Predicting Medical Malpractice of Healthcare Practitioners

Hezhi Wang, Center for Data Science, hw1567@nyu.edu
Han Zhao, Center for Data Science, hz1411@nyu.edu

Abstract

The aim of this paper is to build a predicative model to measure the risk of malpractice at physician level, and to determine the most important factors that affect the probability of malpractice. The datasets we used are Medicare data from 2006 to 2012, Florida malpractice dataset of malpractice claims filed, and AMA master file of demographic features of national practitioners. We focus on information from physicians' medical practice including prescription and procedures, demographic features of their patients, as well as their own characteristics. Using the data, we then build generalized linear regression and regression tree models. Finally, we test our model on Texas malpractice record.

1 Introduction

Medical malpractice occurs when a hospital, doctor or other health care professional, through a negligent act or omission, causes an injury to a patient. In the United States, patients who believe they have been the victims of medical malpractice can recover damages by bringing lawsuits against the providers. In this project, we aim to explore whether malpractice can be predicted by certain factors, including patient outcome, payment information as well as demographic information of physicians.

2 Related Work

Many studies have been done to explore the characteristics, causes and effects of malpractice. Some focus on exploratory analysis of malpractice payments. Tehrani et al (2013) provides a summary for US malpractice claims for diagnostic errors, Mello et al (2010) measures the national cost of the medical liability system. A large part of literature is dedicated to the impact of tort reform on medical malpractice from the perspective of practitioners, patients as well as insurers. (Janet Currie et al, 2006; Ronan Avraham, 2007; David A. Matsa, 2007; Patricia Born et al, 2009; Seabury et al, 2014;; Andrew I. Friedson, 2015).

Another significant part aim to reveal the effect of malpractice. Malpractice can have remarkable market responses (David Dranove et al, 2012) and welfare effects (Lakdawalla & Seabury, 2009). It changes the behavior of physicians (Ity Shurtz, 2012). David Dranove and Anne Gron (2005) states that high malpractice risk may drive doctors away from high-risk procedures. Some specialties may be more heavily affected, like obstetrics and gynecology (Jessica Wolpaw Reyes, 2010; Gilbert W. Gimm, 2010). It can have further impact on the sorting of medical students across medical occupations (Pascal Courty & Gerald R. Marschke, 2008). Fear of malpractice claims can lead to defensive medicine (Daniel Kessler and Mark McClellan, 1996) which then induces rising cost of health care.

However, there isn't much empirical research on the influencing factors of malpractice risk. Bernard S. Black et al (2017) explore the association between malpractice risk and patient safety indicators (PSI rates) aggregated at hospital level. Michael Greenberg et al (2011) examine

malpractice risk and electronic technology used to enhance clinical decision making. Malpractice can also be associated with specialties of practitioners. Anupam B. Jena et al (2011) conducts exploratory analysis of the proportion of physicians who had malpractice claims across 25 specialties and divided the specialties into high and low risk groups. Sandeep Mangalmurti et al (2014) explore malpractice liability among cardiologists and conclude it to be of significantly high risk. Bovbjerg & Petronis (1994) discuss whether past history of malpractice claim will affect future claims.

In this article, we aim to build a predicative model to measure the risk of malpractice at physician level. We also intend to determine the most important factors that affect the probability of malpractice of physicians. Features to explore include information from their medical practice, prescription and procedures, demographic features of their patients, as well as their own characteristics.

3 The Data

Our data consist of 4 major parts. The first part is malpractice data in Florida which contains 12904 malpractice claims which were settled from 2003 to 2015. Each record provides details about the malpractice claim, including but not limited to, id of the practitioner, which is the Florida state license number, date and location of injury occurred, lawsuits filed and settled, and descriptions of court decisions and final disposition. It comes together with Health Care Practitioner Data Portal in FL which provides more demographic information about each physician, like first and last name, address line, zip code and city. These two datasets are linked by FL license number.

The second part is the Medicare data from 2006 to 2012. We do not have direct access to this dataset, and we worked with Professor Daniel L. Chen who can access the datasets in the format of a government SAS server. The codebooks of Medicare date are listed online on Chronic Conditions Data Warehouse. The Medicare dataset contains rich information about every beneficiary and their associated claims. It includes date, payment amount, diagnosis and many other details of Medicare claims from Medpar (inpatient claims), Part B Carrier claims and line file (outpatient and clinic claims) and Part D Event (prescription and pharmacy claims). It also lists the demographic features of every beneficiary enrolled in Part AB including gender, date of birth, race in Master Beneficiary Summary File. For our analysis, we only considered these 4 major parts of Medicare data mentioned above.

