National Inpatient Sample: Big Data Issues

M B Rao
Division of Biostatistics and Epidemiology
And
Department of Biomedical Engineering
University of Cincinnati

A Seminar Delivered
Under the Aegis of BERD
And
University of Cincinnati Children’s Hospital
Department of Biostatistics and Epidemiology

June 10, 2014
1. Exordium

The amount of money spent on health care runs into trillions of dollars seemingly out of control. A question arose how much money Americans spend on being treated in hospitals.

2. Big Data

Nationwide Inpatient Sample (NIS)

The Healthcare Cost and Utilization Project (HCUP) is funded by the Agency for Healthcare Research and Quality (AHRQ). Federal and State Governments along with Industry provide money to AHRQ. The Nationwide Inpatient Sample (NIS) is one of the major databases compiled and maintained by the HCUP.

What is NIS?

It is the largest all-payer inpatient care database in the United States. NIS data are available from 1988 to 2011 (24 years). If one wants to examine trend over time, one needs at least 20 years data. This data base is adequate to examine the trend of any phenomenon of interest over time with reference to hospital admissions.

3. Sampling frame and strata

4. Structure of the data

5. Variables of Interest

6. Output

7. Future Work

8. Excursus

This is an example of Big Data.

What is Big Data?
In Statistics departments, traditionally, they deal with ‘small n – small p’ data. (n is the number of observations and p is the number of variables.) A new discipline emerged, namely Bioinformatics, to handle ‘small n – large p’ data. (Genome Wide Association Data, Gene Expression Data, Protein Expression Data, Metabolomics, etc.) ‘Large n’ data come under the purview of Big Data or Data Science.

In 2013, ~ 3000 exabytes of data existed on the internet. Of the data that exists in the world now, 90% was created in the last two years. The growth is exponential with an estimated growth rate of 10%. (Source: Dr. Eric Rozier, Head of the Trustworthy Systems Engineering Laboratory, Coral Gables, FL.)

Basic Unit of Data: a Byte

KB (Kilobyte) $10^3$ bytes

MB (Megabyte) $10^6$ bytes

GB (Gigabyte) $10^9$ bytes

TB (Terabyte) $10^{12}$ bytes

PB (Petabyte) $10^{15}$ bytes

EB (Exabyte) $10^{18}$ bytes

ZB (Zettabyte) $10^{21}$ bytes

YB (Yottabyte) $10^{24}$ bytes

XB (Xenottabyte) $10^{27}$ bytes

SB (Shiletnobyte) $10^{30}$ bytes

DB (Domegemegrottebyte) $10^{33}$ bytes

How do we handle vast data sets?

We need a fusion of Statistics, Computer Science, and Mathematics. NSF and NIH created special divisions to encourage proposals on big data.
A word of exhortation from Bin Yu, Berkeley, ex-president of the Institute of Mathematical Statistics:

Statisticians are data scientists, but so are other people from Computer Science, Electrical Engineering, Applied Mathematics, Physics, Biology, and Astronomy. In my view, the key factor of gain success in data science is human resource: we need to improve our interpersonal, leadership, and coding skills. There is no doubt that our expertise is needed for all big data projects, but if we do not rise to the big data occasion to take leadership in the big data projects, we will likely become secondary to other data scientists with better leadership and computing skills. We either compute or concede.

What is going on in our neighborhood?

1. University of Northern Kentucky is now offering a Bachelor’s degree program in Data Science.
2. Ohio State University has created a new department of data science offering graduate degree programs in data science.
3. Computer Science Department and Business School at UC are offering a 20-credit certificate program in Big Data.
4. Division of Epidemiology and Biostatistics at UC is contemplating a Ph.D. program with Big Data track.
5. I am offering a 3-credit class on ‘Introduction to Data Science’ next Spring semester.

Back to NIS data …

**Population and Sampling Scheme**

**Year 2008**

The basic sampling unit for this project is a hospital admission and discharge, called ‘episode,’ in every year of interest. Consequently, information about the episodes should come from our hospitals. The population of interest is the collection of all episodes. Episodes that occurred in VA hospitals were excluded. Episodes that occurred in hospitals in the Indian Reservations were excluded.
Some states did not participate in the study. Of course those states’ hospitals were excluded. We modify the definition of our population. The population of interest is all episodes in all hospitals excluding those mentioned. The size of the population is about 95% of all episodes that occurred in all the hospitals.

