**Abstract**

The following study concerns Medicare healthcare insurance in the US by analyzing data from the year 2014. The goal of the study was to address the question of whether there is a difference in extent or “quality” of coverage across different regions and different types of providers across the US. The extent of coverage was measured in 2 ways: the amount a patient needs to pay from his pocket to settle a provider’s bill and the difference between the maximum Medicare allows for a given service and what it actually pays out. The regions used were the South, Northeast, Midwest and the west and the types of providers were segmented into males, females and organizations.

In the preliminary analysis by observing scatter plots it was seen that there was slightly worse coverage on the Northeast region however the differences as seen in these plots were not significant. On the other hand, it appeared that the quality of coverage was more poor amongst males and was much better with female or organization providers. While there was roughly a uniform distribution of records across the 4 regions, there was a much larger number of male providers than females, a cause for concern.

K-means clustering was then used to group the data into clusters and hence help me point out the reasons behind the differences between quality of coverage. A cluster size of 9 was seen to be the most efficient. Analysis of the clustering revealed that on average Medicare coverage was poorer in the Northeast region especially and lacking in the south and Midwest too. In these regions patients were required to pay more and there was a larger gap between Medicare allowed and actual payments. In the west coverage was much better with respect to both measures. Additionally, the quality of coverage was much lower amongst male providers and the best in organizations. Females also promoted good coverage.

In order to improve Medicare, a major concern for healthcare officials, the following steps were recommended. First male providers who charge high fees for typically low Medicare coverage services should be identified especially in the north east region. Medicare should work with such providers and encourage them to drive down fees or Medicare must increase their funding. Secondly, Medicare should try target more female providers or encourage the formation of more organizations as these typically have better coverage. The funding of the aforementioned programs could partially be generated by cutting down coverage in the west, a region that is currently benefiting very well from Medicare.

**Problem Statement**

Medicare is a national health insurance program that is intended to aid elderly patients in the US to cover their medical bills. Medicare however will neither cover total costs as charged by healthcare providers nor will cover all types of medical treatments. Medicare usually has a maximum amount of coverage allowed depending on the medical service and this coverage will vary depending on regions and types of providers. The rest of the charge is to be settled by the patient. In fact, in most cases, Medicare does not pay out its maximum allowance, forcing patients to pay more out of their pocket. This difference again has been seen to vary across regions and medical providers. Elderly patients are hence victims of this insufficient coverage by Medicare. **To better understand the extent or the “quality” of coverage I want to determine if there is a difference in extent of coverage across different regions or amongst different provider types and if so, where are the deficits mostly prevalent. The extent of coverage will be measured by the difference between the maximum Medicare is allowed to pay and what it actually pays as well as the amount patients need to pay to fully settle their bill.**

**Methodology**

Before the data can be analyzed, the data had to be cleaned and prepared. The first step was to reduce the data to make it more manageable. Using R 1 million rows were randomly selected which I believe gives a good representation of the overall data. Next several attribute columns were discarded such as Provider M.I, Street 1, Street2, Zip Code, Place of Service, HCPCS Code, description and Drug Indicator, line service count and benefactor service count because they were not required for the clustering analysis that would address my problem statement. Next, all rows with missing information were deleted as to prevent any biases along the way. This did not affect the reliability of the dataset.

Then, since the research question deals with coverage depending on the region, I decided to divide the state column into 4 key regions: N. EAST, SOUTH, WEST and MIDWEST. Lumping states into respective regions made it easier to understand the data and draw conclusions and is based on a valid assumption that states in the same region will have similar Medicare and healthcare policies and charges. These 4 regions were later converted into categorical variables in preparation for clustering. Provider gender was also similarly converted into categories.

In the final step of data preparation, 2 new attributes were created. The first one named PAY was created as the difference between the charge by a provider and the amount pad by Medicare. This is essentially the amount to be paid by a patient. The second one is called DIFF which is the difference between what Medicare pays and the maximum allowed amount by Medicare. Both these new attributes are measures of extent or quality of coverage.

