Predicting Medical Malpractice Injury Settlements using the National Practitioner Data Bank

Hailie Dono, Marlee Sherrod, & Michelle Uroza

Messy Data and Machine Learning (Spring 2023)

New York University

May 11, 2023

Introduction

Providing high quality and professional care should be the goal for the healthcare community. However, in the United States alone, an average of 85,000 medical malpractice cases are filed each year (“US Medical Malpractice Case Statistics”, 2022). Malpractice is defined as professional activity that is deemed improper, illegal and/or negligent in any manner. A common consequence of these malpractice acts is a monetary compensation paid to the victim that is determined by a judge, or the parties involved in the incident. In the United States alone, the average malpractice payment grew approximately 52 percent between the years 1991 and 2003 (Chandra et al. 2005). While these rising costs are consistent with increases in the cost of healthcare, the specific attributes of an individual malpractice claim are the primary determinants of the final payment owed.

Thus, the substantive question of our research then comes down to: What are the best predictors for the probability of a malpractice case resulting in a higher payout? Intuitively, the next immediate question may be: who are the stakeholders of this research? The primary stakeholders we may imagine are medical insurers. From 2011 to 2015, insurer payouts were up across the dollar amount spectrum; and, payments of $500,000 to $1 million rose almost fivefold (“Legal Malpractice” , 2016). As the primary payor in a majority of malpractice claims, we may identify insurers as a primary stakeholder group to this substantive question. The intended model could assuage insurance related concerns such as whether or not a new report will result in a high payout (over $500,000.00). Another potential secondary stakeholder in this research is the practitioners themselves.

According to medical malpractice insurance broker Max Schloemann, practitioners who have a malpractice indemnity payment in the past five years are likely to experience a significant increase in medical malpractice insurance rates than those with no claims (Schloemann, 2022). In addition to increasing insurance rates, some insurers only cover part of the payout– that means the rest would be paid out of the practitioner’s pocket. In this situation we perceive a practitioner would be interested in the probability that their case results in a particularly high settlement.

To conclude, through statistical analysis and machine learning, our research goals are:

1) To predict the probability of receiving a compensation payment of $500,000 based on the following malpractice payment report characteristics: severity, patient age, patient gender,

age group of practitioner, practitioner's field of license, past malpractice and adverse action history of the practitioner; etc.

2) Through the use of randomForest feature selection, determine the best predictors of compensation payment outcome;
3) Conduct model comparisons to determine the most optimal model for predicting malpractice cases that result in a high monetary settlement.

Data

The data comes from the National Practitioner Data Bank. It is a database, specifically a web-based repository, operated by the U.S. Department of Health and Human Services that contains medical malpractice payment and adverse action reports on healthcare professionals. The NPDB was established by Congress in 1986 with the goal of preventing medical professionals from moving to different states to practice without disclosing their past performance ("The NPDB - About Us"). The process of how the NPDB works is detailed in Figure 1.

Specifically, this research uses the National Practitioner Data Bank's Public Use File data set and examines the data set within a ten consecutive year time frame (the current file contains 1,676,479 cases from September 1, 1990 to December 31, 2022) (U.S. Department of Health, 2023). The data contains information about healthcare professionals such as nurses, doctors, dentists, psychologists, as well as case information about the adverse action or malpractice payout. According to the Department of Health and Human Services, by law this information is required to be reported to NPDB by state licensing boards, medical malpractice payers and others.

The Public Use Data File contains two types of data: reports of the malpractice payment and adverse actions. The National Practitioner's Data Bank defines adverse actions as the outcome or consequence of the claim other than medical malpractice payments ("NPDB Research Statistics") . Some examples of adverse action include suspension of licensure and practice closure. For the purpose of this research project, we will solely look at malpractice reports.

Selected variables from the Public Use Data File include payment, adverse licensure, clinical privileges, professional society membership, and Drug Enforcement Administration

(DEA) reports received by the NPDB concerning physicians, dentists, and other licensed health care practitioners. It also includes reports of Medicare and Medicaid exclusion actions taken by the Department of HHS. Further information on all the variables included in NPDB can be found in Appendix Table 1.

Literature Review

Through the research of medical malpractice within the United States, we concentrated on several emerging trends. These trends in malpractice literature vary in cases relating to: the payments themselves, the severity of malpractice claims, physician specialties/field, and reports relating to diagnosing errors.

A study conducted over a twenty-year span revealed that diagnostic errors were the leading type of malpractice and accounted for the highest proportion of all total payments (Tehrani et al., 2013). This finding is consistent with other research that also found diagnosis errors to be one of the common malpractice claims, as well as the most dangerous yet preventable (Sandstrom, 2007). This knowledge furthered our research in the preliminary stages as we sought to determine the most important features in predicting the probability of high total payment outcomes. Ultimately, the results of this paper led us to consider malpractice allegation types as a predictor in our model. Another analysis spanning twenty-years researched the characteristics of malpractice reports, which concluded that the number of malpractice claims paid on physicians' behalf in the US decreased significantly between 1992 and 2014 (Schaffer et al., 2017). Using this information, we were able to determine another key predictor of malpractice payouts– the specialty, or field of the practitioner reported in the claim. Examples of fields include occupational therapy and examples of specialized practitioners like pharmacists, dentists, and nurse practitioners.

Another trend identified through the research concerns the severity of these malpractice claims; a ten-year long study, primarily composed of significant cases of "Significant Permanent Injury", "Major Permanent Injury" or "Death", found that for the majority of outcomes and various severity levels, claims have decreased overall, including claims by the severity of malpractice injury (Goode et al., 2022). Severity of malpractice injury are classified from 1- emotional injury….6- significant permanent injury, to 9-death. 10 cannot be determined from available records. For a full list of the severity of malpractice injury, see Table 2. We can expect the greater the severity of the malpractice instance, the higher the total payment– therefore, understanding the most recent trends in these severity-levels was influential in how we developed our model.

Finally, through the research we discovered information concerning our outcome variable – total malpractice payments. Researchers Chandra, Nundy, and Seabury (2005) concluded that the average payment grew by 52% percent in a twelve year span (p. 240). As stated prior, while these increases are mostly consistent with increases in the cost of health care, a combination of extreme awards, or certain specializations and other unforeseen influences has resulted in an incomplete understanding of the growth of physician malpractice payments. This conclusion only further motivates our research on selecting the best features for predicting high medical malpractice payments.

Although malpractice data are not the most ideal source for determining the frequency, causes, or remedies for the majority of cases– especially diagnostic errors– they do provide a valuable, measurable indicator of the medicolegal expenses associated with injuries caused by negligent medical practice (Tehrani et al., 2013). In addition to overarching themes and general trends, our literature review also revealed some limitations previous researchers encountered when dealing with NPDB data. This is often due to the fact that the correlation between paid malpractice claims and medical errors is known to be imperfect and therefore many studies are unable to assess specific rates of different types of allegations, the number of events, the tangible and intangible burden of the harm done and the preventability of such harms (Tehrani et al., 2013). Another limitation to the NPDB is that some of the data regarding adverse actions against practitioners have been known to be incomplete (Romano, 2005). Ultimately, our literature review reached a general consensus that the foremost limitation is that the NPDB data recorded on malpractice payments do not always provide a complete and accurate depiction of actual healthcare ineptitude or medical misconduct.