The goal is to draw a 20% random sample of episodes. With an estimated number of episodes to be about 40,000,000, the task of drawing a sample is daunting. A simple random sample is not practical. For a simple random sample, one needs to number the episodes serially and then set about drawing a random sample of about 8,000,000 episodes. Implementation is impossible. HCUP followed a stratified cluster random sampling method. From the view point of getting a representative sample and better inference, stratified random sampling beats simple random sampling heads and shoulder. A stratified random sampling scheme can be devised in many different ways. The basic idea is to divide the entire population into strata in an illuminating way, and then draw a random sample from each stratum.

HCUP sampling procedure

There were 4,310 hospitals in the United States excluding VA hospitals, Indian Healthcare hospitals, and those hospitals that belong to states which did not participate. Stratification was done on hospitals. A 20% sample of hospitals amounted to 862 hospitals. Stratification was done with respect to 4 categorical variables on the hospitals.

A. Geographic region
   1. Northeast
   2. Midwest
   3. West
   4. South
B. Control
   0. Government or Private
   1. Government, nonfederal
   2. Private, not-for-profit
   3. Private, investor-owned
4. Private, either not-for-profit or investor-owned

C. Location/Teaching
   1. Rural
   2. Urban nonteaching
   3. Urban teaching

D. Bedsize
   1. Small
   2. Medium
   3. Large

Identify all hospitals that fit the description of one level of each categorical variable. For example, the symbol 1311 indicates all those hospitals located in the Northeast, private (investor-owned), rural and with a small number of beds. This is one stratum.

Total number of strata: $4 \times 5 \times 3 \times 3 = 180$. In some strata, there were no hospitals or very few hospitals. Some of these strata were merged. The final tally of strata was 60. In other words, all hospitals were segregated into 60 strata.

From each stratum of hospitals, a 20% random sample of hospitals was chosen. For this they have used systematic sampling. How does this work? Suppose a stratum has 100 hospitals listed in some order. We want a sample of 20 hospitals. Choose a number at random from 1 to 5. Suppose we get 4. Choose the 4th hospital in the list, then 9th, 14th, etc.

All the episodes in the chosen hospitals constitute HCUP sample. Each of the hospitals in the sample collected data on each inpatient admission.

Information sought is divided into four groups.

   1. Core information
      a. Date of admission
      b. Date of discharge
      c. LOS (length of stay)
d. Reason for admission (coded-APSDRG)
e. Co-morbidities (coded-APSDRG)
f. Insurance details
g. Cost of stay
h. Zip code of his hospital
i. ICD-9 code
j. Etc.

2. Groups
3. Severity
4. Hospitals

I have looked at 2008 NIS data.

The data come in 4 Ascii files.

<table>
<thead>
<tr>
<th>ASCII File Name</th>
<th># episodes</th>
<th># variables</th>
<th>File size</th>
<th>Primary focus of data</th>
</tr>
</thead>
<tbody>
<tr>
<td>Or records</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2008_NIS_Core</td>
<td>8,158,381</td>
<td>135</td>
<td>2.77 GB</td>
<td>Patient</td>
</tr>
<tr>
<td>2008_NIS_DX_PR_GRPS</td>
<td>8,158,381</td>
<td>47</td>
<td>490 MB</td>
<td>Disease</td>
</tr>
<tr>
<td>2008_NIS_Severity</td>
<td>8,158,381</td>
<td>40</td>
<td>850 MB</td>
<td>Severity</td>
</tr>
<tr>
<td>2008_NIS_Hospitals</td>
<td>1,056</td>
<td>33</td>
<td>205 KB</td>
<td>Hospitals</td>
</tr>
</tbody>
</table>

The data are not free. One can buy any particular year’s data.

Cost:

Student: $ 50

Non-student: $ 250

When you buy the data, you get the data in two CDs and an information booklet.

