 72.8% and F1 score was 30.6%.

Table 1. Predictive models results

Logistic Baseline	No Sampling (83.7%, 2.13%)		
	Down (Accu, F1)	Up (Accu, F1)	SMOTE (Accu, F1)
Logistic	(59.1%, 30.7%)	(60.6%, 30.9%)	(74.3%, 24.5%)
Random Forest	(57.7%, 30.6%)	(81.7%, 13.8%)	(81.4%, 11.8%)
XGBoost	(58.5%, 31.0%)	(61.6%, 30.0%)	(82.8%, 10.6%)
XGBoost (Best)	Up and Down Sampling (72.8%, 30.6%)		

4. Discussion

The final predictive model supports the previous discovery that payment amount, dentists' years of experience and length of disputes are major variables in determining whether a dentist is at risk of multiple malpractice cases in the near future. In addition, allegation type, dentists' working states and patient ages are also important factors for prediction.

A highly imbalanced dataset and categorical variables with multiple levels are the two main challenges for constructing predictive models. We employed different sampling techniques and used robust models that are suitable for categorical data to solve these issues. In future studies, we plan to optimize the probability threshold to overcome the imbalanced class problem. Furthermore, summary information of the dataset will be passed into the XGBoost model as additional features to improve model accuracy and F1 score.

5. Conclusion

In the past 15 years, over 1.7 Billion dollars of dental malpractice payments were awarded to patients. While the number of malpractices claims decreased through years, the payment amount per claim has increased. Practitioners' age was also an important variable. Dentists with more years of experience and older age were subjected to higher risk of having multiple malpractice cases in the near future. Additional standardized training should be provided to overcome improper performance issue.

XGBoost was the best machine learning model to predict the risk of dentists having several malpractice reports. The best performing model had an accuracy of 72.8% with 30.6% F1 score.

The NPDB database is a valuable dataset to study dental malpractice claims. Further analysis of information extracted from this dataset is warranted.

References

[1] National Health Expenditure Data: Historical, U.S. Centers for Medicare & Medicaid Services, CMS.gov.

[2] Sekhar, Msonal, and N. Vyas, Defensive Medicine: A Bane to Healthcare, *Annals of Medical and Health Sciences Research* **3(2)** (2013), 295.

[3] Motta S., Testa D., Cesari U., Quaremba G., & Motta G., Medical Liability, Defensive Medicine and Professional Insurance in Otolaryngology, *BMC Research Notes* **8** (2015), 343.

[4] Guardado J, Policy Research Perspectives: Medical Liability Claim Frequency among U.S. Physicians, American Medical Association.

[5] Saber Tehrani, A. S., Lee, H., Mathews, S. C., Shore, A., Makary, M. A., Pronovost, P. J., & Newman-Toker, D. E.,25 Year Summary of US Malpractice Claims for Diagnostic Errors 1986–2010: An Analysis from the National Practitioner Data Bank, BMJ Quality & Safety 22(8) (2013), 672– 680.

[6] Rubin J.B., Bishop T.F., Characteristics of Paid Malpractice Claims Settled In and Out of Court in the USA: a Retrospective Analysis, BMJ Open 3(6) (2013), e002985.

[7] National Practitioner Data Bank Public Use Data File, 2018, U.S. Department of Health and Human Services, Health Resources and Services Administration, Bureau of Health Workforce, Division of Practitioner Data Bank.