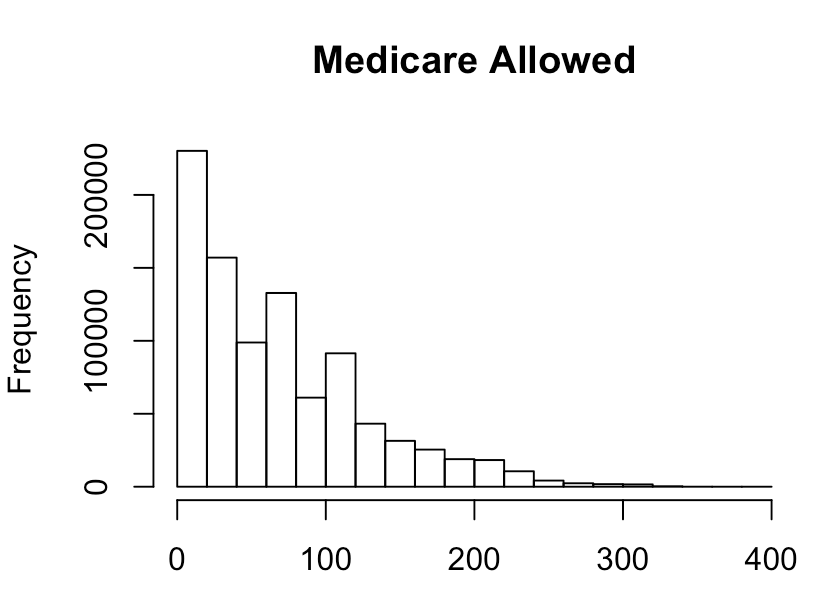
Before PAY and DIFF could be used in clustering they had to be log transformed first and then standardized. Log transformation was applied to make the distribution of the data better and prepare for standardizing. Taking the logs will prevent extremely large values from skewing and overshadowing the rest majority of the data in a given attribute. Standardizing pushes all the values between 0 and 1 hence scaling them appropriately for my analysis. Of course we are omitting the effect of larger data which may be very influential in cluster formation. However, we will assume such outliers are minimal and not important.

Finally categories were created for the state and gender attributes with binary variables indicating if a given record is or is not in a state or provided by a gender. This format is required for the following cluster analysis.

**Analysis**

**Summarizing the important data:**

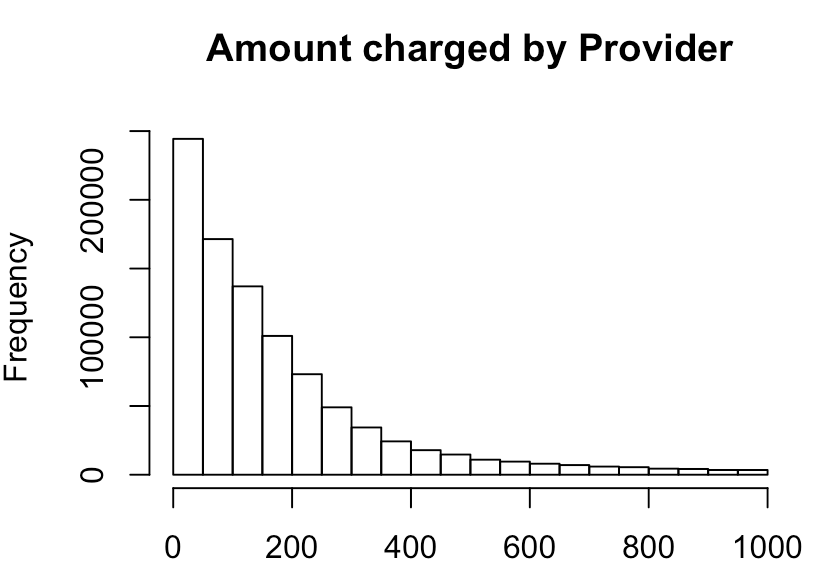
The first step of data analysis is based on summarizing various attributes of the data to better understand the important data and further prepare it for clustering. I start by summarizing and plotting the Medicare allowed, Medicare paid and the amount charged. Outliers were then removed to get a better spread. Plots of these 3 attributes corrected for outliers are depicted below.