One can buy all years data.
Cost:

Student: $250

Non-student: $3000

**DRG code**

This is one of the variables in the data set. For every patient admitted, the hospital determines for what medical condition the patient is treated most predominantly, codified from 001 to 999. DRG = 103 means Headache without complications. DRG code classifies the medical conditions into 999 categories. This coding is specific to our hospitals. Internationally, ICD-9 code (~ 17,000 medical conditions) is used to codify medical conditions.

**ICD-10 codes (~ 180,000 medical conditions)**

**An illustration**

A Master’s student, Xin Wang, is interested on blood disorders for her thesis.

**DRG codes:** 811 = Blood Disorders without complications

812 = Blood Disorders with complications

Year of interest: 2009

Total Number of Episodes: 7,810,762

Number of episodes with DRG = 811 or 812: 62,853

Extract this particular subset from the entire 2009 data.

```r
> RBCD2009<-read.csv("J:/NISDATARBCD/RBCD2009.csv")
> dim(RBCD2009)
[1] 62853   187
> RBCD2010<-read.csv("J:/NISDATARBCD/RBCD2010.csv")
```
> dim(RBCD2010)
[1] 67964  186

> RBCD2011<-read.csv("J:/NISDATARBCD/RBCD2011.csv")
> dim(RBCD2011)
[1] 69264  186

Summary

<table>
<thead>
<tr>
<th>Total</th>
<th>Disorders</th>
<th>Percent</th>
</tr>
</thead>
<tbody>
<tr>
<td>7,810,762</td>
<td>62,853</td>
<td>0.80</td>
</tr>
<tr>
<td>7,800,441</td>
<td>67,964</td>
<td>0.87</td>
</tr>
<tr>
<td>8,023,590</td>
<td>69,264</td>
<td>0.86</td>
</tr>
</tbody>
</table>

What are the variables in the data set?

187 variables

Documentation is available at the HCUP website.

> RBCD2009[1,]

   HOSPID AGE AGEDAY AMONTH ASOURCE ASOURCEUB92 ASOURCE_X ATYPE AWEEKEND DIED DISCWT DISPUB04
1  4005  75 NA     3 NA                      2  0  0 5.346624        1
   DISPUNIFORM DQTR DQTR_X DRG DRG24 DRGVER DRG_NoPOA DSHOSPID D DX1 DX2 DX3 DX4 DX5 DX6
1    1  1  1     812  395 812 MED0204 2800 5789 25000 2111 53011 4580
   DX7 DX8 DX9 DX10 DX11 DX12 DX13 DX14 DX15 DX16 DX17 DX18 DX19 DX20 DX21 DX22 DX23 DX24
1 V5861 V4501 V5866 V5863
   DX25 DXCCS1 DXCCS2 DXCCS3 DXCCS4 DXCCS5 DXCCS6 DXCCS7 DXCCS8 DXCCS9 DXCCS10 DXCCS11 DXCCS12
1  59  153  49  47  138  117  257  105  257  257  NA  NA
   DXCCS13 DXCCS14 DXCCS15 DXCCS16 DXCCS17 DXCCS18 DXCCS19 DXCCS20 DXCCS21 DXCCS22 DXCCS23
1 NA NA NA NA NA NA NA NA NA NA NA NA
   DXCCS24 DXCCS25 ECODE1 ECODE2 ECODE3 ECODE4 ELECTIVE E_CCS1 E_CCS2 E_CCS3 E_CCS4 FEMALE
1 NA NA  9342  8490  8798  8497  0  2617  2621  2616  2621  1
Blood disorder is a broad name. Identify the disease precisely. Get the ICD-9 code. The decimal point is missing in the data.

> ICD9<-table(RBCD2009$DX1)

> ICD9

2800 2801 2808 2809 2810 2811 2812 2813 2818 2819 2820 2821 2822 2823 2825 2827
7674 58 302 8939 166 194 53 42 7 600 65 5 12 7 26 13
2828 2829 2841 2850 2851 2853 2858 2859 2897 9996 23872 23873 23874 23875 28241 28242
6 3 2844 39 4856 133 736 11383 47 1 168 107 32 1644 28 538
28249 28260 28261 28262 28263 28264 28265 28269 28311 28521 28522 28529 79001 79009 99989
110 401 208 13807 27 432 41 540 109 2058 2040 2129 11 1 211
Out of 62,853 patients admitted under the broad name of blood disorders, 13,807 of them had ICD-9 code 282.62. This is the most predominant blood disorder.