And the third part is the AMA master file that provide demographic information of each practitioner across the country. Features of interest include gender, year of birth, year of graduation, license state, and specialty. It can be linked with Medicare by NPI, which is a national unique identifier for practitioners.

The last part is the Texas malpractice dataset. It contains all the closed claims, some important timestamp, location and other descriptions, however no information about the physician.

4 Methods

4.1 Feature Selection

As the size of Medicare data is huge, we need to conduct careful feature selection for pre-processing phase to improve efficiency and save space. Firstly, the demographic features for practitioners can be obtained from AMA master file, and we would like to include gender, year of birth, year of graduation and specialty. The years will be used to calculate physician experience

and age. Secondly, the demographic features of patients, including date of birth, race, gender, date of death, can be obtained from one part of Medicare dataset. And finally, we would like to include other medical information related to activities of physicians, including amount of payments, amount of prescription, length of inpatient stay, number of ER visit. These will then be aggregated at physician level.

4.2 Data preprocessing

4.2.1 Malpractice

One of the major issues we have with this dataset is that it only contains state license number of each practitioner, however the prevailing identifiers for physicians and prescribers in both Medicare and AMA date are NPI. So the first step should be matching license number with NPI. Since license number varies across different states and therefore comes in a messy format, it is intractable to search and map it with NPI directly. To tackle this, we first merge Malpractice with FL practitioner summary file to get the first and last names. And then we utilized the API for NPI lookup, which is provided by CMS National Plan and Provider Enumeration System (NPPES) with first and last name as main search criteria and limit the state to be FL. When there are multiple matches, we would further validate by other information manually. The matched dataset contains around 11000 records. As in our analysis, the malpractice data is only used for labelling, the finalized malpractice data contains only NPI and year of injury occurred. It will be merged with Medicare dataset by NPI.

For labeling, we are assuming that the past does have an impact on the future. This is partly inspired by Sharan, U., & Neville, J. (2008) 's work on incorporating temporal-relational features for prediction. So when the year of malpractice claim matches the year of service for a given record, the label will be set to 1. And if the years do not match, but within a 4-year time lag, the label will be set to be less than 1 but larger than 0, decreasing exponentially. And after that time range, or if there is simply not a match, the label will be set to 0.

4.2.2 Medicare and AMA

The Medicare data is huge and messy. For just Part D, after concatenating all the data from 2006 to 2012, there are over 100 million records. The five parts of Medicare data that will be considered are linked by Beneficiary ID, which means that each record is associated with a beneficiary. However, it is easy to see that it is not a unique key, so if we are to combine these datasets, merging by ID is inefficient, and will also cause the data size to explode. So we write out a whole detailed plan for the whole process of data cleaning. The basic idea here is to pick a unique cohort and pick features of interest instead of merging directly, so that we can reduce data size and improve efficiency quite significantly.

Within each part of Medicare dataset, there are separate files for each year from 2006 to 2012. Thus we begin by concatenating files from each year for each part, only keeping columns that will be considered as features. Then Part D event (PDE) is merged with AMA master file by NPI, and only records within Florida (or Texas, for testing) will be kept. A list of unique physician ids is then extracted from this file, which forms the basis of physician cohort. This cohort is then used to pick records from Part B base claim file and Part B carrier line file. Based on the three files mentioned above, a list of unique beneficiary id is created, which forms the basis of the patient cohort. This list is then combined with master beneficiary file which contains demographic

features of all the patients. It will then be merged back to the 3 datasets mentioned above so that they will contain information of race, gender and other features of patients as features.

Now we turn to Medpar data. This is different from the 3 datasets because it does not contain any information about the physician. Alternatively, the year, month of service and id of the patient are selected as keys to match with Part B base claim data. We choose not to match this with all of the 3 datasets to avoid repetition.

Finally, as all of the 3 main datasets are now set, they are then aggregated at the physician level, which is by physician id and year. By doing this, we will get features from the physician level, like the average payment of all his/her patients, total number of patients treated by him/her in a given year. After the aggregation is done, each of the 3 datasets will now have physician id and year as the unique keys, making it very easy to merge them together. Now malpractice data is merged with Medicare dataset and labels are created according to process discussed above.

4.2.3 Texas Malpractice

The Texas data is recorded yearly. Because it does not contain any information of the practitioner, to align with the Medicare dataset, we use the year and county of the incident as the key. The data is firstly concatenated by the recorded year, and aggregated to create count of each county in each year. The count is then used to compare with the predicted value.