This paper will serve to further current literature on malpractice reports as our goal is to provide greater insight into the factors that can best predict whether or not a malpractice incident will result in a high payout.

Experimental Setup

<u>Data acquisition protocol</u>

As far as the data acquisition protocol is concerned, the National Practitioner Data Bank is a web-based repository. This file is updated four times a year and the most current file available for download contains disclosable reports received from September 1990 through December 2022.

The Data Use Agreement does not permit using the data set in any way that would identify an individual, or entity. And upon request, users may be required to return, delete or permanently dispose of all copies of the data in their possession. In order to download the Public Use Data File, the contact information of the researcher is required in addition to agreement of the Data Use Agreement. Once submitted, the Public Use Data File is sent in the file format of the researcher's choice to the provided email address. For our research purposes we requested a comma-separated values file.

Data annotation

Data annotation is the practice of labeling data in machine learning to demonstrate the outcome you want your machine learning model to anticipate. You are marking a data set with the properties you want your machine learning system to learn to identify, whether by labeling, tagging, transcribing, or processing it. Once your model is put into use, you want it to automatically recognize those features and decide what to do next. Annotated data reveals features that your algorithms will use to recognize the same features in unannotated data. In supervised learning and hybrid, or semi-supervised, machine learning models that incorporate supervised learning, data annotation is used.

In terms of data annotating within our project, we wanted our model to recognize payments that were above a certain threshold. Thus, we engineered an indicator variable that flagged reports with total malpractice payments over $500,000, or half a million U.S. dollars. This threshold was selected as in the time since 2001 medical malpractice claims that resulted in smaller payouts (less than $500,000) have dropped nearly 55%. Over the same time span, there have been substantially fewer major claims ($500,000 or more) against healthcare providers("US Medical Malpractice Case Statistics", 2022. Despite having substantially fewer larger claims, these cases are likely to be of more interest to stakeholders to predict as they have greater monetary consequences.

Data pre-processing/Feature Engineering

Below are the pre-processing and feature engineering steps we took in order to clean the data for modeling. A PDF version of the full codebook, the “NPBD Public Use Data File”, is provided using this link. There is also a full list of all the variables included in the NPDB listed below in Table 1.

First, we decided on what records, malpractice payment records or adverse actions records, we wanted to analyze and in what time frame. While the data file is only about these two types of cases, NPDB overhauled how they recorded adverse actions in 1999 and then how they recorded malpractice payments in 2004. In other words, there are 2 different types of record-reporting styles for malpractice and adverse actions each, thus creating four types of reports in the data file, coded under the “RECTYPE" variable. Adverse actions that were reported to NPDB before November 22, 1999 are recorded as “A” while actions reported after that date are “C”. On the other hand, malpractice cases that were reported to the NPDB before January 31, 2004 are recorded as “M" while payments reported after that data are “P”. Every row/observation has one of these 4 letters as its RECTYPE value.

We decided to analyze malpractice payment records only. Therefore, because of the time frame, all the records in our data set have “P” for RECTYPE. We decided to look at malpractice payments in order to analyze a concrete, objective outcome in what can be a traumatizing, emotion-ladened life event. Also, to keep up with the latest trends in malpractice case reporting and to ensure no overlap between the four RECTYPEs, we decided to only look at data from years 2010 through 2022.

Next, we consolidated all the state-related variables (WORKSTAT, HOMESTAT, LICNSTAT) into one state variable (STATE). While the NPDB file says that reporters must report either the accused practioner’s work state (WORKSTAT) or home state (HOMESTAT), not every report had this information. The LICNSTAT variable, the licensed state of accused practitioner, had the fewest missing values. Practitioners can be licensed in up to 10 states in NPDB, but the Public Use File only reports one. For that reason, the work state value was given priority for our STATE variable and if it was not present, then the HOMESTAT value was used. The practitioner’s licensed state was the final option. Every row in our data set has a STATE value.

Afterward, we removed columns that primarily pertained to adverse action records since many of them have blank/NAs in rows that pertained to malpractice records, and vice versa. We then factored many variables that had numbers to represent categories. For example, the malpractice allegation group (ALGNNATR) lists health sectors with no inherent hierarchical order, like 1 = "Diagnosis Related," 2 = "Anesthesia Related," 3 = "Surgery Related," and so on. We then renamed the levels in these newly-factored variables.

At this point, we started to feature-engineer variables that would provide a time-element to our analysis. We created an additional year variable (ADD_YEAR) that calculates how many additional years a malpractice case took place: looking at the MALYEAR1 and MALYEAR2 variables, most cases only last for a year. This could be because the alleged incident occurred once and MALYEAR1 reports the year it happened. If there is no MALYEAR2, or it is the same year as MALYEAR1, there are no additional years and the value for ADD_YEAR is 0. We also created a similar variable called DURATION where based on the above logic, the base year for all cases is 1. If there temporal differences in MALYEAR1 and MALYEAR2, they are accounted for in both these new variables.

Two other time variables we created were PREVIOUS_CASES and AA_INDICATOR. PREVIOUS_CASES aggregates how many malpractice payment reports were filed against the alleged practitioner before 2010. The AA_INDICATOR aggregates if the practitioner has any adverse actions reports filed against them before 2010. It is a binary, Yes = 1/No = 0 (1/0) indicator.

Another feature we added was if a case had more than one malpractice allegation. Malpractice payment records report the primary allegation, or what the practitioner is accused of doing wrong, in the ALEGATN1 column. Some cases also report a second allegation in the ALEGATN2 column, but the majority do not. We created another binary variable, ADDT_ALLEGNT, to indicate if another allegation was present on the report besides ALEGATN1: It is a True/False variable on whether the ALEGATN2 field is blank or not. We more so want to see if the presence of another allegation impacts our model.

Lastly, we created a feature called LCNFIELD_GRP where we grouped together the 93 different-yet-related practitioner roles from the full data file into 32 groups. Twenty group values were left after cutting down the data to 2010-2020 reporting years and focusing on only "P" malpractice payment records. The logic behind our grouping was based on industry, education or certificate level, and role in health care. We grouped together individual practitioners, i.e. “Dental Group/Practice” into the same “Dentistry” category as “Dentist” and “Dental Assistant”. We came up with 3 different grouping formats and chose the one we hoped would represent the nuance in these fields while also combining them in a logical fashion. We are aware that how we interpreted and grouped these specialities has an impact on how our model predicts. It is a limitation of our data that is worth exploring in future research. The grouping format we decided on can be found in Table 3 in the Appendix.

Other minor cleaning we did was remove the “$” sign from the TOTALPMT and PAYMENT fields; renamed the data set’s original OUTCOME variable to SEVERITY, which records what happened to the patient; renamed the data set’s original TYPE variable to REPRT_ENTITY, which is the type of entity submitting the report to the NPDB, like an insurance company; and we removed the 6 states with the fewest “P” malpractice records in our time frame. We had to do this because our random forest model would not run with more than 53 variable levels.