**Top 5 blood disorders**

1. **282.62**  Hb-SS disease with crisis  
2. **285.9**  Anemia unspecified  
3. **280.9**  Iron deficiency anemia unspecified  
4. **280.0**  Iron deficiency anemia secondary  
5. **285.1**  Acute posthemorrhagic anemia  

The same five ICD-9 codes showed up in the same order as top five for the years 2010 and 2011.

What about gender distribution?

```r
> RBCD2009F<-table(RBCD2009$FEMALE)
> RBCD2009F
     0     1
25008 37743
> RBCD2010F<-table(RBCD2010$FEMALE)
> RBCD2010F
```
Gender Distribution

<table>
<thead>
<tr>
<th>Year</th>
<th>Total Cases</th>
<th>Males</th>
<th>Females</th>
</tr>
</thead>
<tbody>
<tr>
<td>2009</td>
<td>62,853</td>
<td>25,008 (40%)</td>
<td>37,743 (60%)</td>
</tr>
<tr>
<td>2010</td>
<td>67,964</td>
<td>27,175 (40%)</td>
<td>40,700 (60%)</td>
</tr>
<tr>
<td>2011</td>
<td>69,264</td>
<td>27,106 (39.2%)</td>
<td>42,078 (60.8%)</td>
</tr>
</tbody>
</table>

Age distribution

> summary(RBCD2009$AGE)

<table>
<thead>
<tr>
<th>Min.</th>
<th>1st Qu.</th>
<th>Median</th>
<th>Mean</th>
<th>3rd Qu.</th>
<th>Max.</th>
<th>NA's</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.00</td>
<td>36.00</td>
<td>60.00</td>
<td>56.25</td>
<td>78.00</td>
<td>109.00</td>
<td>38</td>
</tr>
</tbody>
</table>

> summary(RBCD2010$AGE)

<table>
<thead>
<tr>
<th>Min.</th>
<th>1st Qu.</th>
<th>Median</th>
<th>Mean</th>
<th>3rd Qu.</th>
<th>Max.</th>
<th>NA's</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.00</td>
<td>32.00</td>
<td>58.00</td>
<td>54.58</td>
<td>77.00</td>
<td>106.00</td>
<td>52</td>
</tr>
</tbody>
</table>

> summary(RBCD2011$AGE)

<table>
<thead>
<tr>
<th>Min.</th>
<th>1st Qu.</th>
<th>Median</th>
<th>Mean</th>
<th>3rd Qu.</th>
<th>Max.</th>
<th>NA's</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.00</td>
<td>36.00</td>
<td>61.00</td>
<td>56.69</td>
<td>78.00</td>
<td>109.00</td>
<td>30</td>
</tr>
</tbody>
</table>

Age distribution gender-wise

> RBCD2009FEMALE<- subset (RBCD2009, RBCD2009$FEMALE==1)
> dim(RBCD2009FEMALE)
[1] 37743 187
> summary(RBCD2009FEMALE$AGE)

       Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
 0.00 37.00  60.00  56.92  78.00 109.00       5

> RBCD2009MALE<-subset(RBCD2009, RBCD2009$FEMALE==0)
> summary(RBCD2009MALE$AGE)

       Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
 0.00 32.00  60.00  55.31  77.00 104.00       1

Look at the histogram of the age distribution of females for the year 2009.
Analysis depends on your imagination and questions you raise ...

**What did I do with the data?**

I started working on the data in collaboration with Dr. Ravi Chinta, Associate Professor, Xavier University. We cannot handle all episodes (over 8 millions) at the same time. Right from the beginning we wanted to focus on one medical condition. We settled for ‘Headache (DRG code = 103)’ and ‘Headache With Complications (DRG code = 102).’ Isolating the episodes pertaining to these conditions netted us over 18,381 episodes for the year 2008. This is the segment
of data we wanted to study. The first thing we did was to convert the Ascii data into SPSS files to SAS files to R files.

What is needed to work on such a project?