Mean = $68

Min = $0

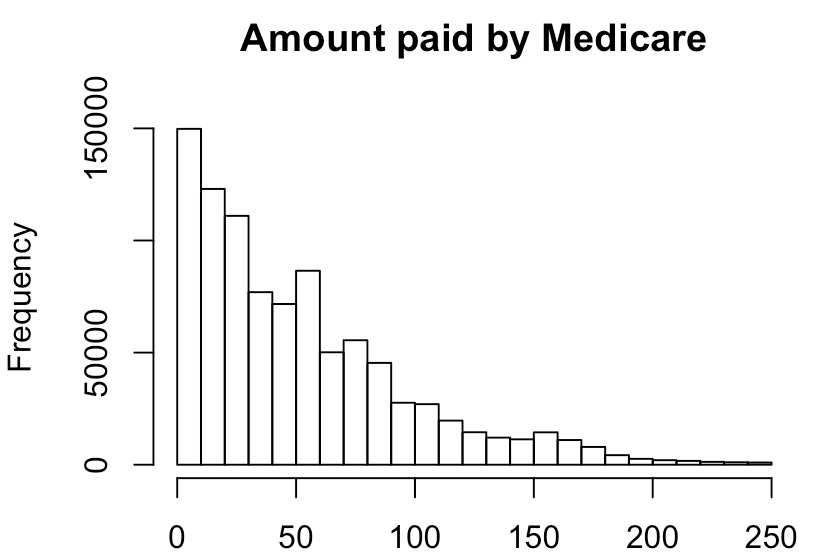
Max = $381



Mean = $175

Min = $0

Max = $1000



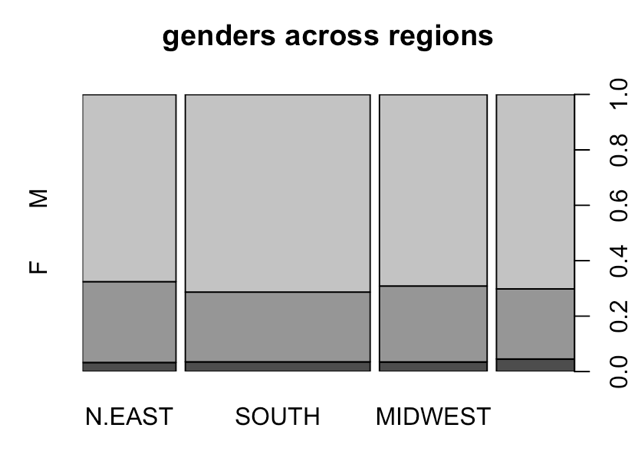
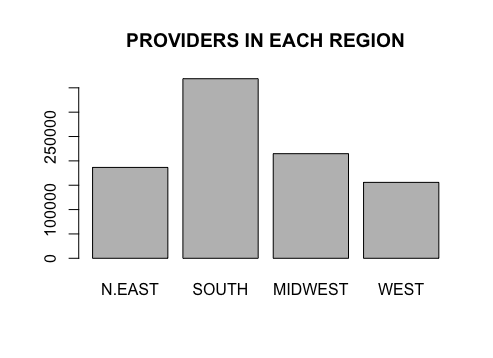
Mean = $52