Finally, as mentioned above, to determine the outcome we wanted, we made our own OUTCOME variable: a binary variable on whether a TOTALPMT amount was greater or equal to $500,000.00 with 1 = “greater than $500,000.00” and 0 = “less than $500,000.00”.

Methods

In order to answer our substantive question our proposed method is a logistic regression model for binary classification. We will model

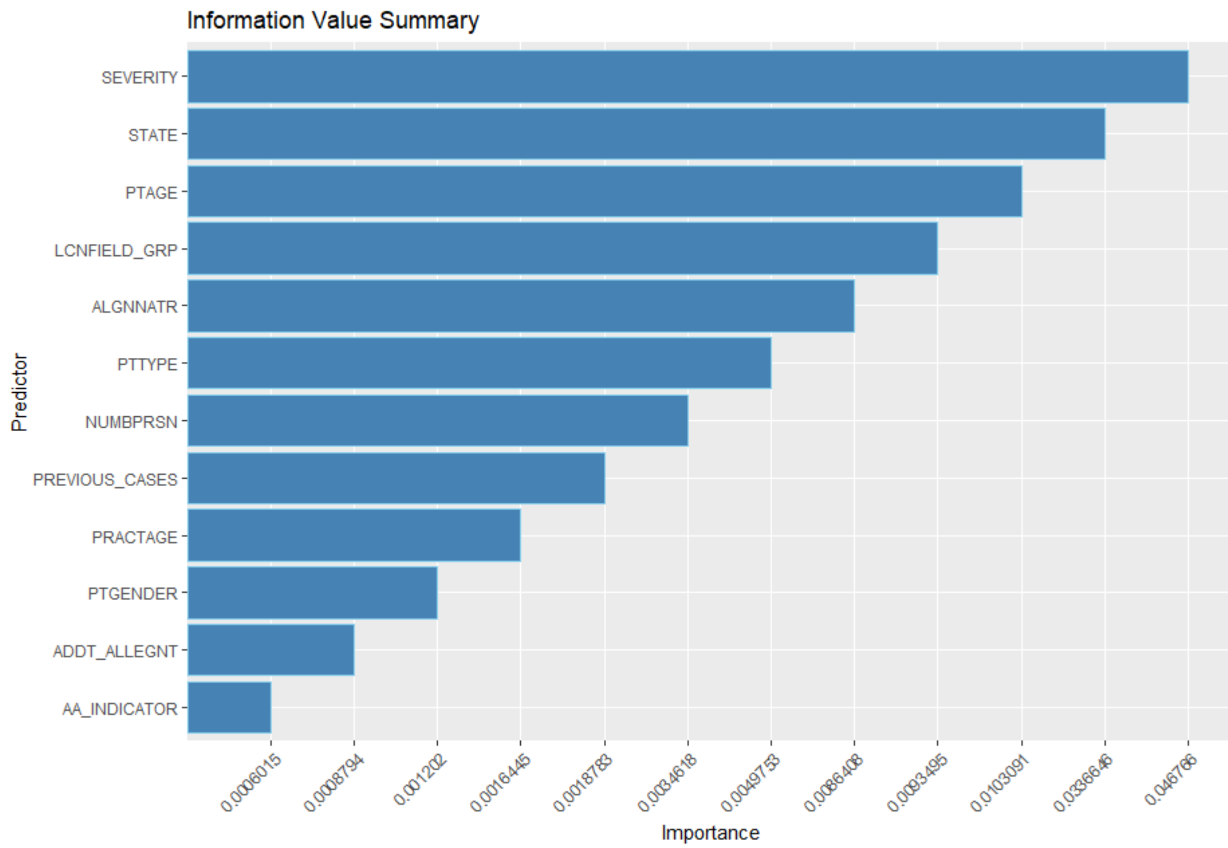
$$Pr(Y_i = 1 | X = (x_{i1}, \ldots, x_{ip})) = logit^{-1}(\beta_1 x_{i1} + \ldots + \beta_p x_{ip}),$$

where Y is a binary indicator of a report having a high payout, and each predictor, X, are various attributes from the data set, such as practitioner age, severity of the report, or patient gender. In this particular model, a one-unit increase in a specific predictor is associated with an additive increase of β_p in the log-odds that $Y_i = 1$, or a multiplicative increase of e^{β_p} in the odds that $Y_i = 1$.

In addition to our logistic regression modeling, we will also employ random forest algorithms for feature selection. As its name suggests, a random forest model is a series of relatively uncorrelated models, or decision trees. In general, the algorithm selects a subset of the observations, builds a tree, and takes the average result.

To reiterate, our proposed methodologies for completing our research goals included classification algorithms such as logistic regression, and random forest classification. For our random forest analysis we relied on the "ranger" package. It is worthy to note that this package cannot model on missing data entries; thus, we began our modeling by removing all rows with missing values from our cleaned data set to keep all model comparisons consistent. Moreover, we will gauge these model comparisons by their classification accuracy. That being said, prior to modeling we split our data temporally into a training and testing set. The training set consists of years 2010 to 2018, and the testing set from years 2019 to 2022. From here we could begin the modeling process.

Our first model of choice was a logistic regression model, with all the predictors we suspected to be important determinants of whether or not a report resulted in a high total payout, or greater than $500,000 U.S. Dollars, as previously stated. These variables were: severity of malpractice injury (SEVERITY); type of malpractice or alleged malpractice (e.g. diagnosis error, treatment error, surgery error,...) (ALGNNATR); the number of malpractice payments prior to 2010 by the practitioner identified in the report (PREVIOUS_CASES); whether or not the practitioner in the report also had an adverse action report against them (AA_INDICATOR); the field of the practitioner (LCINFIELD_GRP); state where report was made (STATE), patient gender (PTGENDER); patient type (e.g. inpatient versus outpatient) (PTTYPE); whether or not there were multiple allegations in a single report (ADDT_ALLEGNT); the age of the patient (PTAGE); the age of the practitioner (PRACTAGE), and the number of practitioners linked to this incident of malpractice (NUMBPRSN). Then we utilized a random forest model for selecting the most important features of our original model. Running this model with a thousand trees yielded importance values for each predictor variable.



We then used the top six most important predictors: severity, state, patient age, practitioner license-field, allegation type-group, and patient type, to run a second model. Finally we ran a second ranger model, with the same predictors as the best fit logistic regression, and in order to make a final model comparison.