1. Dexterity with some computing package.
2. A reasonable grounding in sample survey methodology.

Using data from a stratified random sample, one needs to know how to estimate population parameters and provide confidence intervals. A book by Paul Levy and Stanley Lemeshow (Sampling of Populations, Wiley 1991) is helpful. The booklet by HCUP and the website are helpful in explaining how to build national estimates.

We examined a number of variables and their distributions.

1. Gender
2. Distribution of Gender state by state
3. Distribution of Gender region by region (Northeast; south; Midwest; west)
4. Stratum estimates; national estimates
5. LOS (length of stay)
6. LOS national
7. LOS state by state
8. LOS region by region
9. Average cost per day national
10. Average cost state by state
11. Average cost region by region
12. Who paid?
13. Age national
14. Age state by state
15. Age region by region
16. Headaches versus total nationwide
17. Headaches versus total state by state
18. Headaches versus total region by region
19. Etc.
Goal: **Estimate the distribution of Gender suffering from headache nationally**

Step 1:

Estimate the distribution of Gender stratum by stratum.

<table>
<thead>
<tr>
<th>Stratum</th>
<th>Male</th>
<th>Female</th>
<th>Total</th>
<th>Male%</th>
<th>Female%</th>
<th>Weight</th>
</tr>
</thead>
<tbody>
<tr>
<td>1011</td>
<td>3</td>
<td>13</td>
<td>16</td>
<td>18.8%</td>
<td>81.3%</td>
<td>6.54</td>
</tr>
<tr>
<td>1012</td>
<td>11</td>
<td>19</td>
<td>30</td>
<td>36.7%</td>
<td>63.3%</td>
<td>6.10</td>
</tr>
</tbody>
</table>

Etc.

Note: The weights are proportional to the size (Total Number of Episodes) of the strata. These weights are provided by HCUP. When we want to get a national estimate of the distribution of Gender, we need to calculate the weighted average of strata distributions.

**National estimate**

Gender: Male Female

Percentage: 25.6% 74.4%

Goal: **How the distribution of the gender varies from state to state?**

Each hospital is identified by the state in which it is located. Pull out all the episodes that occurred in all hospitals in the state of interest.

<table>
<thead>
<tr>
<th>State</th>
<th>Male</th>
<th>Female</th>
<th>Total</th>
<th>Male%</th>
<th>Female%</th>
</tr>
</thead>
<tbody>
<tr>
<td>AR</td>
<td>67</td>
<td>177</td>
<td>244</td>
<td>27.5%</td>
<td>72.5%</td>
</tr>
<tr>
<td>AZ</td>
<td>130</td>
<td>360</td>
<td>490</td>
<td>26.5%</td>
<td>73.5%</td>
</tr>
</tbody>
</table>

Etc.

A technical note: Recording the data state by state is also stratification. This is post-stratification. One can use the post-stratified data to get a national estimate of the distribution of gender suffering from headaches. This is not a problem. The
daunting task is to obtain standard errors. The methodology comes under ‘Domain Analysis.’

Here is the bar plot of percentage of women with headache admitted to hospital state by state and sorted from the lowest to the highest.

There is some variation in the percentage of women with headache admitted to hospital, with the least percentage from Vermont at 50% and highest in South Dakota at 85.7%.

Let us look at regional variations.

<table>
<thead>
<tr>
<th>Region</th>
<th>Male</th>
<th>Female</th>
<th>Total</th>
<th>Male%</th>
<th>Female%</th>
</tr>
</thead>
<tbody>
<tr>
<td>Northeast</td>
<td>1071</td>
<td>2932</td>
<td>4003</td>
<td>26.8%</td>
<td>73.2%</td>
</tr>
<tr>
<td>Midwest</td>
<td>997</td>
<td>2828</td>
<td>3925</td>
<td>25.4%</td>
<td>74.6%</td>
</tr>
<tr>
<td>South</td>
<td>1909</td>
<td>5696</td>
<td>7605</td>
<td>25.1%</td>
<td>74.9%</td>
</tr>
<tr>
<td>West</td>
<td>739</td>
<td>2011</td>
<td>2750</td>
<td>26.9%</td>
<td>73.1%</td>
</tr>
</tbody>
</table>
Inter-regional variation is not much.