Min = $0

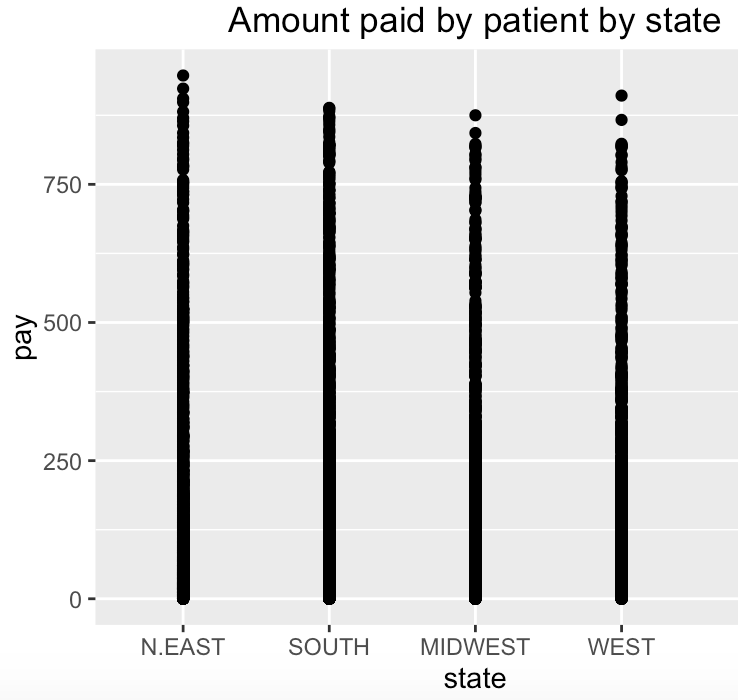
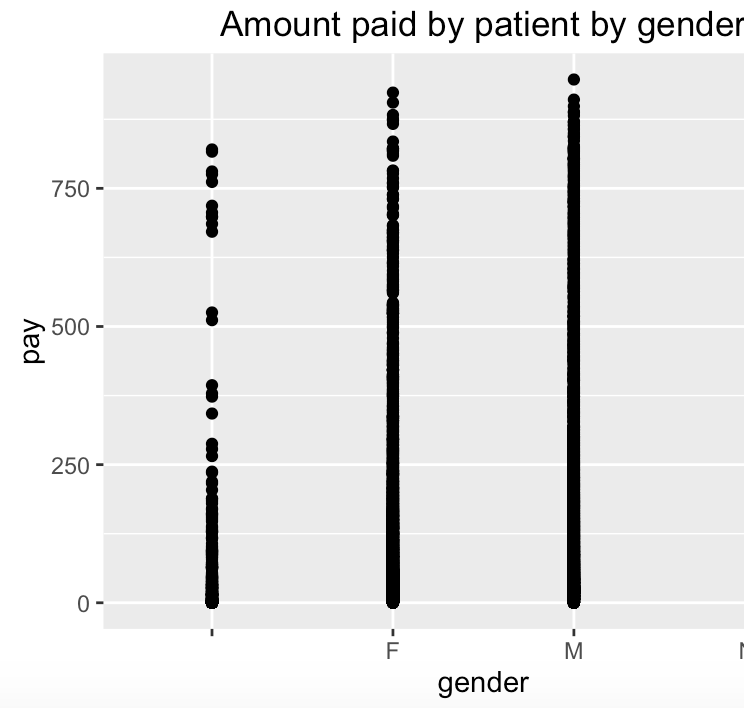
Max = $250

The plots above confirm the problem statement concerning quality of coverage of Medicare. There is an obvious difference between what providers charge and what Medicare pays as well as a discrepancy between Medicare’s maximum allowed and what it actually pays.

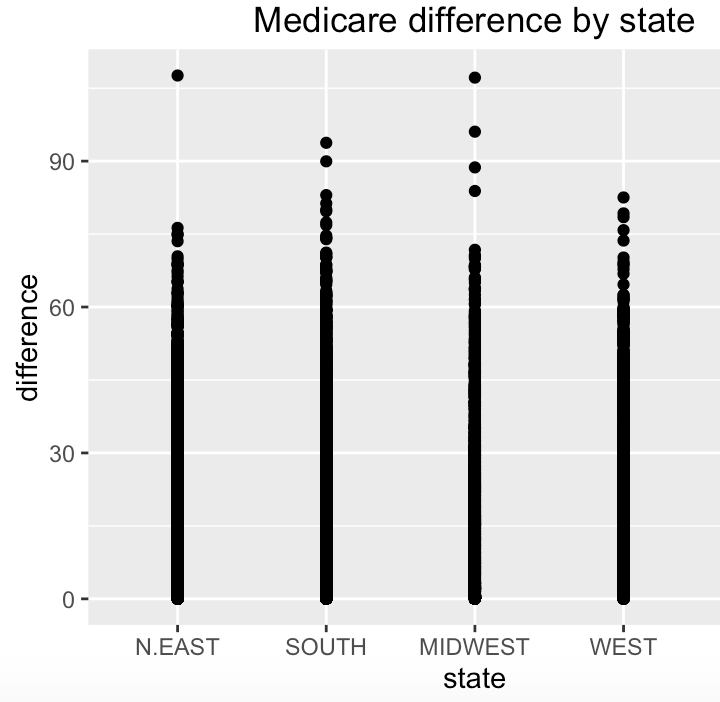
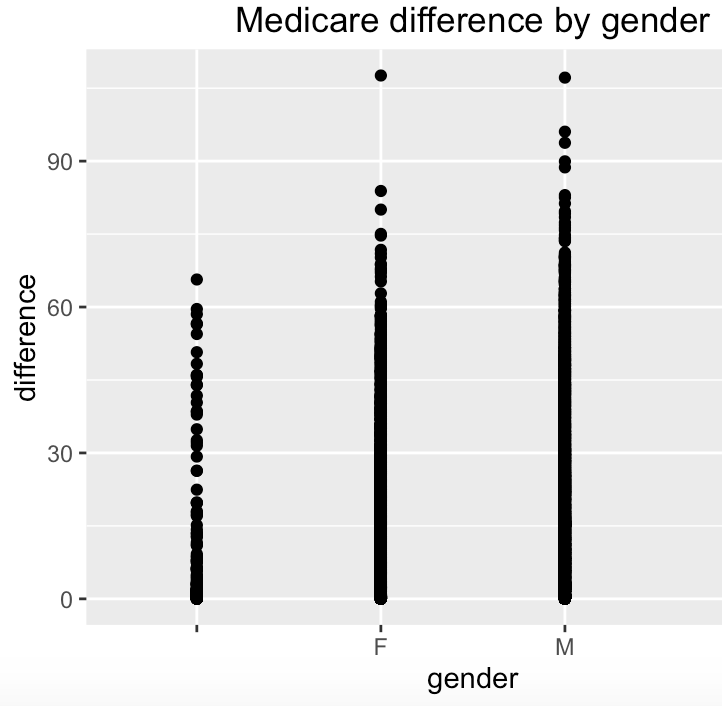
The plots below show that although there are the most number of Medicare records from the South, the number of records is distributed fairly uniformly. Additionally, the proportion of females and males is roughly equal for each region (females making up about 25% on average). It may be concerning that the overall number of female records are much lower than that of males and hence may impact clustering. However due to a large data set, these concerns can be overcome.



