– Independent Work Report Fall, 2017 –

Creation and Analysis of a Risk Adjustment Dataset

Grace Guan
Adviser: Mark Braverman

January 9, 2018

Abstract

This paper presents a first step into a Big Data analysis of the Department of Health and Human Services-Hierarchical Condition Categories (HHS-HCC) risk adjustment transfer payments ("risk adjustment") under the Affordable Care Act (ACA). Such an analysis may reveal areas to establish data-driven incentives that help insurers better predict pricing and become more competitive in the market, which reduces inefficiencies and lowers costs. The advent of Big Data in the form of high-dimensional government records allows for investigation of metrics previously not looked at, but these extensive records are disjoint and incomplete. We combine 9,265 "Insurer Report" Excel files with the "Summary Report" PDF files from the 2014 and 2015 Benefit Years from the Centers for Medicare and Medicaid Services' (CMS) website to create a new, easier-to-analyze risk adjustment dataset. Additionally, we perform preliminary data analysis in the individual marketplace and find that total loss from the 2014 benefit year has weak correlation ($r = -.11439$) with the difference in individual risk adjustment between the 2014 and 2015 benefit years. The negative correlation suggests that insurers may need multiple years to respond to market signals.

Contents

1. Introduction

1.1. Motivation and Goal

The goal of risk adjustment is to compensate health insurance plans for differences in the health of their enrollees.[4] Plans with healthier enrollees are supposed to transfer funds to plans that enroll less healthy individuals through "transfer payments."[7] The health mix of each plan is represented by the average of each patient's "risk score," which measures how costly each patient is expected to be relative to all other patients based on demographic factors and pre-existing conditions.[1] These transfer payments are meant to ensure that plan premiums do indeed reflect differences in plan factors, not differences in enrollees' health statuses.[7]

To truly understand how risk adjustment transfers play out in the wild, we originally aimed to backwards engineer the risk adjustment transfer payment formula. Transforming the health data of anonymous enrollees in a plan into the transfer payments of that same plan would allow for insights into how insurers could prevent paying out large transfers due to bad estimations or inappropriate reporting of their enrollee health mix. More significantly, we would also gain an intuition of weaknesses within the formula that may be rectified in the future to prevent larger losses. However, due to a lack of data, we shifted gears to creating and analyzing a dataset containing the

estimated and actual risk adjustment payments of every company that reported in the 2014 and 2015 benefit years.

On the whole, lack of functioning datasets appears to be an issue in the healthcare insurance analysis field. The CMS website provides the public with three datasets with data from the 2011-2015 benefit years. The first dataset consists of "Public Use" Excel files, which are supposed to contain all data from every company for each benefit year. The second dataset includes more than 20,000 Medical Loss Ratio "Insurer Report" Excel files, one for each insurer, which contain each insurer's reported estimated values of risk adjustment, among other useful things such as reported member months. The third dataset is comprised of "Summary Report" PDF files, which contain the actual risk adjustment transfer payouts from every company for each benefit year.

These datasets are obviously disjoint. Our initial aim is to turn these three datasets, which span multiple benefit years, into one CSV file per benefit year. Each CSV file would contain estimated and reported values for every insurer that reported data that year. Then, we can take a preliminary glance at the correlation between past years' losses and risk adjustment for the next year in the individual marketplace. We create this new dataset for the 2014 and 2015 benefit years because data from the 2016 benefit year is not yet available, and the 2011-2013 files do not contain risk adjustment transfer values. Our year to year analysis compares the 2014 and 2015 benefit years.

Creation of a complete dataset that contains risk adjustment transfer payment information contributes to the healthcare insurance field. Analyzing this newly created data may provide insights into how insurers can price plans more competitively to avoid large losses in the risk adjustment system. This paper focuses on the process for creation of this dataset and some preliminary statistical data analysis. The hope is that bringing to light these public government records may encourage insurance companies to take a more active approach in competitively pricing plans, which in turn can make the market more efficient.

1.2. Roadmap

The format of this paper is as follows. Section 2 provides an overview of the relevant topics at hand and prior literature. Then, section 3 discusses the general approach taken to solve this problem. Next, sections 4 and 5 discuss the implementation of this approach through the creation of the dataset and evaluate its effectiveness through preliminary data analysis. Section 6 concludes the paper and discusses further areas for research. Lastly, section 7 provides acknowledgments to my mentors.

2. Problem Background and Related Work

2.1. The Patient Protection and Affordable Care Act (2010)

The passage of the Patient Protection and Affordable Care Act (PPACA) and the Health Care and Education Reconciliation Act of 2010 (HCERA) drastically altered the health insurance plan pricing game in the United States.[1] New requirements on insurers prevented plans from turning away consumers with pre-existing conditions, expanded coverage to individuals who would have never been insured before, and allowed greater access to preventative services.[8] These changes are significant because prior to the ACA, insurers profited from enrolling more healthy people who cost less than they paid for insurance and turning away sicker individuals who cost more than they paid for insurance. Now that insurers could no longer turn away presumed higher-cost individuals based on their pre-existing conditions, there were fewer healthy people to balance out the risk pool. Additionally, the sicker individuals, who now found themselves insured, had deferred usage of insurance costs, making the insurers' transition to incorporating them in the system while not losing healthy consumers very risky and pricey.