4.2.4 Data Summary

The finalized FL dataset has 49 columns with 157860 records. Of all the records, 115968 rows, which is about 73% are labeled as zero. Of all the columns, 45 will be used as features, with 2 being classification variables and others being numerical variables. A summary for numerical values of our dataset is attached below. Note that for some columns with missing values, we then fill it with 0.

Page Break

Summary of Practitioner information_1

The MEANS Procedure

Year	N Obs	Variable	N	Mean	Maximum	Minimum	Range	Std Dev
2006	7	phy_exp	7	22.4286	38.0000	7.0000	31.0000	10.8452
		phy_age	7	50.7143	63.0000	43.0000	20.0000	7.7613
2007	12192	phy_exp	12192	22.4669	58.0000	0	58.0000	9.4944
		phy_age	12192	49.5040	107.0	26.0000	81.0000	9.3610
2008	26622	phy_exp	26622	21.5680	62.0000	-1.0000	63.0000	10.5007
		phy_age	26622	48.8312	108.0	24.0000	84.0000	10.5842
2009	28464	phy_exp	28464	21.8953	63.0000	0	63.0000	10.8221
		phy_age	28464	49.2215	109.0	24.0000	85.0000	10.9876
2010	29327	phy_exp	29327	22.3718	64.0000	0	64.0000	11.0691
		phy_age	29327	49.7755	110.0	24.0000	86.0000	11.3416
2011	30305	phy_exp	30305	22.7908	65.0000	0	65.0000	11.3138
		phy_age	30305	50.2592	111.0	23.0000	88.0000	11.6760
2012	30943	phy_exp	30943	23.2635	66.0000	0	66.0000	11.5556
		phy_age	30943	50.7705	112.0	24.0000	88.0000	11.9719

Page Break

Summary of Medical information

The MEANS Procedure

Variable	N	Mean	Maximum	Minimum	Range	Std Dev
Num_of_Patients_CLAIM	157860	650.5	16591.0	1.0000	16590.0	972.3
Avg_PMT_AMT_CLAIM	157860	117.1	16863.9	0	16863.9	116.9
Avg_PRVDR_PMT_AMT_CLAIM	157860	116.5	16863.9	0	16863.9	117.1
Num_of_Male_CLAIM	157860	279.3	8922.0	0	8922.0	445.7
Num_of_Female_CLAIM	157860	371.2	10731.0	0	10731.0	568.9
Avg_patient_birth_year_CLAIM	157860	1937.8	2011.0	1905.0	106.0	8.1196
Race_Unknown_CLAIM	157860	1.2105	626.0	0	626.0	6.7684
Race_White_CLAIM	157860	565.7	16541.0	0	16541.0	894.7
Race_Black_CLAIM	157860	47.5666	7666.0	0	7666.0	121.0
Race_Other_CLAIM	157860	7.6861	794.0	0	794.0	21.6720
Race_Asian_CLAIM	157860	3.0981	1416.0	0	1416.0	14.2956
Race_Hispanic_CLAIM	157860	24.7201	6626.0	0	6626.0	115.4
Race_North_American_Native_CLAIM	157860	0.4784	328.0	0	328.0	4.4244
NUM_OF_DEATH_PATIENT_CLAIM	157860	13.0976	3178.0	0	3178.0	61.9945
avg_num_phyvis_CLAIM	134348	1.8061	28.0000	1.0000	27.0000	1.3829
avg_er_vis_CLAIM	13262	1.0220	5.3939	1.0000	4.3939	0.0907
avg_loscnt	0	.	.	.	.	.
avg_mdcr_pmt_amt	0	.	.	.	.	.
Num_of_Patients_PDE	157860	549.6	36861.0	1.0000	36860.0	1106.0
Num_of_Male_PDE	157860	205.1	15879.0	0	15879.0	429.3
Num_of_Female_PDE	157860	344.5	20982.0	0	20982.0	714.8
Avg_Quantity_Dispensed_PDE	157860	69.9150	28941.8	0.00500	28941.8	236.0
Avg_Days_Supply_PDE	157860	30.1143	93.0000	1.0000	92.0000	14.9574
Avg_Drug_Cost_PDE	157860	80.5037	10650.9	0.0100	10650.9	160.1
Avg_patient_birth_year_PDE	157860	1940.3	2011.0	1905.0	106.0	9.2345
Race_Unknown_PDE	157860	1.4587	505.0	0	505.0	10.4453
Race_White_PDE	157860	413.7	28756.0	0	28756.0	819.1
Race_Black_PDE	157860	68.2306	10591.0	0	10591.0	243.0
Race_Other_PDE	157860	10.8977	1895.0	0	1895.0	46.8153
Race_Asian_PDE	157860	4.1344	2511.0	0	2511.0	31.6988
Race_Hispanic_PDE	157860	50.6239	11800.0	0	11800.0	282.7
Race_North_American_Native_PDE	157860	0.5419	984.0	0	984.0	7.8434
NUM_OF_DEATH_PATIENT_PDE	157860	1.6762	1087.0	0	1087.0	15.3205
avg_num_phyvis_PDE	120058	1.4619	28.0000	1.0000	27.0000	0.9382
avg_er_vis_PDE	10072	1.0263	3.0000	1.0000	2.0000	0.0872