Results

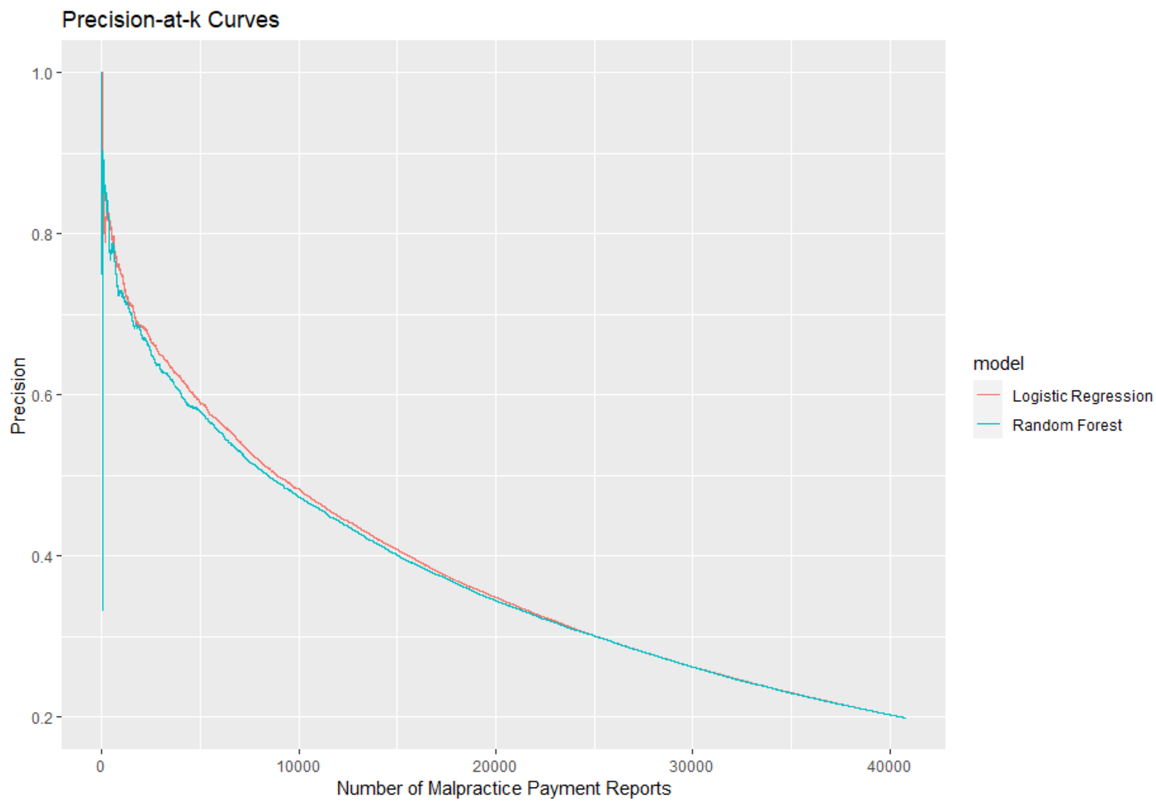
An overall measure of performance across all potential classification criteria is provided by AUC, or "Area Under Curve". AUC can be seen as the likelihood that the model values a randomly chosen positive example higher than a randomly chosen negative example. We will rely on this performance measure to make conclusions about our various models. Where we cannot rely on AUC scores, we will use Akaike's Information Criteria (AIC) and Bayesian Information Criteria (BIC). Any estimated statistical model's quality of fit can be measured using the AIC. The BIC is one way to choose a model from a group of parametric models with various numbers of parameters.

Our first comparison was between our first logistic regression, where we selected all the predictors we were most interested in and our second logistic regression, which used the top six most important predictors as determined by our random forest model. The former regression yielded a score of 0.82 and the latter yielded the same! To combat this lack of information, we found the AIC and BIC of the first logistic regression was 73368.35 and 74490.73, respectively. And the AIC and BIC of the second logistic regression was determined to be 73594.18 and 74591.85, respectively. Typically, A lower AIC or BIC value indicates a better fit. Therefore, we were able to conclude that our first logistic regression was the better model for predicting a high total payment outcome.

Now that we have our optimal model, we were able to compare across a different classification method– random forest classification. Using the same predictors as our best fit logistic model, we ran a random forest using 1,000 trees.

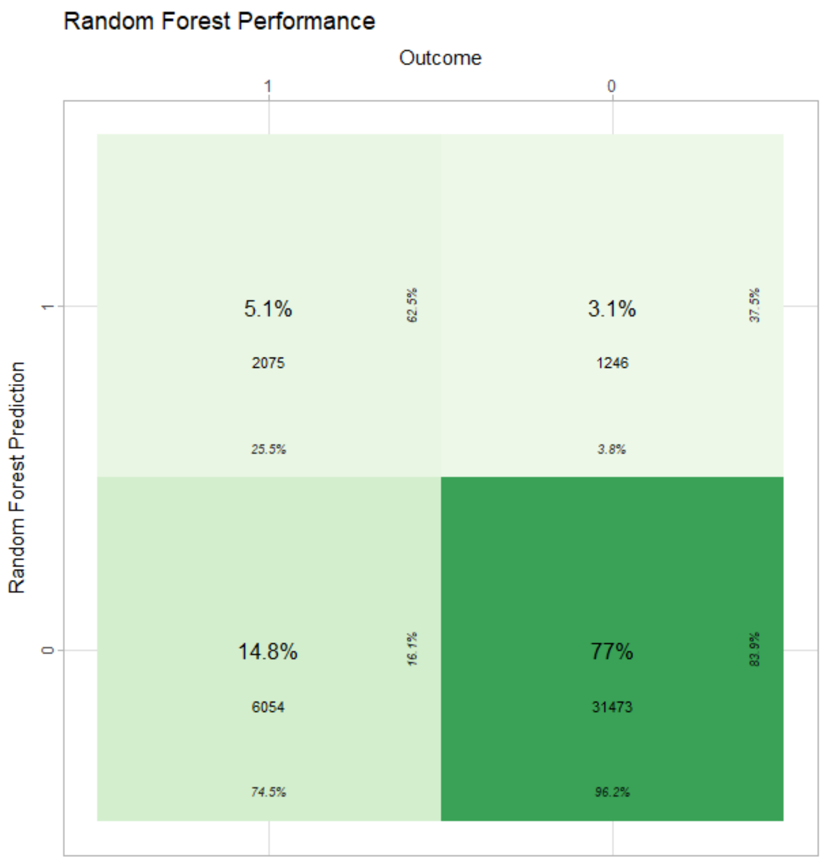
```
Type:                             Regression
Number of trees:                  1000
Sample size:                      108328
Number of independent variables:  13
Mtry:                             3
Target node size:                 5
Variable importance mode:         none
Splitrule:                        variance
OOB prediction error (MSE):       0.1027358
R squared (OOB):                  0.2706476
```

The Out of Bag (OOB) Error Estimate of the R- Square and Mean Square Error were determined to be 0.10 and 0.27, respectively. These measures tell us the model’s estimate on how it will perform on future data. An out of bag error estimate of 10% means that about 90% of the out of bag samples were correctly classified by the random forest. Using this model on test data (Test.NA) the results found that the AUC performance is high at 81– very similar to our logistic regression score! In order to circumvent these close estimates we may look at a precision-at-k plot which divides the total number of reports by the number of reports that were successfully predicted as a high outcome payout.