There were many predicted side effects to this health insurance pricing game. For example, since insurers could no longer turn away sicker individuals, premiums would have to go up to account for the sicker peoples' higher costs. Further, if no incentives are given, no insurer would want to be the one taking care of all of the sick individuals due to their costly nature, and quality of care could

decrease. Lastly, no insurer would want to be the first to raise prices in fear of turning away healthy enrollees. These three examples, among many other possible changes, illustrate the difficulty in playing the new insurance pricing game.

Additionally, on the consumer side of the game, there has been evidence of adverse selection in ACA markets.[6] Adverse selection occurs in insurance when those seeking insurance are those likely to be higher-cost, and plan selection under the new ACA rules illustrated adverse selection. The ACA had an "individual mandate" that all people in America, minus a few exceptions, must be covered by insurance or pay a penalty fine otherwise.[1] These additional healthy people in the insurance system provide premiums that help balance the "risk pool" and pay for sicker patients. However, healthy people, doubting that they will use as much money as they pay into the system, are more likely to choose less expensive plans that do not cover as much, and sick people are more likely to choose more expensive plans that cover more. The changes in each plan's enrollee mix in turn change each plan's pricing. This has been found to reduce customer welfare on the whole.[2]

The serious changes in health insurance pools for both insurers and individual enrollees led to other mechanisms, specifically risk adjustment and reinsurance, being put in place to ensure the market functions properly.[1]

2.2. Risk Adjustment and Reinsurance

Risk adjustment is redistribution of money from plans with lower-cost enrollees to plans with higher-cost enrollees through "transfer payments."[7] These transfer payments are supposed to sum to zero for every state.[7] Risk adjustment is intended to offset variations in each plan's enrollee health mix due to risk selection by measuring the difference between plan premiums with and without risk selection.[7] This is to ensure that plan premiums do not reflect enrollees' health and pre-existing conditions, removing any incentives that insurers had to avoid sicker patients.

Similarly, reinsurance is payment from the government to plans with higher-cost enrollees to account for the premium increases that may come with enrolling sicker patients.[3] The temporary reinsurance program under the ACA was established from 2014-2016 to help stabilize premiums

in the individual health market. Since reinsurance protects companies against large losses if they enroll more high-cost enrollees than expected or if enrollees cost more than expected, reinsurance will act as a buffer so that the insurers will not have to raise premiums drastically.[5]

2.3. The Individual and Small Group Marketplaces

Payment transfers are separated between the individual, small group, and large group marketplaces. Health insurance bought by an individual is pooled in the individual marketplace. Group coverage is provided by an individual's employer or association.

First, we narrow down our focus to the individual and small group marketplaces since they have been reformed under the ACA. The large group marketplace was not affected by the ACA.[1] The individual and small group marketplaces have been modified to be "a single health insurance pool in each state, populated by all lawful residents in the state who do not have health benefits through a government program or a large employer, serviced by health insurance plans that provide all essential health care benefits and compete on the basis of cost and quality, with guaranteed access and identical premiums for all, subject to a few narrowly tailored exceptions that do not include health status."[1]

In this paper, we specifically focus on the individual marketplace since enrollees can choose their own plans and take full control over their health benefits. Enrollment by a small employer into the small group marketplace does not allow enrollees to choose the plan under which they feel the best covered. Thus, because group coverage does not fully reflect a competitive insurance-pricing game, we tailor our analyses to the individual marketplace.

2.4. Related Work

We believe that risk adjustment payment data has never been publicly cleaned nor analyzed before for multiple reasons. First, we were unable to find any other attempts in the literature to analyze or clean this data. Second, the current "Public Use" Excel files provided on the CMS websites are missing insurer data. Third, some of the current "Insurer Report" Excel files provided on the CMS

website are encrypted and we could not find any previous attempt to unencrypt these files. Lastly, no comprehensive dataset is available in the literature or on the internet to analyze.

As the government's public data is relatively inaccessible, there has been very little work done on analyzing ACA risk adjustment data. Therefore, since the current literature does not meet our goals, we would like to focus on analyzing the publicly available CMS data.

3. Approach

3.1. General Methodology

CMS has provided the public with three datasets: "Public Use" Excel files, "Insurer Report" Excel files, and "Summary Report" PDF files. Each dataset has a different format and contains different data. Our approach will be guided by two goals: first, converting our final output to CSV format, which is smaller in size and easier to parse, and second, combining as much relevant data as possible from all three sources for each insurer for each benefit year.

For our data analysis, our goal is to look at year-to-year correlation between loss and risk adjustment between 2014 and 2015. Our hypothesis is that larger losses in 2014 will lead to an insurer making active changes in how they price their plans, which leads to a positive risk adjustment transfer payment in the next year. On one hand, the insurer want to play it safe and enroll more healthy people, thus lowering the cost of their plans and also attracting sick people that they cannot turn away. Both healthy and sick people, attracted by the lower cost plan, cost more than they pay, but the sick people will increase the amount of risk transfer payments in the next year. On the other hand, if the insurer raises the cost of their plan, fewer healthy people will enroll and more sick people will enroll, which also increases risk transfer payments.