Now I look at the amount paid by patients and the difference between Medicare allowed and Medicare paid across different states and amongst different genders. The plots are displayed below.

****

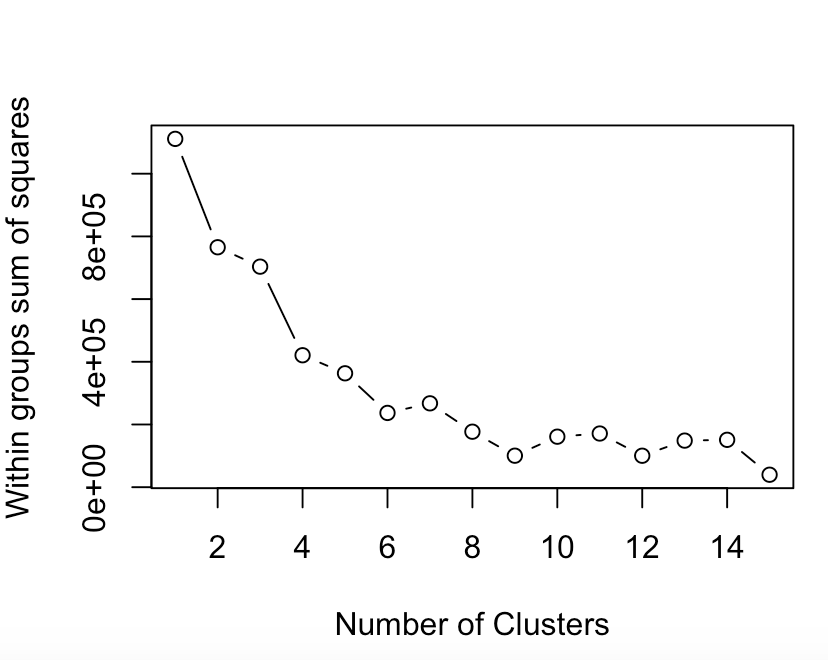
Looking at the two plots above it appears that in the North East region, patients may have to pay slightly more out of their pocket. However, the amount paid are quite similar as seen from this plot especially within the other 3 regions. Additionally, patients seem to pay slightly more when they consult male providers as compared to female providers. This is often as a result of males charging more than females. Organizations as shown by the first bar clearly require patients to pay the least from their pocket. This is perhaps because they charge the least. We will look out for such trends if more pronounced during clustering.

****

Looking at the difference between Medicare allocated and Medicare actually paid suggests that there is no significant variation by state. Expect for a few outliers, all states appear to have the same distribution of gaps. On the contrary, there is a more pronounced gap amongst genders. For males, Medicare appears to have the largest difference indicating poorest coverage quality. Females have a lower gap and organizations have the lowest. This suggests females and organizations promote better coverage by Medicare as measured by a smaller difference between Medicare allowed and Medicare paid. Again we will look out for such trends during clustering.