Goal: Examine the length of stay

1. The length of stay varied from 0 to 62 days. The length is 0 means that the person was discharged on the same day.
2. The mean length of stay is 2.68 days.
3. The mean length of stay for males is 2.50 days.
4. The mean length of stay for females is 2.75 days.

The distribution of the length of stay in the hospital is given below.

<table>
<thead>
<tr>
<th>Length</th>
<th>Total Frequency</th>
<th>Male Frequency</th>
<th>Female Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0</td>
<td>738</td>
<td>226</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>4894</td>
<td>1477</td>
</tr>
<tr>
<td>3</td>
<td>2</td>
<td>5274</td>
<td>1365</td>
</tr>
<tr>
<td>4</td>
<td>3</td>
<td>3166</td>
<td>729</td>
</tr>
<tr>
<td>5</td>
<td>4</td>
<td>1688</td>
<td>382</td>
</tr>
<tr>
<td>6</td>
<td>5</td>
<td>1019</td>
<td>213</td>
</tr>
<tr>
<td>7</td>
<td>6</td>
<td>572</td>
<td>116</td>
</tr>
<tr>
<td>8</td>
<td>7</td>
<td>330</td>
<td>79</td>
</tr>
<tr>
<td>9</td>
<td>8</td>
<td>187</td>
<td>39</td>
</tr>
<tr>
<td>10</td>
<td>9</td>
<td>129</td>
<td>20</td>
</tr>
<tr>
<td>11</td>
<td>10</td>
<td>84</td>
<td>17</td>
</tr>
<tr>
<td>12</td>
<td>11</td>
<td>51</td>
<td>12</td>
</tr>
<tr>
<td>13</td>
<td>12</td>
<td>36</td>
<td>8</td>
</tr>
<tr>
<td>14</td>
<td>13</td>
<td>28</td>
<td>6</td>
</tr>
<tr>
<td>15</td>
<td>14</td>
<td>20</td>
<td>9</td>
</tr>
<tr>
<td>Length</td>
<td>MalePer</td>
<td>FemalePer</td>
<td></td>
</tr>
<tr>
<td>--------</td>
<td>---------</td>
<td>-----------</td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>4.79</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>31.32</td>
<td></td>
</tr>
</tbody>
</table>

The Distribution of the Length of Stay Gender-wise in Percentages

<table>
<thead>
<tr>
<th>Length</th>
<th>Male ePer</th>
<th>Female ePer</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>4.79</td>
<td>3.77</td>
</tr>
<tr>
<td>2</td>
<td>31.32</td>
<td>25.19</td>
</tr>
</tbody>
</table>
A bar plot

Goal: How much each patient was charged?

A column in the data with the heading ‘TOTCHG’ gives total charge levied for each episode. This is what we did with this column.
1. Look at all the episodes in which the patient was discharged on the same day. Take the average of all charges levied.
2. Look at all the episodes in which the patient stayed for one day. Take the average of all charges levied.
3. Look at all the episodes in which the patient stayed for two days. Calculate the charge per day for each patient. Then average.
4. And so on.

The standard deviation of these per day total charges is also calculated. Is it the best way to convey the cost of staying in a hospital when the ailment is headache?