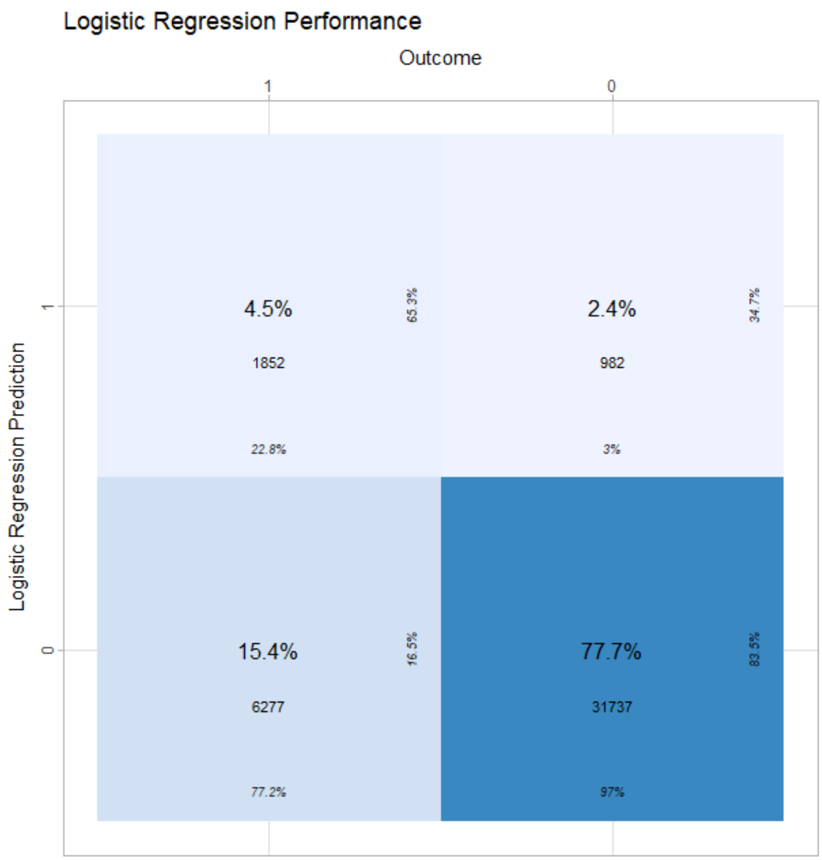


From our precision-at-k curves we may see– while still exceedingly similar in successful classification– the logistic regression curve lies slightly above the random forest curve. This suggests that we may confirm what our AUC scores were telling us– that the logistic regression model performed better in predicting a report with a high total payout.

In a further investigation we may look at the True Positive/Negative Rates of our models.



Random Forest Performance
Outcome
1
0
Random Forest Prediction
1
0
5.1%
2075
25.5%
62.5%
3.1%
1246
3.8%
37.5%
14.8%
6054
74.5%
16.1%
77%
31473
96.2%
83.9%



From these confusion matrices, we can see that while the logistic regression has the more accurate True Negative Rate (77.7%), the random forest actually has a higher True Positive Rate by 0.6%. Meanwhile, our random forest model also maintains a higher False Positive Rate and lower False Negative Rate in comparison to our logistic regression.

Alongside logistic regression, we decided to test these same variables using random forest. This machine learning method is generally very accurate by fitting many decision trees and bagging. Through this process random forest is also randomly sampling columns at each tree. We can then examine the results from the collection of trees. For a visual of the decision trees, see Figure 2 in the Appendix.

Conclusions

Our research team managed to take an original data set from NPDB, clean it, and process it to create predictive models. Through research on malpractice and investigation of various variables, the research question became clear: What are the best predictors for the probability of a malpractice case resulting in a higher payout? In the end, we fulfilled our research goals in predicting the probability of a high malpractice settlement payment as well as by naming our first logistic regression model as the most optimal for such prediction.

Our goal for this research is to provide valuable insights for insurance companies and practitioners. Through our data analysis we narrowed down on the various malpractice trends and we identified the contributing factors that lead to a high malpractice payment (over 500k).

Our findings indicate that the logistic model does a great job of predicting high malpractice payout. However, we may imagine some ways we can improve our model. As previously stated, data collected on malpractice payments does not provide a complete and accurate depiction of actual healthcare ineptitude or medical misconduct. Therefore, if we used NPDB data alongside more insightful studies, or surveys on malpractice cases, we could strengthen our model. Another way we could improve our model, albeit a computationally intensive one, would be to try all possible predictors and then perform a random forest for feature selection– our current model was purely made of predictors we suspected to be most important but there is a chance a better predictor, or predictors, exist in our data set.

The theoretical implication of these findings is that we can expect any trends in specific, or significant, predictors to have a considerable impact on the probability of a malpractice report resulting in a payout over half a million dollars. More consequentially, these implications also affect our stakeholders as they face high-stakes, including rising insurance rates and great monetary loss.

As far as our research goes, we can also imagine expanding upon this model and topic. It may be the case that some predictors are more important than others depending on the characteristics (e.g. a certain age group, or racial group) of the patient themselves. Findings of that nature would then have greater sociological implications. Collectively, the National Practitioner’s Data Bank has substantial research potential.

References

Chandra, A., Nundy, S., & Seabury, S. A. (2005). The Growth Of Physician Medical Malpractice Payments: Evidence From The National Practitioner Data Bank. *Health Affairs*, *24*, 240–249. https://doi-org.proxy.library.nyu.edu/10.1377/hlthaff.W5.240

Goode, V., Douglas, C., & Merwin, E. (2022). The National Practitioner Database Malpractice Study of Severity of Alleged Malpractice Injuries Trends 2008-2018. *American journal of medical quality : the official journal of the American College of Medical Quality*, *37*(5), 396–404. https://doi.org/10.1097/JMQ.0000000000000060

Legal malpractice claims costing more, settling sooner, research shows. Business Trial Group. (2016, October 27). https://www.businesstrialgroup.com/news/legal-malpractice-claims-costing-more-settling-sooner-research-shows/

NPDB research statistics. The NPDB - Research Statistics. (n.d.). https://www.npdb.hrsa.gov/resources/definitions.jsp

Romano M. (2005). System failure. Most U.S. hospitals have never filed a report with the databank that records doctor suspensions; critics say it's time for a new method. *Modern healthcare*, *35*(30), 6–1. https://www.modernhealthcare.com/article/20050725/NEWS/507250315/system-failure

Saber Tehrani, A. 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. https://pubmed.ncbi.nlm.nih.gov/23610443/

Schaffer, A. C., Jena, A. B., Seabury, S. A., Singh, H., Chalasani, V., & Kachalia, A. (2017). Rates and Characteristics of Paid Malpractice Claims Among US Physicians by

Specialty, 1992-2014. *JAMA internal medicine*, *177*(5), 710–718. https://doi.org/10.1001/jamainternmed.2017.0311

Schloemann, M. (2022, June 7). *Why did my medical malpractice insurance rate increase?: MEDPLI*. MEDPLI Professional Liability Insurance. https://medpli.com/why-did-my-medical-malpractice-insurance-rate-increase/

The NPDB - About Us. (n.d.). https://www.npdb.hrsa.gov/topNavigation/aboutUs.jsp

U.S. Department of Health and Human Services. (2023, February). *The NPDB - Public Use Data File*. National Practitioner Data Bank. https://www.npdb.hrsa.gov/resources/publicData.jsp