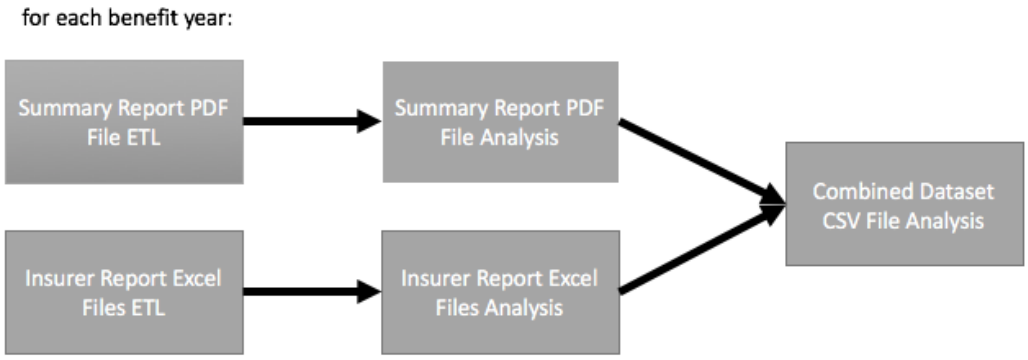

Figure 1: Flowchart of the data extract, transform, load process.

4. Implementation

4.1. Key Steps

Figure 1 illustrates flowchart of the data transformation process. We extract data from both the Summary Report and the Insurer Reports, and merge the data on the HIOS ID field to create the combined dataset CSV File. One CSV file is generated for each benefit year.

We used Python and the Python pandas library for all implementations. To make the Python code readable and accessible to the public, we used Jupyter Notebook. The full implementations and results can be found on GitHub.

4.2. "Public Use" Excel File Analysis

Prior to the creation of the dataset, we analyzed the "Public Use" Excel files, finding that they were unsuitable for use in our dataset. These "Public Use" Excel files were in the easiest format to extract data from due to their nature of being in an Excel file. However, the first problem was that these Excel files lacked the companies' names, listing the companies only by their Health Insurance Oversight System (HIOS) identification number. Disregarding the lack of proper company identification, we proceeded with our analysis of this data.

First, we extracted all data of the member months, reinsurance payment, and risk adjustment payment. Then, we dropped any companies which had null values for any of these data fields. We were left with 533 insurers, which is fewer than the 833 reported to have received transfer payments

in the "Summary Report" PDF File. Further, the risk transfer payments within this file did not sum to 0, as they should have. Thus, we determined that the data contained in this file was incomplete and inadequate for creating a dataset, and we ignored it in all future work.

4.3. "Summary Report" PDF File ETL

Our next challenge was parsing the tables within each "Summary Report" PDF file, which spanned multiple pages within each PDF file. The PDF file contained values for HIOS ID, company name, company state, reinsurance payment, and individual and small group risk adjustment payments. Each benefit year had two PDF files, one for estimated data that matched the "Insurer Report" Excel files as well as the "Public Use" Excel files, and one with the updated final values of the payments that ended up being made for each benefit year. We used the updated files because we wanted to see how they differed from the estimated files.

The approach we chose to extract data from this PDF file was first copying the data into a plain text file, then parsing the values within the plain text file and adding them into a CSV format dataframe. Since company names had line breaks in them, we parsed each raw text file by whitespace and determined what each field was through reading every word until finding a value that represented a payment. Of the words that we skipped, the last word represented the state, and previous words represented the company name. Algorithm 1 illustrates the implementation of this process.

4.4. "Summary Report" PDF File Analysis

To analyze the "Summary Report" PDF Files, we matched the updated PDF files that contained the final risk adjustment values to the "Public Use" Excel file analyzed in section 4.2 through a pandas dataframe merge. Because the "Public Use" file contained estimated numbers, and the "Summary Report" file contained final numbers, the values did not match exactly. We rectified this by matching the closest values of a hash together, where the hash was given by the following formula:

$$Hash = 7 * Reinsurance + 31 * RiskTransferIndividual + 17 * RiskTransferSmallGroup$$

The constant multipliers are randomly generated prime numbers of different magnitudes. Through the hash, we were able to match all 533 insurers described in the "Public Use" Excel files, but found

Algorithm 1 4.3 "Summary Report" PDF File ETL

```
procedure PARSEPDF
    read raw text from pdf into rawdata
    dataframe ← [HIOS ID, Company Name, State, Reinsurance, Risk Adjustment]
    split rawdata into array by whitespace
    index ← 0
    length ← length of rawdata
    while index < length - 1 do
        HIOS ID ← rawdata[index++]
        temp ← index
        while rawdata[index] != $ and rawdata[index] != "Not Eligible" do
            index++
        Company Name ← rawdata[temp:index - 1]
        State ← rawdata[index - 1]
        Reinsurance ← rawdata[index++]
        Risk Adjustment ← rawdata[index++]
        append [HIOS ID, Company, State, Reinsurance, Risk Adjustment] to dataframe
    save dataframe to csv
```

Algorithm 2 4.5 "Insurer Report" Excel File ETL

```
procedure PARSEINSURERS
    for file in all "Insurer Report" files do
        parse the reporting year
        parse all fields in sheet 0
        parse all fields in sheet 1
        parse all fields in sheet 2
        add all fields into dataframe for that reporting year
    drop all fields that were null for every insurer
    save all CSVs to csv
```

that there were 300 insurers that did not have data in the "Public Use" Excel file. Thus, we turned to the "Insurer Report" Excel files.

4.5. "Insurer Report" Excel Files ETL

There were over 20,000 "Insurer Report" Excel files for the 2011-2015 benefit years. Our first difficulty in parsing these files was differentiating between three different templates for these files, one for the 2011 and 2012 benefit years, one for the 2013 benefit year, and one for the 2014 and 2015 benefit years. We found that the templates for the 2011, 2012, and 2013 benefit years did not include adequate risk adjustment transfer data. Thus, we decided to only parse files from the 2014

and 2015 benefit years. We were able to filter files by their "Last Modified" date to find their benefit years, as files must have been last edited and submitted to CMS by July 31 of the following calendar year. We chose an arbitrary date of any file last modified after January 1, 2015, was considered to be from 2014 or 2015.