**Clustering**

K means algorithm was run and the following scree plot results. It is clear that 9 perhaps the ideal number of clusters as shown by the lowest within group sum of squares before the curve begins to level off.



The following clustering ratio is obtained:

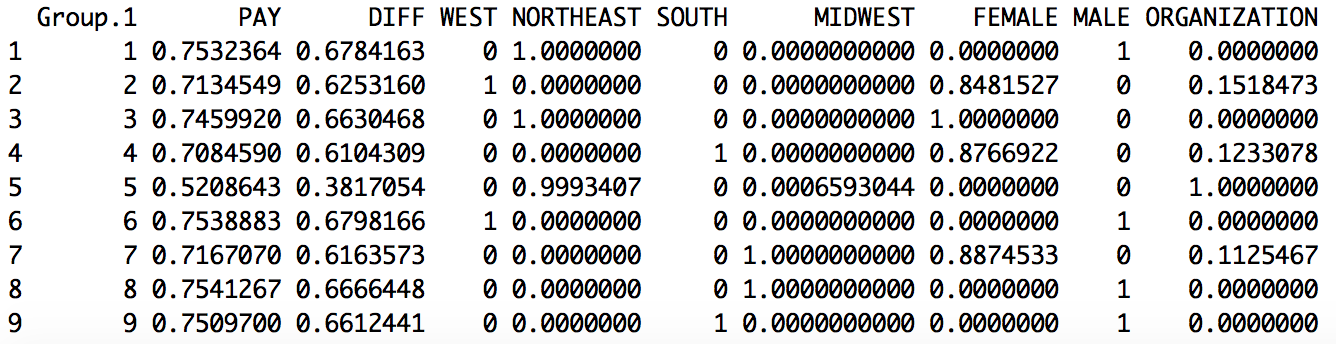
between\_SS / total\_SS = 91.4 %

Now the Silhouette score is determined for various cluster sizes around 9 to confirm that this is ideal:

|  |  |
| --- | --- |
| Cluster size | Silhouette score |
| 10 | 0.9972 |
| 9 | 0.9981 |
| 8 | 0.9968 |

This seems to confirm the fact that 9 is the ideal number of clusters.

The following clusters result with their given centroids for the respective attributes:



Looking at the amount paid by patients (PAY) it appears that the centroid is highest for group 1,3,6,8 and 9 with values about 0.75. In these clusters, patients are required to pay more from their pocket as compared to the rest. Group 5 has the lowest centroid of 0.52 and groups 2 and 4 have moderately high centroids around 0.7. Looking at clusters 1,6,8 and 9 we notice that all are covered by males. This in fact agrees with our previous analysis looking at the plots and confirms that patients tend to pay more out of their pocket when consulting male healthcare providers. Additionally, clusters 1 and 3, amongst the high paying clusters are located in the northeast region. Custer 8 and 9 are in the Midwest and south respectively. This also supports the initial hypothesis that the northeast region appears to require patients to pay the most. On the other hand, another important finding sis the organizations are responsible for requiring least money from patient’s pockets.

Now looking at the difference between Medicare allowed and that actually paid, the same clusters as before: 1,3,8,8 and 9 have the highest centroids of about 0.67. In these clusters there is the largest discrepancy between Medicare allowed and that paid. Again group 5 has the lowest centroid of about 0.38. This is similar to the results of PAY as described above. In cluster 5, Medicare pays an amount closest to its maximum allowances. These results combined reveal that cluster 5 appears to have the highest quality of Medicare coverage

**Concerns/ limitations**

Of course one would be concerned if the results above translates to poor coverage for 1,3,6,8 and 9 and great coverage for 3. This is because cluster 5 may typically have low allowances and hence a small difference as well as lower fees by doctors. Another concern to look at is the lack of data for organizations and females relative to males. This may cause us to not obtain a clear picture on the true differences between men and women.

**Recommendations**

Clustering reveals the characteristics of groups that appear to have high quality of Medicare coverage and those that have lower quality of coverage. The biggest concern was that male providers appear to charge the most resulting in patients paying the most out of their pocket. I would recommend that that healthcare officials negotiate with high charging males especially in the Northeast region to help reduce charges. Perhaps Medicare should also work more with male practices in the Northeast and Midwest and south too where they can increase their allowed coverage as well as what they actually pay. To fund this Medicare can reduce their funding in the West where patients seem to be covered pretty well.