4.3 Model Building

With real-valued target variable, our model should be supervised learning by regression. Due to limitations of the version of SAS server, here we trained 2 types of models.

4.3.1 Generalized Linear Regression

With so many columns, lasso regression is a natural way to do feature selection and reduce dimensionality. We thus include all the variables and their cross terms and perform stepwise regression. The stepwise regression stops at step 19, and the selected features are shown below.

Data Set	DCH070SL.MERGED_ALL
Dependent Variable	Target
Selection Method	Stepwise
Select Criterion	SBC
Stop at Specified Number of Steps	20
Choose Criterion	C(p)
Effect Hierarchy Enforced	None

Number of Observations Read	157860
Number of Observations Used	157860

Class Level Information

Class	Levels	Values
Sex	2	F M
PrimarySpecialty	188	A ACA ADL ADM ADP AHF AI ALI AM AN APM AR AS ASO ATP BBK CAP CCA CCM CCP CCS CD CFS CHN CHP CHS CLP CN CRS CS D DBP DIA DMP DR DS EFM EM END ES ESM ETX FMP FOP FP FPG FPP FPR FPS FSM GE GO GP GPM GS GYN HEM HEP HMP HNS HO HOS HPE HPF HPI HPM HPO HS HSO ...

Dimensions	
Number of Effects	562
Number of Parameters	6898

The GLMSELECT Procedure
Selected Model

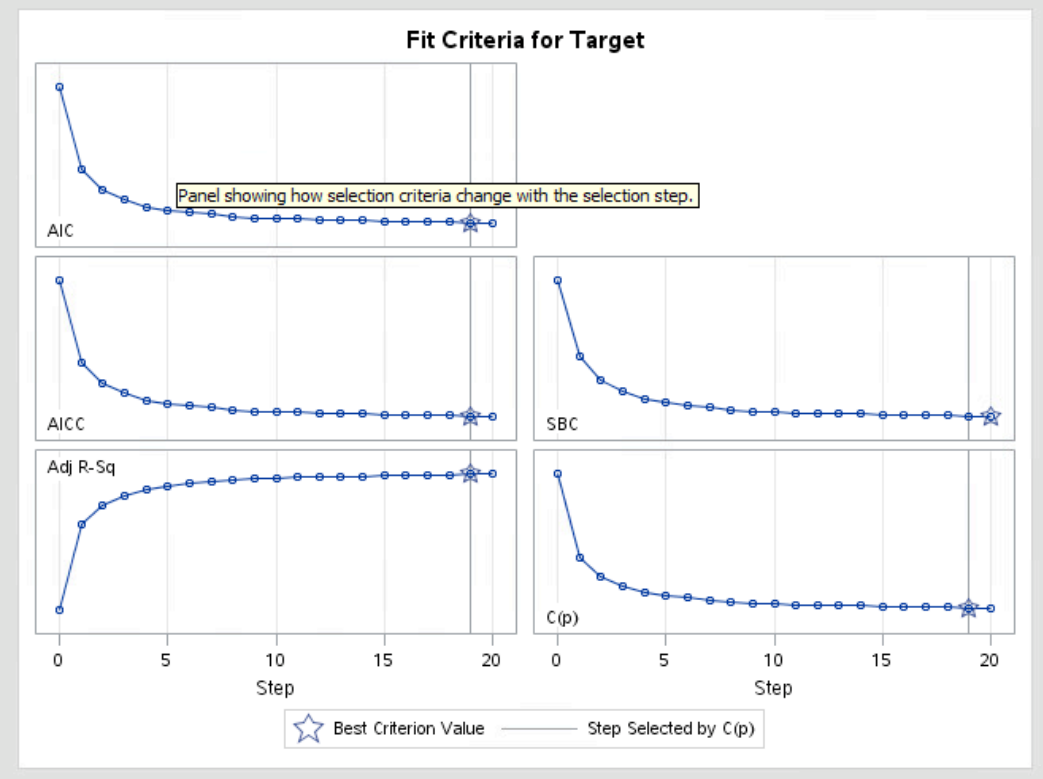
The selected model, based on C(p), is the model at Step 19.