The next problem that we faced was file encryption on a minority of the files. While the files were not visibly encrypted to a human observer, we could not access the files programmatically through Python, R, or Java. These languages could not access data in the file because each sheet was protected to avoid editing errors. Possible solutions include manually decrypting the files, programmatically opening and closing the files, and converting each encrypted file to a csv to parse manually. We tried every solution, but we were unable to programmatically open and close each file due to memory constraints on the virtual Windows 32 client that we were running through Python. Thus, we manually decrypted the 470 files that were encrypted. For an XLSX file, decryption was enacted by opening and closing the file. For an XLS file, the steps for decryption were as follows. First, the files were converted to XLSX. Then, the files were individually zipped into archives. Next, 7-Zip was used to remove the "sheetProtection" XML tags within the zipped folders. Lastly, the files were unzipped and converted back to XLS format.

A third problem that we faced was accessing all of the data within the file. Each Excel file had 8 or 9 sheets within the file, only the first 3 of which were relevant to our data analysis. Since Excel sheets are 0-indexed, this corresponded to sheets 0, 1, and 2. Algorithm 2 illustrates the script we took to parse all of the data from the file, generating one file each for the 2014 and 2015 benefit years. The 2014 file had 5260 columns and 5059 rows. The 2015 file had 5260 columns and 4206 rows. We named each column in the new Excel spreadsheet by concatenating the row and the column names from the original 2014 Excel sheet. Each row represented a different Excel file. Some insurers spanned multiple Excel files, since they must report separately for every state they operate in.

Lastly, we deleted all columns that had null entries for all of the insurers who reported an Excel file, which halved the size of the CSV file. The entire parsing process from 9,265 Excel files to 2

CSV files took 10.5 hours.

4.6. “Insurer Report” Excel Files Analysis

We cannot make blanket statements about analyzing the contents of these files because different insurers filled out the files differently.

Each insurer has a "Grand_Total" File which is supposed to represent the sums of the values from each state the insurer operates in. Some of these Grand_Total Files added up correctly; others reported their grand totals as 0. Similarly, some insurers filled out obviously incorrect values. For example, some insurers even reported negative expected values of reinsurance, when reinsurance is a strictly positive value.

4.7. Analysis of the Dataset

We wanted to compare total loss per member month in the 2014 benefit year with the difference in risk adjustment per member month between the 2014 and 2015 benefit years. We calculated total loss by subtracting Total Incurred Claims, Prescription Drugs, Pharmaceutical Rebates, and State stop loss market stabilization and claim from the Total Direct Premium Earned for the individual market. We calculated the difference in risk adjustment per member month between the 2014 and 2015 benefit years by separately calculating the risk adjustment per member month in 2014 and 2015 and then subtracting the values. We deleted all companies that had a Null value for any of these entries. We also deleted all companies that had 0 values for all of their entries.

We then looked at the following:

1. The distribution of individual-group risk adjustment per member month in the 2014 and 2015 benefit years
2. The distribution of total loss per member month in the 2014 and 2015 benefit years
3. The distribution of the difference in individual-group risk adjustment per member month between the 2014 and 2015 benefit years
4. The correlation between total loss and the difference in individual-group risk adjustment per member month between the 2014 and 2015 benefit years

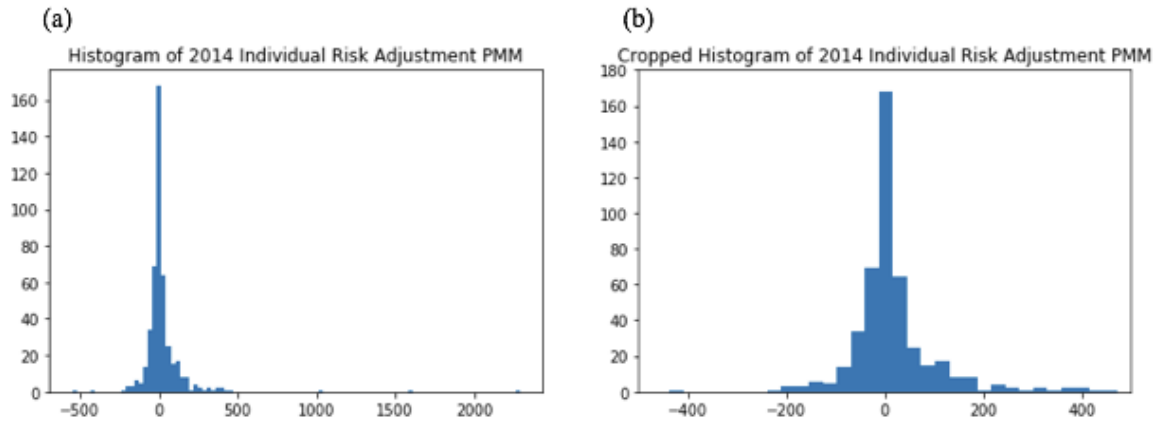

Figure 2: 2014 Risk Adjustment Per Member Month With and Without Outliers, 100 buckets