<table>
<thead>
<tr>
<th>No. of Days Stayed</th>
<th>Total Charge Per Day</th>
<th># Episodes</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>9370</td>
<td>729</td>
</tr>
<tr>
<td>1</td>
<td>10565</td>
<td>4831</td>
</tr>
<tr>
<td>2</td>
<td>6839</td>
<td>5237</td>
</tr>
<tr>
<td>3</td>
<td>5477</td>
<td>3139</td>
</tr>
<tr>
<td>4</td>
<td>5076</td>
<td>1695</td>
</tr>
<tr>
<td>5</td>
<td>4642</td>
<td>1016</td>
</tr>
<tr>
<td>6</td>
<td>4559</td>
<td>568</td>
</tr>
<tr>
<td>7</td>
<td>4201</td>
<td>322</td>
</tr>
<tr>
<td>8</td>
<td>4199</td>
<td>184</td>
</tr>
<tr>
<td>9</td>
<td>4045</td>
<td>129</td>
</tr>
<tr>
<td>10</td>
<td>4134</td>
<td>84</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>---</td>
<td>-----</td>
<td>-----</td>
</tr>
<tr>
<td>11</td>
<td>3898</td>
<td>51</td>
</tr>
<tr>
<td>12</td>
<td>4027</td>
<td>36</td>
</tr>
<tr>
<td>13</td>
<td>3549</td>
<td>28</td>
</tr>
<tr>
<td>14</td>
<td>4161</td>
<td>19</td>
</tr>
<tr>
<td>15</td>
<td>4172</td>
<td>13</td>
</tr>
<tr>
<td>16</td>
<td>4660</td>
<td>8</td>
</tr>
<tr>
<td>17</td>
<td>3544</td>
<td>12</td>
</tr>
<tr>
<td>18</td>
<td>3696</td>
<td>8</td>
</tr>
<tr>
<td>19</td>
<td>4974</td>
<td>4</td>
</tr>
<tr>
<td>20</td>
<td>4362</td>
<td>2</td>
</tr>
<tr>
<td>22</td>
<td>1824</td>
<td>1</td>
</tr>
<tr>
<td>23</td>
<td>5433</td>
<td>4</td>
</tr>
<tr>
<td>24</td>
<td>3634</td>
<td>1</td>
</tr>
<tr>
<td>26</td>
<td>2897</td>
<td>2</td>
</tr>
<tr>
<td>27</td>
<td>4064</td>
<td>1</td>
</tr>
<tr>
<td>28</td>
<td>8092</td>
<td>1</td>
</tr>
<tr>
<td>29</td>
<td>1455</td>
<td>1</td>
</tr>
<tr>
<td>30</td>
<td>6580</td>
<td>2</td>
</tr>
<tr>
<td>31</td>
<td>3062</td>
<td>2</td>
</tr>
<tr>
<td>35</td>
<td>3803</td>
<td>1</td>
</tr>
<tr>
<td>36</td>
<td>6981</td>
<td>1</td>
</tr>
<tr>
<td>37</td>
<td>12018</td>
<td>1</td>
</tr>
</tbody>
</table>
What factors influence these charges?

One strong predictor is the number of co-morbidities each episode entails.

<table>
<thead>
<tr>
<th>No. of co-morbidities</th>
<th># Episodes</th>
<th>Percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>5577</td>
<td>30.3</td>
</tr>
<tr>
<td>1</td>
<td>5371</td>
<td>29.2</td>
</tr>
<tr>
<td>2</td>
<td>3829</td>
<td>20.8</td>
</tr>
<tr>
<td>3</td>
<td>2131</td>
<td>11.6</td>
</tr>
<tr>
<td>4</td>
<td>947</td>
<td>5.2</td>
</tr>
<tr>
<td>5</td>
<td>374</td>
<td>2.0</td>
</tr>
<tr>
<td>6</td>
<td>104</td>
<td>0.6</td>
</tr>
<tr>
<td>7</td>
<td>35</td>
<td>0.2</td>
</tr>
<tr>
<td>8</td>
<td>10</td>
<td>0.1</td>
</tr>
<tr>
<td>9</td>
<td>3</td>
<td>0.0</td>
</tr>
</tbody>
</table>

**Total 18381 100.0**

**Gender distribution (percentages)**

<table>
<thead>
<tr>
<th>Year</th>
<th>Females</th>
<th>Males</th>
</tr>
</thead>
<tbody>
<tr>
<td>2005</td>
<td>54</td>
<td>46</td>
</tr>
<tr>
<td>2006</td>
<td>55</td>
<td>45</td>
</tr>
<tr>
<td>2007</td>
<td>65</td>
<td>35</td>
</tr>
<tr>
<td>2008</td>
<td>74</td>
<td>26</td>
</tr>
<tr>
<td>2009</td>
<td>73</td>
<td>27</td>
</tr>
</tbody>
</table>
What’s going on?

**Challenging problems**

1. Trend analysis
2. Incidence of headaches in relation to total number of admissions – stratum by stratum – state by state. What is the trend like?
3. When data collection began in 1988 only 8 states participated in the survey. In 2008, 42 states participated. In 2009, 44 states participated. The size of the target population is not the same over the years.
4. Integrating two or more data sets.

