**NPDB Variable Definitions**

**Description of data and subsetting**

The National Practitioner Data Bank (NPDB) contains data about malpractice claims and adverse action events reported by insurance companies, hospitals, and other third parties, as required by law.

The complete data set includes 1,198,266 records, but includes several record types as below -

rectype # cases

Judgment or Conviction Report, 11/22/1999 and later 0

Adverse Action Report (Legacy Format) 51,301

Malpractice Payment Report, 1/31/04 and later 167,027

Malpractice Payment Report, 9/1/90 to 1/31/04 250,685

Consolidated Adverse Action Report, 11/22/1999 and later 729,253

Given the definitions according to the NPDB website, we feel the most complete and viable data to use is the subset for rectype=”Malpractice Payment Report, 1/31/04 and later”.

Based on this rectype, we are further focusing on doctors only – i.e. MD or DO designations.

licnfeld # cases

Allopathic Physician (MD) 119,501

Physician Resident (MD) 825

Osteopathic Physician (DO) 9,092

Osteopathic Physician Resident (DO) 135

Since we are interested in building a model that predicts malpractice payment severity (totalpmt), it makes sense to choose this for one allegation type because of the many factors. For this reason, we are choosing “Diagnosis Related” allegations, as they have the most within the physician subset. This also minimizes the number of alegatn1 values, which holds more specific allegation detail beneath Diagnosis Related allegations.

alegnntr # cases

Diagnosis Related 41,268

Surgery Related 34,551

Treatment Related 25,214

Obstetrics Related 9,952

Medication Related 6,709

We also thought it would be interesting to compare different regions of the country. We know that New York and California both have high numbers of claims, but wanted to increase number of records. Following a state by regions map from the “Physician Census 2014” article published by the Federation of State Medical Boards (source: <http://www.fsmb.org/Media/Default/PDF/Census/2014census.pdf>, accessed 5/9/2016), we chose the regions including California and New York. Each subgroup will perform analysis on the separate regions.

Region 1: WA, OR, CA, HI, AK (Pacific region)

Region 5: ME, VT, NH, MA, CT, RI, NY, PA, NJ (New England + Middle Atlantic)

**Final Subset Criteria**

* rectype = ”Malpractice Payment Report, 1/31/04 and later”
* licnfeld in ("Allopathic Physician (MD)","Physician Resident (MD)","Osteopathic Physician (DO)","Osteopathic Physician Resident (DO)")
* algnnatr = "Diagnosis Related"
* Region subsets
  + Region 1: licnstat in ("Hawaii", "Alaska", "Washington", "California", "Oregon")
  + Region 5: licnstat in ("Maine", "Vermont", "New Hampshire", "Massachusetts", "Connecticut", "Rhode Island", "New York", "Pennsylvania", "New Jersey")

**Dependent Variable**

* TotalPmt: total malpractice payment made per record

**Independent Variables**

For our final variables, we transformed some data into numerical values to leverage in our analysis, and limited allegations to the top 5 types.

* practage
  + Originally designated in bins, we substitute the bin mean in order to leverage the numerical value
* MDexp = malyear1 – grad
  + A new variable describing years of physician experience at the time of the malpractice event; we substituted the mean year of the grad bin to leverage the numerical number
* alegatn1 dummy variables (dx is a medical abbreviation for “diagnosis”)
  + a1failToDx = 1, if alegatn1 = “Failure to Diagnose”
  + a1delayInDx = 1, “Delay in Diagnosis”
  + a1wrongDx = 1, if alegatn1 = “Wrong or Misdiagnosis (e.g. Original Diagnosis is Incorrect)”
  + a1failToOrder = 1, if alegatn1 = “Failure to Order Appropriate Test”
  + a1radError = 1, if alegatn1 = “Radiology or Imaging Error”
* alegatn2 dummy variables (tx is a medical abbreviation for “treatment”)
  + a2failToTx = 1, if alegatn2 = “Failure to Treat”
  + a2delayInTx = 1, if alegatn2 = “Delay in Treatment”
  + a2delayInDx = 1, if alegatn2 = “Delay in Diagnosis”
  + a2failToDx = 1, if alegatn2 = “Failure to Diagnose”
  + a2failToTest = 1, if alegatn2 = “Failure to Order Appropriate Test”
* outcome
  + Because outcome is rated on an ordered scale – low (1) to high (9), we can use the numerical values. We remove any case where the outcome is unknown (10).
* ptage
  + We substitute the bin mean of patient age in order to leverage the numerical value
* ptgender
  + ptMale = 1, where ptgender = “male”
  + ptFemale = 1, where ptgender = “female”
  + omitted if ptgender is unknown
* pttype – based on pttype, we create new variables:
  + inpatient = 1, if pttype=”inpatient” or pttype = “both”
  + outpatient = 1, if pptype=”outpatient” or pttype = “both”
  + omitted if pttype is unknown

Our final subsets have 19 variables – 5 numerical variables and 14 dummy variables.