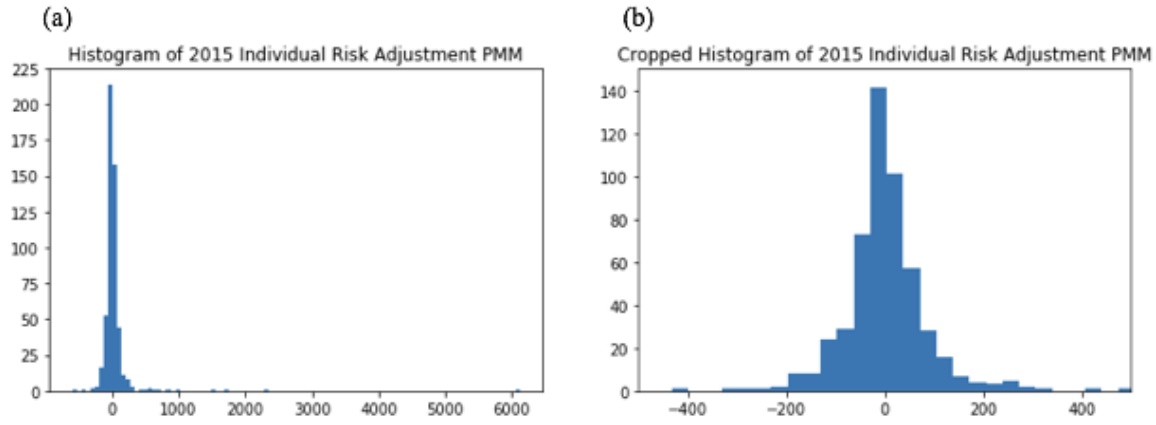

Figure 3: 2015 Risk Adjustment Per Member Month With and Without Outliers, 100 buckets

5. The correlation between individual-group risk adjustment per member month in the 2014 benefit year and individual-group risk adjustment per member month in the 2015 benefit year
6. Basic statistics (such as mean, median, standard deviation, etc.) and outliers

5. Results

5.1. Experiment Design

The following results were generated in similar fashion, using Python, the Python pandas library, and Jupyter Notebook. First, we extracted the relevant columns to each problem from the combined dataset from both the 2014 and 2015 benefit years. We stored these columns in two different dataframes. Then, we merged these two dataframes, one containing 2014 values and one containing

2015 values, on HIOS ID. Next, we dropped all rows that had null entries in any of the columns. We also dropped all rows that had reported entries of 0 in all of the columns manually. There were 993 insurers in 2014 and 911 insurers in 2015 that reported entries of 0 in all columns.

We used Matplotlib to generate plots and SciPy to generate correlation values.

5.2. The distribution of individual-group risk adjustment per member month in the 2014 and 2015 benefit years

Figures 2 and 3 illustrate a distribution of individual-group risk adjustment transfer payments per member month for all reporting insurers who did not report a 0 value. The (a) part of each figure depicts the distribution with outliers, and the (b) part of each figure depicts the distribution without outliers. Even after removing the insurers who had inputted all zeros in their forms, there was an obvious peak around 0 in both 2014 and 2015. After removing the outliers, the distributions for 2014 and 2015 are pretty similar, though 2014 has more transfers in the bucket around 0, which represents values of [-12.5, 12.5].

5.3. The distribution of total loss per member month in the 2014 and 2015 benefit years

Similar to the individual risk adjustment histograms, Figures 4 and 5 illustrate a distribution of total loss–calculated by subtracting all individual market claims from the total direct premium earned in the individual market–per member month for all reporting insurers who did not report a 0 value. The (a) part of each figure depicts the distribution with outliers, and the (b) part of each figure depicts the distribution without outliers. While 2015 had outliers of greater magnitude, after removal of the outliers, the distributions became similar. It makes sense that these distributions mirror the distributions of individual group risk adjustment. Plans that enroll healthier individuals who do not use as much cost as they pay into insurance will have a negative total loss but a positive risk transfer payment, and vice versa.

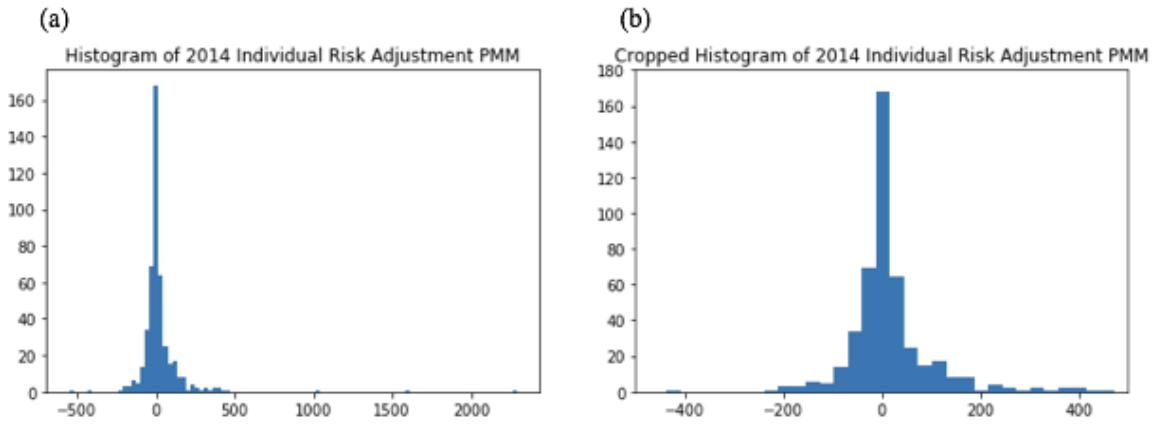

Figure 4: 2014 Total Loss Per Member Month With and Without Outliers, 100 buckets