Effects:	Intercept Num_of_Pa*Num_of_Mal Avg_patient_birth_ye Race_White_CLAIM*Sex Avg_patie*Race_White Avg_PMT_A*Race_Black Avg_PRVDR*Race_Black Race_Blac*Num_of_Fem Num_of_Pa*Avg_Days_S Avg_patie*Avg_Days_S Avg_PMT_A*Race_White Avg_PRVDR*Race_White Avg_Quant*Race_Hispa Avg_patie*Race_Hispa phy_exp*PrimarySpeci Num_of_Patie*phy_exp Num_of_Male_*phy_exp phy_age*Sex Num_of_Patie*phy_age phy_exp*phy_age

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value
Model	208	2032.15550	9.76998	78.93
Error	157651	19513	0.12378	
Corrected Total	157859	21546		

Root MSE	0.35182
Dependent Mean	0.19538
R-Square	0.0943
Adj R-Sq	0.0931
AIC	-171743
AICC	-171743
BIC	-329615
C(p)	5223.59582
SBC	-327522



4.3.2 Regression Tree

The baseline model is a regression tree with default settings. The tree is built with cross-validation and the splitting criterion is square error. Data is randomly split into training and testing subset, with 70% and 30% of the original data respectively. After the model is built on training data, we do predictions on the testing dataset. Since SAS does not calculate explicitly the results and evaluation metrics of this model, we are unable to recover the most important features of this model.

To tune hyper-parameters, we then calculate manually the sum of squared error between the real value and the predicted value (equivalent to residual). The parameters include tree depth and leaf size. Leaf size ranges from 2 to 1024, and max depth ranges from 10 to 100. The model with the least square loss on testing dataset is when leaf size is 2, and max depth is 20, with total square loss 8179.46.

4.4 Results

Our aim is to predict the probability of malpractice, as well as to identify the most important factors.

The two models can be both used to predict malpractice. For the regression tree, the predicted value is guaranteed to be within the bound of [0,1], which is reasonable to be interpreted as a probability measure. However, for generalized linear regression, there is no bound for the predicted value. We may need to do some transformations to make it easy to understand.

And for identify most features, the two models can actually both be used. The linear regression can select features by stepwise regression, adding or dropping features following certain criteria. For the regression tree, the closer a certain feature is to the root, the more important it is, as it reduces more of the total square error. However, we are unable to get that information with this version of SAS server.

It is hard to compare the 2 models given information we have now, though. We cannot compare by their square loss (residual) because they are not built on same amount of data. An alternative way is to use yet another labeled and independent dataset to see their predicting power.

4.5 Test with Texas Dataset

We apply the models which are built on the Florida Medicare dataset and apply to Texas Medicare dataset and get the predict value. To match with TX malpractice records, we aggregate the predicted results at county and year level. And finally we calculate the percentage error of our predicted value with the real count. The results for the 2 models are shown below.

Analysis Variable : error_per					
N	Mean	Maximum	Minimum	Range	Std Dev
317	16.8759	338.0	-0.9976	339.0	35.7379

Regression Model

Analysis Variable : error_per					
N	Mean	Maximum	Minimum	Range	Std Dev
317	17.6149	348.6	-0.9937	349.6	37.4658

Tree Model

We can see that the performance of the 2 models are approximately the same. However, predicted values of both are significantly larger than the real value.

5 Challenges and Future Work

The main challenges we faced during the whole process are the data cleaning phase. The Medicare dataset is huge and not directly accessible for us. Also, the version of SAS Enterprise Guide leaves us few choices of available models, and also makes model tuning very difficult, especially for the regression tree model.

Given more time, we would try to improve the performance of our model. We would like to look deeper into the malpractice dataset. For simplicity we used whole malpractice dataset, it is doubtful whether the practitioner actually did something wrong even with a claim filed. Also, we can include more features from Medicare dataset. Besides, we notice that in our current model,

there are 188 specialties for practitioners. It could be better if we categorize them into smaller buckets to reduce dimensionality, but it requires professional knowledge so we are unable to include that now.