US medical malpractice case statistics. Knowledge Base Justpoint. (2022, November 2). https://justpoint.com/knowledge-base/us-medical-malpractice-case-statistics/

What is the NPDB?. The NPDB - What is the NPDB? Infographic.. https://www.npdb.hrsa.gov/resources/whatIsTheNPDB.jsp

Appendix

Figure 1. "What is the NPDB?" infographic from NPDB website

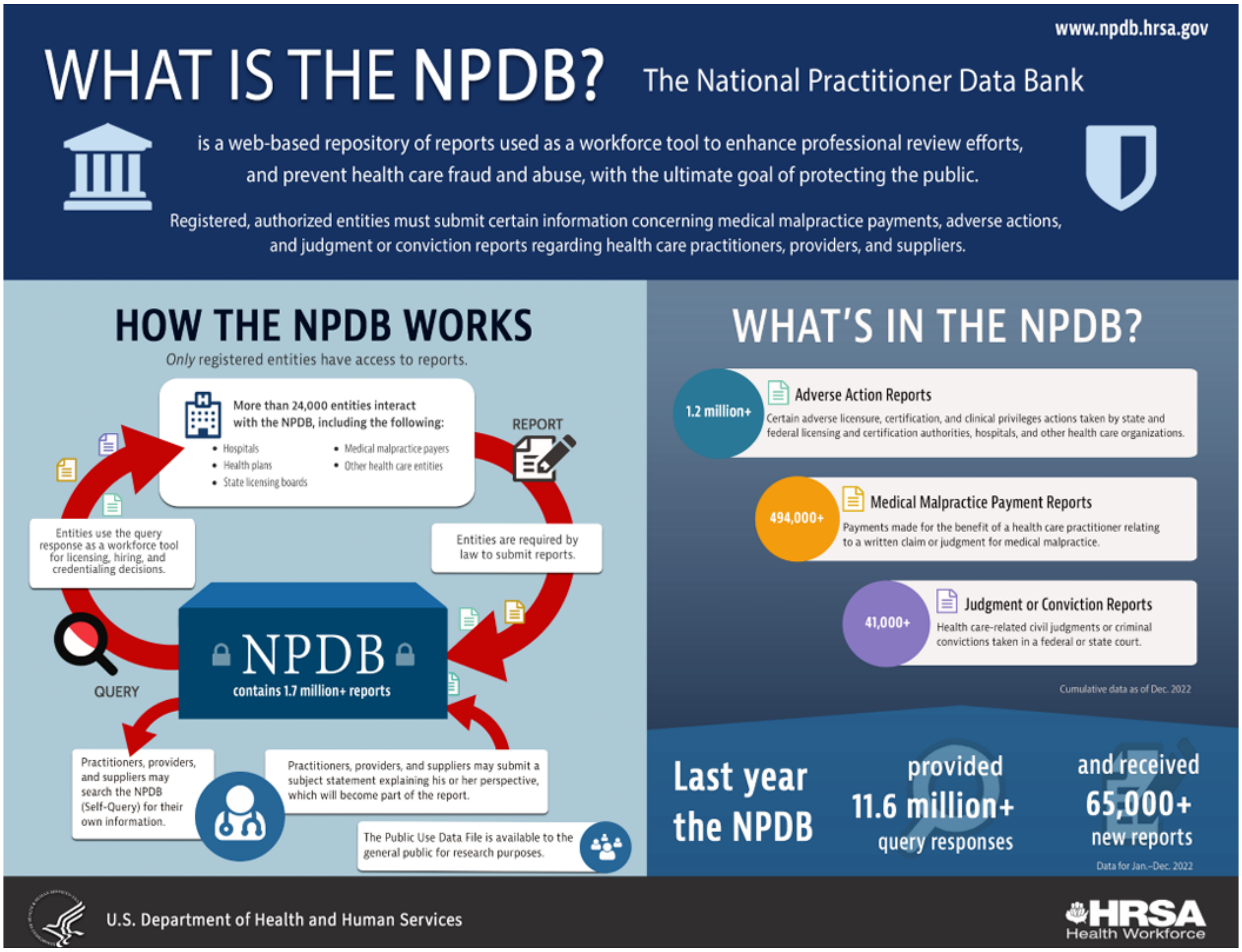


Figure 2. Random Forest. *Decision Trees*

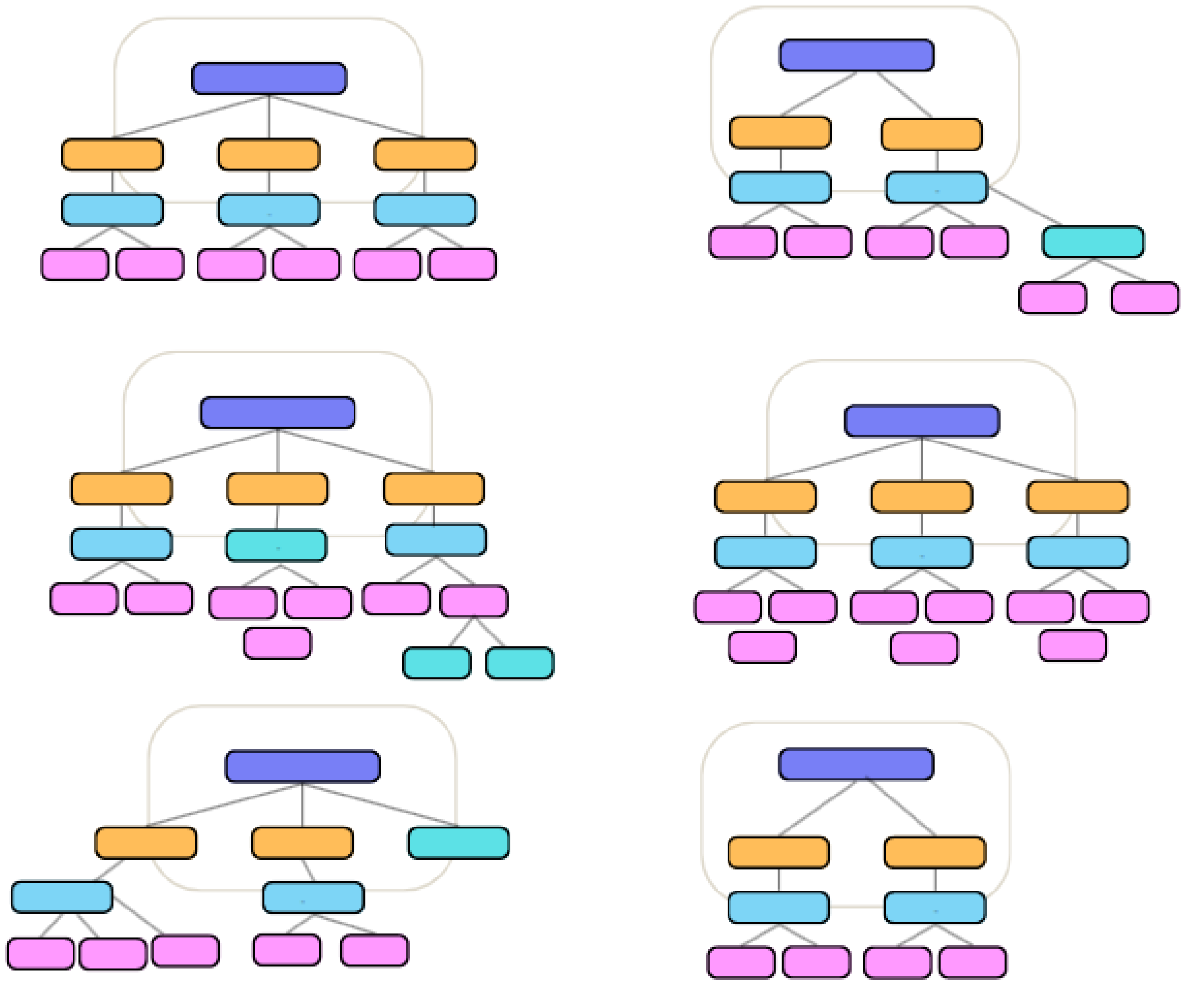


Table 1. Variables in the NPDB *variables highlighted were used in the final dataset

Variable	Type	Variable Label
seqno	Numeric	Sequence Number -- Unique to Each Record
rectype	String	Record Type
reptype	Numeric	Report Type
origyear	Numeric	Year original report processed
workstat	String	Practitioners Work State
workctry	String	Practitioners Work Country
homestat	String	Practitioners Home State
homectry	String	Practitioners Home Country
licnstat	String	Practitioners State of License (First Listed)
licnfeld	Numeric	Practitioners Field of License
practage	Numeric	Age Group of Practitioner
grad	Numeric	Graduation year group
algnnatr	Numeric	Malpractice Allegation Group
alegatn1	Numeric	Specific Malpractice Allegation 1
alegatn2	Numeric	Specific Malpractice Allegation 2
outcome	Numeric	Severity of Alleged Malpractice Injury [available for use 1/31/2004]
malyear1	Numeric	Year of Act or Omission 1
malyear2	Numeric	Year of Act or Omission 2
payment	Dollar	Payment Amount (this payment only)
totalpmt	Dollar	Total Payment by this Payer for This Practitioner [available for use 1/31/2004] *
paynumbr	String	Single of Multiple Payment *
numbprsn	Numeric	Number of Practitioners Payment For *
paytype	String	Payment A Result of ... *

pyrrltns	String	Relationship of Paying Entity to the Practitioner [available for use 1/31/2004] *

ptage	Numeric	Age Group of Patient [available for use 1/31/2004]
ptgender	String	Gender of Patient [available for use 1/31/2004]
pttype	String	Patient Type (Inpatient, Outpatient)
aayear	Numeric	Year of Adverse Action
aaclass1	Numeric	Adverse Action Classification1 [available for use 11/22/1999/old records converted]
aaclass2	Numeric	Adverse Action Classification2 [available for use 11/22/1999]
aaclass3	Numeric	Adverse Action Classification3 [available for use 11/22/1999]
aaclass4	Numeric	Adverse Action Classification4 [available for use 11/22/1999]
aaclass5	Numeric	Adverse Action Classification5 [available for use 11/22/1999]
basiscd1	String	Basis for Action [available for use 11/22/1999]
basiscd2	String	Basis for Action2 [available for use 11/22/1999]
basiscd3	String	Basis for Action3 [available for use 11/22/1999]
basiscd4	String	Basis for Action4 [available for use 11/22/1999]
basiscd5	String	Basis for Action5 [available for use 9/9/2002]