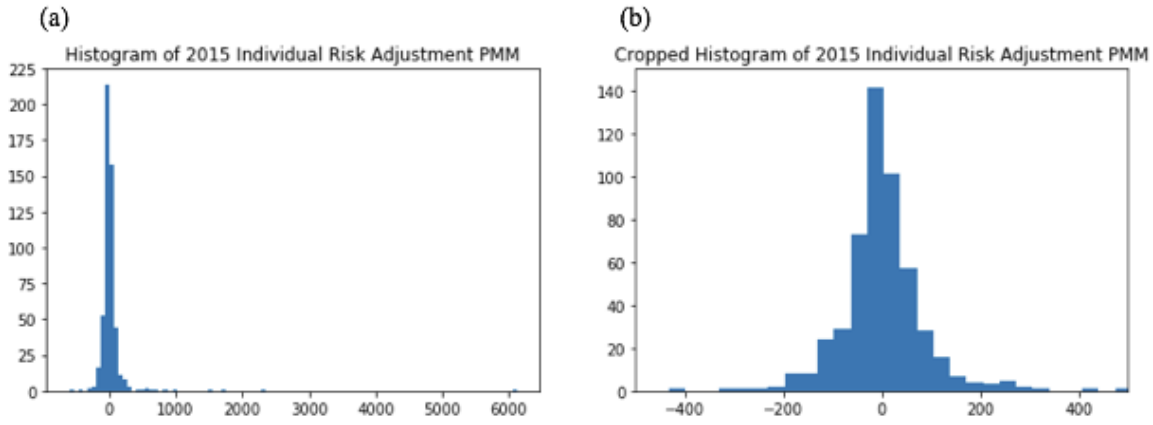

Figure 5: 2015 Total Loss Per Member Month With and Without Outliers, 100 buckets

5.4. The distribution of the difference in individual-group risk adjustment per member month between the 2014 and 2015 benefit years

Figure 6 depicts the distribution of the difference in individual-group risk adjustment per member month between the 2014 and 2015 benefit years. There is a higher peak around zero than each of the total loss or risk adjustment figures, indicating that most companies' risk adjustments did not change much between the two years.

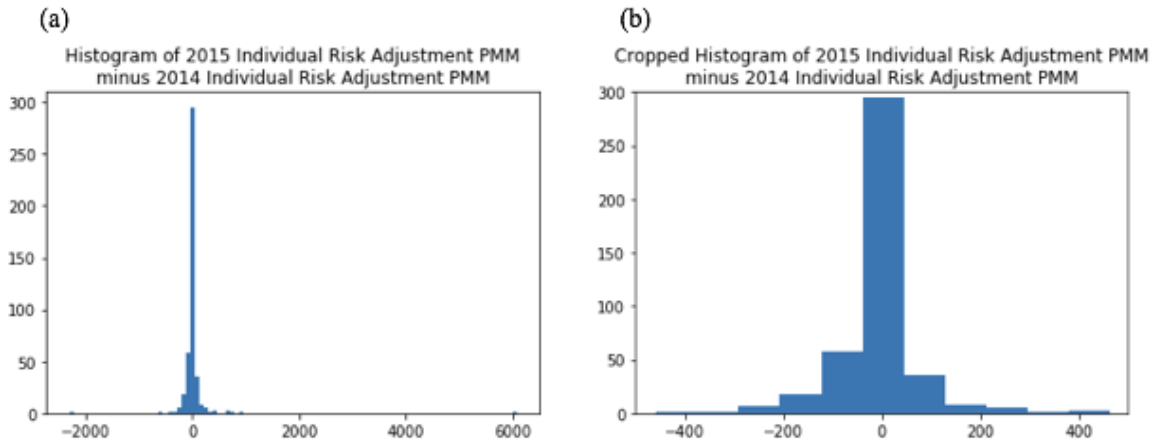

Figure 6: Difference between 2014 and 2015 Risk Adjustment Payments Per Member Month With and Without Outliers, 100 buckets

5.5. The correlation between total loss and the difference in individual-group risk adjustment per member month between the 2014 and 2015 benefit years

Figure 7 is a scatter plot depicting individual risk adjustment per member month compared to total loss. After removing all $(0,0)$ points from this dataset, we calculated the correlation coefficient $r = -0.11439$, which represents a very weak linear correlation. This is the opposite of what we expected. We hypothesized that the more an insurer lost in 2014, the more that they would attempt to be active in the marketplace and set premiums that increased the amount of risk adjustment that they received. Thus, the difference between risk adjustment in 2015 and risk adjustment in 2014 would be positive, and we would expect a moderate positive correlation. Because this is not the case, either incentives may take multiple years to show up in marketplace data, or insurers are currently still ineffective in setting prices, or insurers did not take a much more active approach in setting prices in 2015 compared to 2014, even if they lost money.