**Elaboration of Idea 4**

We have HCUP data.

EPA has \( \text{PM}_{2.5} \) Concentration data.

There are more than 1000 monitors around the country monitoring \( \text{PM}_{2.5} \) (Particulate Matter 2.5). At each site, how much \( \text{PM}_{2.5} \) accumulated is measured 4 times a day every day.

Here is the idea.

Is headache environmental?

Look at an episode → Look up the Zip code

→ Identify all monitors within 6 mile radius of the zip code

→ Average \( \text{PM}_{2.5} \) concentrations from all the monitors over the previous ten days from the date of admission

Case - Headache

Control - No headache

Choose control well-matched with the case.
We have PM$_{2.5}$ average for both case and control.

Explore.

We focused on ‘headaches.’ What about working on other medical conditions? How to get the data?

**Connection to Biomedical Engineering**

I work with the Tissue Engineering Group.

Team

Dr. David Butler

Dr. Jason Shearn

Andrew Breidenbach, Ph.D. student

Andrea Lalley, Ph.D. student

Steve Gilday, Ph.D. student

I also work with the Biomechanics group.

Dr. Jason Shearn

Rebecca Nesbit, Ph.D. student

Nate Bates, Ph.D. student

The tissue engineering group is concerned with musculoskeletal injuries. They have information on the total number of patients who had injuries of this type for a year or two. Can we use the HCUP data to fine tune the extent of incidence of these injuries over the years? Cost? Length of Stay? Gender? Etc.

DRG = 477: Biopsies of Musculoskeletal System and Connective Tissue

DRG = 478

DRG = 479
**Other data sets**

**KIDS: Data from pediatric hospitals**

**SIDS: State-wide Inpatient Discharge data**

Emergency data

---

**Introduction to Data Science: Syllabus**

Division of Biostatistics and Epidemiology

Department of Environmental Health

College of Medicine

University of Cincinnati

**Syllabus**

**Title:** Introduction to Data Science

**Course:** BE 7082

Likely to be offered in Spring, 2015 after Curriculum Committee approval

**Introduction**

Traditionally, Statistics departments work within the environment of ‘small n – small p’ data, where n stands for the number of observations and p for the number of variables. A new discipline ‘Bioinformatics’ arose with the objective of handling ‘small n – large p’ data. Analyses of gene expression data, protein data, polymorphism data, etc. come under the purview of ‘Bioinformatics.’ The next
step is dealing with ‘large n’ data. This is what ‘Big data’ is made of. A fusion of Computer Science, Mathematics, and Statistics is needed to handle big data. A data scientist needs more than the fusion. He should be able to harness the following, in order of importance, to be a successful data scientist.

1. Statistics
2. Mathematics
3. Computer Science
4. Machine Learning
5. Domain Expertise (He/she needs to know the field from which the data comes from.)
6. Communication and Presentation skills
7. Data visualization

The purported class is intended to provide an introduction to big data. The students are trained to harness critical skills to become a successful data scientist. The following is an outline of the contents of the course.

Introduction

1. What is Data Science?
2. Examples

Computing skills

3. Introduction to R (ff and bigmemory packages)
4. Python and R
5. Hadoop and R
6. MapReduce, Pregel
7. Cloudera

Machine Learning Tools from Statistics

8. Cluster Analysis
9. Decision Trees and Random Forests
10. Bagging and Boosting
11. Regression
12. Logistic Regression
13. Pattern recognition
14. Naive Bayes
15. Bayesian Networks
16. Outlier Detection
17. Exploratory data analysis

Applications

18. Text mining
19. Social network analysis
20. Designing a spam filter
21. Forecasting in time series

Data visualization

22. Interactive graphs
23. Spatial graphs
24. Trend graphs

Evaluation

1. Homework – 10 homework sheets – 30 points
2. Project (Presentation is required.) 20 points
3. Mid-term exam - 25 points
4. Final exam - 25 points

Grades

A – 90 and above
B – 80 – 89
C – 70 – 79
D – 60 – 69
F – Below 60
Text book


References