aalentyp	String	Adverse Action Length Type
aalength	Numeric	Length of Adverse Action Penalty, in Years
aaefyear	Numeric	Effective Year of Adverse Action
aasigyr	Numeric	Year of AA Report Signature
type	Numeric	Entity Type (assigned)
practnum	Numeric	Practitioner Number Unique to This File
accrrpts	Numeric	Subjects Number of Accreditation Reports
npmalrpt	Numeric	Practitioners Number of Malpractice Payment Reports

		submitted under Title IV.
nplicrpt	Numeric	Practitioners Number of Licensure Reports submitted under Title IV and/or Section 1921
npclprpt	Numeric	Practitioners Number of Clinical Privileges Reports submitted under Title IV
nppsmrpt	Numeric	Practitioners Number of Prof. Soc. Membership Reports submitted under Title IV
npdearpt	Numeric	Practitioners Number of DEA Reports submitted under Title IV
npexcrpt	Numeric	Practitioners Number of Exclusion Reports submitted under Title IV and/or Section 1921
npgarpt	Numeric	Practitioners Number of Government Administrative Reports submitted under Section 1921
npctmrpt	Numeric	Practitioners Number of Contract Termination Reports submitted under Section 1921

Table 2. Severity of Alleged Malpractice Injury. (Previously "OUTCOME" in NPDB file)

Value	**Label**
1	Emotional Injury Only
2	Insignificant Injury
3	Minor Temporary Injury
4	Major Temporary Injury
5	Minor Permanent Injury
6	Significant Permanent Injury
7	Major Permanent Injury
8	Quadriplegic, Brain Damage, Lifelong Care
9	Death

10	Cannot Be Determined from Available Records (These were recoded as "unable_tbd" cases in the final report)

Table 3. LCNFIELD_GRP categories. Based off original LICNFELD listing in the NPDB file.

GENERAL DOCTOR (MD)
10 Allopathic Physician (MD)
15 Physician Resident (MD)
1365 Medical Group/Practice
GENERAL DOCTOR (DO)
20 Osteopathic Physician (DO)
25 Osteopathic Physician Resident (DO)
DENSTRY
30 Dentist
35 Dental Resident
606 Dental Assistant
607 Dental Therapist/Dental Health Aide [available 6/15/09]
609 Dental Hygienist
612 Denturist
613 Other Dental Occupation - Not Classified, Specify [available 11/17/2014]
1362 Dental Group/Practice
PHARMACY
50 Pharmacist
55 Pharmacy Intern [available 9/9/2002]
60 Pharmacist, Nuclear
70 Pharmacy Assistant
75 Pharmacy Technician [available 9/9/2002]
76 Other Pharmacy Service Occupation - Not Classified, Specify [available

1345 Pharmacy
1346 Pharmaceutical Manufacturer
END OF LIFE
1382 Hospice/Hospice Care Provider
1398 End Stage Renal Disease Facility
NORMAL NURSE
100 Registered Nurse
110 Nurse Anesthetist
120 Nurse Midwife
NON-MD PRACTIONERS
130 Nurse Practitioner
134 Doctor of Nursing Practice [available 11/8/2010]
135 Advanced Nurse Practitioner [3/5/02 - 9/9/02]
141 Clinical Nurse Specialist [available 9/9/02]
642 Physician Assistant
645 Phys. Asst., Osteopathic