5.6. The correlation between individual-group risk adjustment per member month in the 2014 benefit year and individual-group risk adjustment per member month in the 2015 benefit year

Figure 8 is a scatter plot depicting risk adjustment in 2014 and 2015. Visually, we can see that the rough shape matches an oval stretched diagonally along the $y = x$ line, which represents a

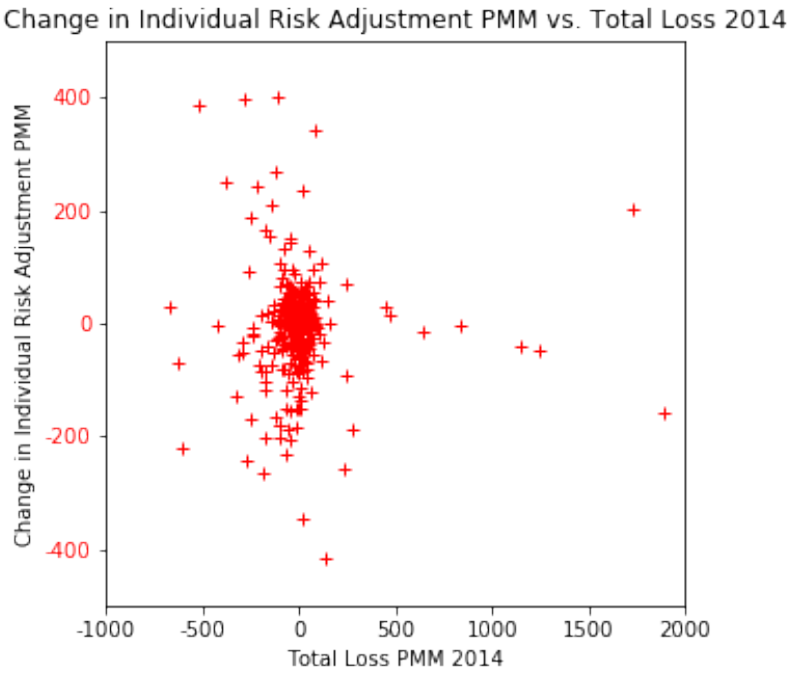

Figure 7: Scatter Plot of Change in Individual Risk Adjustment Per Member Month Between 2014 and 2015 vs. Total Loss in 2014

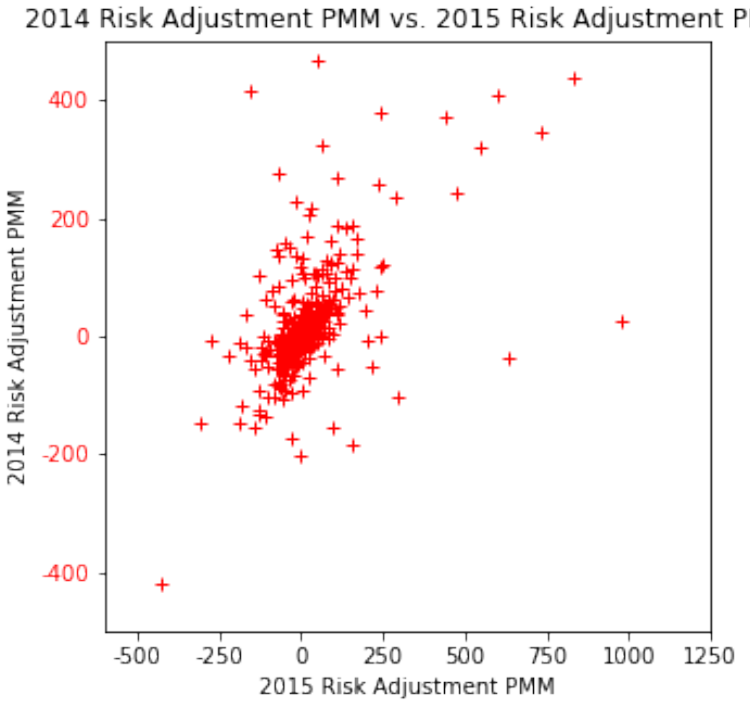

Figure 8: Scatter Plot of 2014 Individual Risk Adjustment Per Member Month vs. 2015 Individual Risk Adjustment Per Member Month

moderate positive correlation. Indeed, after dropping all $(0,0)$ values, the correlation coefficient is $r = 0.31868$. The fact that many of these values are close to 0 may have weakened the correlation. This graph illustrates that losses or gains in one year are still a pretty strong predictor for whether similar scale losses are gains should be predicted in the next year.

5.7. Basic statistics (such as mean, median, standard deviation, etc.) and outliers

Table 1 in the Appendix prints the statistics from the individual market data that we have analyzed in this paper. It is interesting to see that the means of the expected, actual, and difference in risk adjustment per member month have all increased between 2014 and 2015. Also, the actual risk adjustment values do not sum to zero because we dropped any insurers that did not have data for both expected and actual. This is because their lack of expected data means that they did not file a insurer report, so we would not have information about their member months.

6. Conclusion

6.1. Summary

Risk adjustment was an important and interesting addition in the ACA. Not only complicating the game that insurers play with respect to pricing their plans every year, risk adjustment also created new issues in the area of adverse selection. The analysis of risk adjustment payment data is important as a first step in finding limitations within risk adjustment in hopes of making the market more efficient. We present a new, clean dataset containing estimated and actual risk adjustment data from the 2014 and 2015 benefit years.

In analyzing the potential CMS sources for data, we found that the "Public Use" Excel files were missing data from 300 insurers, and thus was inadequate for our needs. Further, the "Summary Report" PDF files contained the final risk adjustment values for every company who received non-zero transfer payments, but did not include other valuable information such as non-profit status and member months insured. Lastly, the "Insurer Report" Excel files contained all the reported data necessary, but lacked the final transfer values, were disjoint, and contained inconsistencies due to

the fact that different companies were filling them out. In the end, we combined the 2014 and 2015 "Summary Report" files with the 9,265 "Insurer Report" Excel files, and analyzed distributions of the result.