Acknowledgements

We are grateful to our advisor and also collaborator, Professor Daniel L. Chen for his guidance and patience and all the timely meetings during the project. We are also very thankful to Ling Li, biostatistician at the Dana Farber Cancer Institute of Harvard Medical School for offering generous help in understanding and pre-processing the Medicare data. We would also like to thank our professor David Rosenberg for advice and support all along.

References

[1] Avraham, R. (2007). An empirical study of the impact of tort reforms on medical malpractice settlement payments. *The Journal of Legal Studies*, *36*(S2), S183-S229.

[2] Black, B. S., Wagner, A. R., & Zabinski, Z. (2017). The Association between Patient Safety Indicators and Medical Malpractice Risk: Evidence from Florida and Texas. *American Journal of Health Economics*.

[3] Born, P., Viscusi, W. K., & Baker, T. (2009). The effects of tort reform on medical malpractice insurers' ultimate losses. *Journal of Risk and Insurance*, *76*(1), 197-219.

[4] Bovbjerg, R. R., & Petronis, K. R. (1994). The relationship between physicians' malpractice claims history and later claims: does the past predict the future?. *Jama*, *272*(18), 1421-1426.

[5] Courty, P., & Marschke, G. R. (2008). *On the sorting of physicians across medical occupations* (No. w14502). National Bureau of Economic Research.

[6] Currie, J., & MacLeod, W. B. (2006). *First do no harm?: Tort reform and birth outcomes* (No. w12478). National Bureau of Economic Research.

[7] Dranove, D., & Gron, A. (2005). Effects of the malpractice crisis on access to and incidence of high-risk procedures: evidence from Florida. *Health Affairs*, *24*(3), 802-810.

[8] Dranove, D., Ramanarayanan, S., & Watanabe, Y. (2012). Delivering bad news: Market responses to negligence. *The Journal of Law and Economics*, *55*(1), 1-25.

[9] Friedson, A. I. (2015). Medical Malpractice Damage Caps and Provider Reimbursement. *Health economics*.

[10] Gimm, G. W. (2010). The impact of malpractice liability claims on obstetrical practice patterns. *Health services research*, *45*(1), 195-211.

[11] Greenberg, M., & Ridgely, M. S. (2011). Clinical decision support and malpractice risk. *JAMA*, *306*(1), 90-91.

[12] Jena, A. B., Seabury, S., Lakdawalla, D., & Chandra, A. (2011). Malpractice risk according to physician specialty. *New England Journal of Medicine*, *365*(7), 629-636.

[13] Kessler, D., & McClellan, M. (1996). Do doctors practice defensive medicine?. *The Quarterly Journal of Economics*, *111*(2), 353-390.

[14] Lakdawalla, D. N., & Seabury, S. A. (2012). The welfare effects of medical malpractice liability. *International review of law and economics*, *32*(4), 356-369.

[15] Mangalmurti, S., Seabury, S. A., Chandra, A., Lakdawalla, D., Oetgen, W. J., & Jena, A. B. (2014). Medical professional liability risk among US cardiologists. *American heart journal*, *167*(5), 690-696.

[16] Matsa, D. A. (2007). Does malpractice liability keep the doctor away? Evidence from tort reform damage caps. *The Journal of Legal Studies*, *36*(S2), S143-S182.

[17] Mello, M. M., Chandra, A., Gawande, A. A., & Studdert, D. M. (2010). National costs of the medical liability system. *Health affairs*, *29*(9), 1569-1577.

[18] Reyes, J. W. (2010). *The effect of malpractice liability on the specialty of obstetrics and gynecology* (No. w15841). National Bureau of Economic Research.

[19] Seabury, S. A., Helland, E., & Jena, A. B. (2014). Medical malpractice reform: Noneconomic damages caps reduced payments 15 percent, with varied effects by specialty. *Health Affairs*, 10-1377.

[20] Sharan, U., & Neville, J. (2008, December). Temporal-relational classifiers for prediction in evolving domains. In *Data Mining, 2008. ICDM'08. Eighth IEEE International Conference on* (pp. 540-549). IEEE.

[21] Shurtz, I. (2013). The impact of medical errors on physician behavior: Evidence from malpractice litigation. *Journal of health economics*, *32*(2), 331-340.

[22] Tehrani, A. S. S., Lee, H., Mathews, S. C., Shore, A., Makary, M. A., Pronovost, P. J., & Newman-Toker, D. E. (2013). 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), 672-680.