OTHER LICENSED NURSES
140 LPN or Vocational Nurse
142 Other Nurse Occupation - Not Classified, Specify [available 11/17/2014]
ASSISTANTING MED STAFF/CAREGIVERS
148 Certified Nurse Aide/Nursing Assistant [available 10/17/05]
150 Nurse Aide/Nursing Assistant
160 Home Health Aide (Homemaker)
165 Health Care Aide/Direct Care Worker [available 10/17/05]
618 Medical Assistant
699 Other Health Care Pract, Not Classified [available 11/22/99]
175 Certified or Qualified Medication Aide [available 10/17/05]
176 Other Aide Occupation - Not Classified, Specify [available 11/17/2014]
1353 Nursing/Health Care Staffing Service
1393 Home Health Agency/Organization
PSCYCHOLOGY
170 Psychiatric Technician
370 Clinical Psychologist [last use 9/9/02]
371 Psychologist [available 9/9/02]
372 School Psychologist [available 9/9/02]
373 Psychological Asst., Assoc., Examiner [available 9/9/02]
374 Other Psychologist/Psychological Assistant Occupation - Not Classified,

DIETIANS
200 Dietitian
210 Nutritionist
211 Other Dietitian/Nutritionist Occupation - Not Classified, Specify
EMTs
240 Emergency Medical Responder (EMR) [available 4/5/2019]
250 Emergency Medical Technician (EMT)
260 EMT, Cardiac/Critical Care
270 Advanced Emergency Medical Technician (AEMT)
280 Paramedic
281 Other Emergency Medical Services Provider - Not Classified, Specify
1390 Ambulance Service/Transportation Company

MENTAL HEALTH
300 Clinical Social Worker
621 Counselor, Mental Health
651 Prof. Counselor
652 Sex Offender Counselor [available 11/17/2014]
653 Pastoral Counselor [available 11/17/2014]
654 Prof. Cnslr., Alcohol
657 Prof. Cnslr., Family/Marriage
660 Addictions Counselor
661 Marriage and Family Therapist [available 9/9/02]
662 Art Therapist [available 11/17/2014]
668 Other Behavioral Health Occupation - Not Classified, Specify [available
1366 Mental Health/Substance Abuse Group/Facility
1383 Intermed. Care Fclty For Intellectually Disabled/Substance Abuse
1386 Residential Treatment Facility/Program
1388 Outpatient Rehab. Fclty./Comprehensive Outptnt. Rehab. Fclty
1303 Rehabilitation Hospital
1308 Rehabilitation Unit
1395 Mental Health Center/Community Mental Health Center
FEET
350 Podiatrist
639 Orthotics/Prosthetics Fitter
648 Podiatric Assistant
649 Other Podiatric Service Occupation - Not Classified, Specify [available
1364 Podiatric Group/Practice

HEARING
400 Audiologist
460 Speech/Language Pathologist
470 Hearing Aid/Instrument Specialist [available 10/17/05]
471 Other Speech, Language and Hearing Service Occupation - Not Classified,
REC THERAPY
402 Art/Recreation Therapist
665 Dance Therapist [available 11/17/2014]
667 Music Therapist [available 11/17/2014]
664 Recreation Therapist [available 11/17/2014]
PHYSICAL THERAPY
405 Massage Therapist
430 Physical Therapist
440 Phys. Therapy Assistant
1367 Physical/Occupational Therapy Group/Practice
OCC THERPAY
410 Occupational Therapist
420 Occup. Therapy Assistant
450 Rehabilitation Therapist
658 Other Rehabilitative, Respiratory and Restorative Service Occupation - Not
663 Respiratory Therapist
666 Resp. Therapy Technician

MED TECHS
500 Medical Technologist [changed to 501(6/15/09)]
501 Medical or Clinical Lab Technician/Technologist [available 6/15/09]
502 Medical/Clinical Lab Technician [available 6/15/09]
503 Surgical Technologist/Assistant [available 6/15/09]
504 Surgical Assistant [available 6/15/09]
505 Cytotechnologist [available 11/22/99]
510 Nuclear Med. Technologist
520 Rad. Therapy Technologist
530 Radiologic Technician/Technologist
540 X-Ray Technician or Operator [available 6/15/09]
550 Limited X-Ray Machine Operator [available 11/8/2010]
551 Other Technologist/Technician - Not Classified, Specify [available 11/17/2014]
647 Perfusionist [available 11/22/99]
1397 Mammography Service Provider
1399 Radiology/Imaging Center
LICENSED SPECIALIST
600 Acupuncturist
601 Athletic Trainer [available 11/22/99]
603 Chiropractor
604 Chiropractic Assistant [available 11/17/2014]
605 Other Chiropractic Occupation - Not Classified, Specify [available 11/17/2014]
1361 Chiropractic Group/Practice
NON_LICENSED_SPEC
615 Homeopath
624 Midwife, Lay (Non-Nurse)
627 Naturopath

EYES
630 Ocularist
633 Optician
636 Optometrist
637 Other Eye and Vision Service Occupation - Not Classified, Specify [available
1363 Optician/Optometric Group/Practice
ADMIN AND ASST_LIV_FACILITY
755 Hospital Administrator [available 11/22/99]
758 Health Care Facility Administrator [available 6/15/09]
800 Researcher, Clinical [available 11/22/99]
1370 Research Center/Facility
752 Adult Care Facility Administrator [available 11/22/99]
759 Assisted Living Facility Administrator [available 6/15/09]
1381 Adult Day Care Facility
1389 Nursing Facility/Skilled Nursing Facility

BIZ_OPS
810 Insurance Agent/Broker [available 11/22/99]
812 Insurance Broker [available 11/22/99]
820 Corporate Officer [available 11/22/99]
822 Business Manager [available 11/22/99]
830 Business Owner [available 11/22/99]
840 Salesperson [available 11/22/99]
850 Accountant [available 11/22/99]
853 Bookkeeper [available 11/22/99]
1320 Health Insurance Company/Provider
1331 Health Maintenance Organization
1351 Fiscal/Billing/Management Agent
1352 Purchasing Service

UNKNOWN_MISSING
998 Subject of Report Not Reportable (missing value)
999 Unspecified or Unknown
899 Other Individual, Not Classified [available 11/22/99]
GENERAL HOSPITALS
1301 General/Acute Care Hospital
1304 Federal Hospital
1391 Ambulatory Surgical Center
1392 Ambulatory Clinic/Center
PSCYHIATRIC FACILITY
1302 Psychiatric Hospital
1307 Psychiatric Unit
LABORATORY
1310 Laboratory/CLIA Laboratory
3RD_PARTY_ORGS
1335 Preferred Provider Organization
1336 Provider Sponsored Organization
1338 Religious, Fraternal Benefit Society Plan
SUPPLIERS_MANUFACTURERS
1342 Blood Bank
1347 Biological Products Manufacturer
1348 Organ Procurement Organization
1343 Durable Medical Equipment Supplier
1344 Eyewear Equipment Supplier
1349 Portable X-Ray Supplier

COMMUNITY HEALTH OUTREACH
1394 Health Cntr/Fedrlly. Qualified Hlth Cntr./Cmmnty Hlth Cntr.
1396 Rural Health Clinic
OTHER_NOTSPECIFIED_ORG
1999 Other Type not classified - specify
9999 Org. Type not specify