Overall, we found that there is a negative correlation between total loss in 2014 and the change in risk adjustment between 2014 and 2015. This is significant because it illustrates that insurers may not be able to react to these market signals quickly enough to make an impact in their losses for the next year. The weak correlation between risk adjustment transfer payments in 2014 and 2015 further supports this argument.

6.2. Limitations

The primary limitation to this work was that not all insurers did report their risk adjustment data to the CMS. It is reasonable that some small insurers could not be bothered to do so. At the same time, all large insurers had accurate data accounted for. However, in this paper, all of the data from the PDF file was adequately matched to risk adjustment data from insurer files. Additional small insurers probably would not have affected the magnitude of correlation or the risk adjustment distributions too drastically, but smaller insurers would probably benefit the most by taking a more active role in the plan pricing game.

6.3. Future Work

In the future, I would like to develop this dataset further and work on more sophisticated analyses. Specifically, I would like to see if there is a significant statistical distinction between for-profit and non-profit insurers with regard to risk adjustment and reinsurance values. Additionally, once data from the 2016 benefit year comes out, I would like to repeat these analyses for the difference between the 2016 and 2015 benefit years. Lastly, at some point, it would be helpful to implement the full Risk Transfer Formula implementation in Python in hopes of back-calculating the values given in these Excel spreadsheets.

7. Acknowledgments

I would like to acknowledge my adviser, Professor Mark Braverman, for suggesting this interesting project in the space of applying computer science to healthcare and medicine. Professor Braverman was incredibly helpful and available, and he always had many insights into where to proceed when I seemed to be stuck, which I very much appreciate. I could not have asked for a better mentor to do my first independent work project with.

I would additionally like to thank Jérémie Lumbroso for his assistance in scraping the data from the CMS websites and suggesting ways and means to access and to handle such a large amount of data.

Lastly, I would like to thank the Princeton Department of Computer Science through the Applications of Computing certificate program for granting me a space within the CS system to host files for this project, and my home Department of Operations Research and Financial Engineering for allowing me to pursue this research project as a sophomore.

References

[1] T. Baker, "Health insurance, risk, and responsibility after the patient protection and affordable care act," *University of Pennsylvania Law Review*, vol. 159, no. 6, 2011.

[2] B. Handel, "Adverse selection and switching costs in health insurance markets: When nudging hurts," *National Bureau of Economic Research*, 2011.

[3] T. S. Jost, "Health insurance exchanges and the affordable care act: Key policy issues," *Washington and Lee University School of Law*, 2010.

[4] J. Kautter, G. C. Pope, M. Ingber, S. Freeman, L. Patterson, M. Cohen, and P. Keenan, "The hhs-hcc risk adjustment model for individual and small group markets under the affordable care act," *Medicare and Medicaid Research Review*, vol. 4, no. 3, 2014.

[5] J. Kautter, G. C. Pope, and P. Keenan, "Affordable care act risk adjustment: Overview, context, and challenges," *Medicare and Medicaid Research Review*, vol. 4, no. 3, 2014.

[6] M. Panhans, "Adverse selection in aca exchange markets: Evidence from colorado," *SSRN*, 2017.

[7] G. C. Pope, H. Bachofer, A. Pearlman, J. Kautter, E. Hunter, D. Miller, and P. Keenan, "Risk transfer formula for individual and small group markets under the affordable care act," *Medicare and Medicaid Research Review*, vol. 4, no. 3, 2014.

[8] S. Rak and J. Coffin, "Affordable care act," *Journal of Medical Practice Management*, vol. 28, no. 5, 2013.

Table 1: Basic Statistics from the Individual Market Data

Column	Mean 2014	Mean 2015	Median 2014	Median 2015
Member Months	129068.3535	197319.3718	49	1650
Expected Risk Adj	-30588.87645	-55273.84991	0	0
Actual Risk Adj	-11429.01659	142675.3889	71612.47	1509.65
Expected Risk Adj PMM	14.6653789	21.88552542	0	0.07
Actual Risk Adj PMM	25.32608225	30.41954373	2.275	0.095
Difference in Risk Adj PMM	0.990372807	3.155304183	1	0
Total Loss	74.54520059	358.8028475	-3.26	-53.695
	Std 2014	**Std 2015**	**Min 2014**	**Min 2015**
Member Months	571308.945	738379.7924	0	0
Expected Risk Adj	13312771	20934625.65	-181649910	-218903904.1
Actual Risk Adj	21906792.98	27963781.79	-181692588	-218903904.1
Expected Risk Adj PMM	137.478212	323.5524114	-551.79	-1557.35
Actual Risk Adj PMM	163.8542688	323.6415277	-551.79	-600.91
Difference in Risk Adj PMM	0.128728666	33.41751368	0	-127.21
Total Loss	1580.206016	12760.95586	-4553.56	-30712.35
	Max 2014	**Max 2015**	**Sum 2014**	**Sum 2015**
Member Months	9694097	11485974	163916809	198503288
Expected Risk Adj	195610300.2	368933331	-35911340.95	-51957418.92
Actual Risk Adj	221628751.9	368933330.5	-5234489.6	74333877.64
Expected Risk Adj PMM	2292.67	6120.33	9869.8	12912.46
Actual Risk Adj PMM	2292.68	6120.33	11700.65	16000.68
Difference in Risk Adj PMM	1.57	641.58	451.61	1659.69
Total Loss	27626.23	308048.38	50168.92	211693.68