Healthcare Bundled Payments

By

Hardik Gupta (71620027)

Richa Saxena (71620058)

Shweta Pisharoti (71620065)

A Report submitted to the

Indian School of Business, Hyderabad

For Certificate Programme in Business Analytics

Under the guidance of

Prof. Subodha Kumar

Project Sponsored by

Deloitte

2017-2018

ISB

Contents

ISB

Executive Summary

In the US, there is an increasing need for devising a payment model for the healthcare sector that rewards providers for delivering quality care to the patients at a lower cost. The current fee-for-service model reimburses providers like hospitals, physicians and nursing facilities individually, for every service provided during the episode of care. This results in high volume, high cost, and inadequate outcome as the providers are paid separately for each service.

Episode based Bundled payment or 'value-based reimbursement' (Michael E. Porter, 2016) is a disbursement model which makes a single payment to different providers involved in the treatment of the patient over a period of time. Bundled payment is an attempt to offer incentives to care providers by making a single payment, for all services provided during an episode of care. However, an effective implementation of this system involves deep knowledge of the episode of care in terms of diagnoses, procedures and cost borne by patients and insurance companies like Centre for Medicare and Medicaid Services (CMS).

Analytical techniques, some of which showcased in this project, are very essential to analyse large patient claim records and find insights for probable bundling. Use of data mining techniques such as Association Rule mining, Sequential Rule Mining and Process Mining can help the medical facility understand the frequent patterns within an episode of care. Patterns include co-occurring diagnoses, sequence of procedures conducted over a period of time and end-to-end visual understanding of patient's medical history, which are all necessary components to form a bundle. In addition, the use of statistical techniques like Logistic Regression is useful to predict the future outcome of a patient belonging to an episode of care. Predicting readmission based on a particular procedure, for example, could help the patient receive better medical facilities (included in the bundle) during early stages of care.

The objective of this project is to devise a framework for bundling medical procedures and diagnoses, based on frequent patterns occurring in the CMS medical claims data. Medical claims data is the collection of records which consists of information regarding the diagnoses, procedures, providers and length of stay across inpatient and outpatient facilities. The project demonstrates bundled payment options for one of the frequently occurring diagnosis – Ischemic Heart Disease- among inpatient records. The project also highlights an analytical way of selecting different providers for the same episode of care based on their cost of treatment, which could help Medicare choose the right set of providers to render these bundled services.

The project discusses several logical steps for probable bundling, however the framework devised is not limited to this. The dataset used is a public synthetic file replicating claims data and hence the insights gained in this report cannot directly be used by the stakeholders. However, all of the concepts and techniques applied in the project could be applied on real claims data (with similar data structure and attributes) to derive actual insights.

Keywords- Healthcare, Bundled Payments, Data Mining

Motivation

- For the Project sponsor: To provide a process on how to tackle the bundling of payments and what approaches could be taken to arrive at an ideal bundle with the right price. An improved version could be a valuable proof of concept for prospective clients.

- For Medicare: To attempt identifying bundles to improve quality of care given by the providers. To also provide the approach for pricing each bundle. Once the process is applied to real data, Medicare could find valuable insights to reduce expenditure.
- For the Team: To apply the analytical concepts and techniques learnt over the course of CBA. To understand the end-to-end process beginning from requirements gathering, data collection and cleaning till deriving insights and models.

Business Problem

The effective implementation of bundled payment system brings in its own set of complications and challenges. This project attempts to provide a framework to address some of these key business problems, using an analytical approach.

1. **Defining the bundle**

   The first step to bundle payment is to define what constitutes the bundle. The key business questions are, what episodes are to be considered for the bundle? How can one define a bundle for a given episode? Should certain patient types be excluded for a bundle? What is the evaluation criteria of a bundle efficiency (the cost of care is minimized and quality of care is maximized)?

2. **Pricing the Bundle**

   Once a bundle is defined, it is important then to price the bundle correctly. The key business problem here is how should the bundle be priced? Is there a minimum volume factor that impacts the price? What factors should be considered for risk adjustment? How should the pricing outlier policy be determined (too low and too high)?

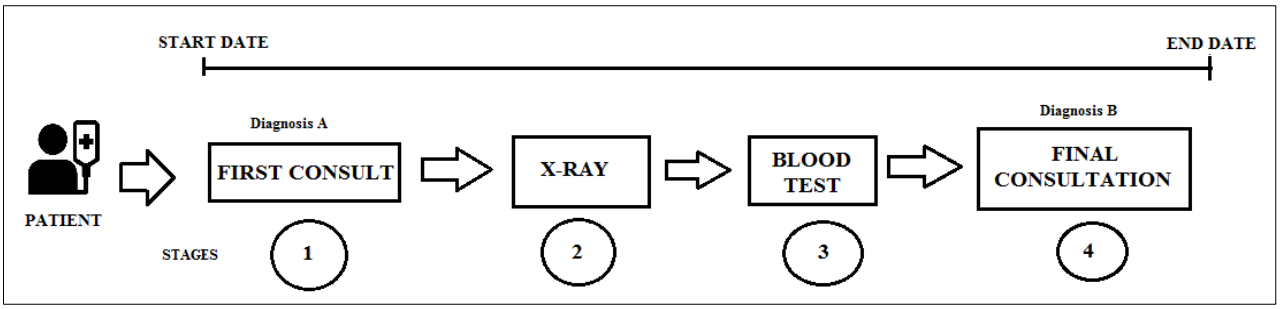
Introduction

Bundled payments could significantly reduce costs for Centre for Medicare and Medicaid Services (CMS), specifically Medicare. Bundled approach aims at paying providers a single payment for all care delivered across the entire episode thereby encouraging providers to deliver better value, improved co-ordination and quality through a patient-centric approach.

The stakeholder of this project is mainly CMS. However, bundled payments is a win-win for everybody be it CMS, the hospital, the outpatient facilities or the patients. The payment model seeks to improve,

1. For Patients, they will avail services from the providers who deliver the best outcomes for their medical condition.
2. For Providers, they will change the way how care is provided currently and focus more on quality and reduced costs.
3. For Pharmaceutical companies, they will have to compete based on their value proposition and price their products carefully.

However, knowing what to bundle requires extensive research of different types of patients, episodes, the diagnoses and procedures performed. A clinical pathway of the patient is described below, where at each stage different diagnoses and procedures are performed.



The quality of care provided by the hospitals or independent physicians depends on effective treatment, identifying correct diagnosis and performing correct procedures at each stage. Healthcare processes involve different types of resources such as physicians, skilled nurses, ambulance facilities etc. and can vary from one provider to another. Overall, healthcare processes are highly variable, complicated and dynamic which ultimately makes bundling these processes complicated. There is a need to understand these pathways of the patients (diagnoses and procedures) to reduce the cost and improve the quality of service.

Data Mining is the analytical technique used to decipher patterns from large, raw datasets and establish relations. The project uses hospital claims data and three data mining techniques, performed on the diagnoses and procedures of the patients to check for frequent patterns which can help in defining the bundles

1. Association Rule Mining – associations between different diagnoses for patients on every visit to the hospital. This is used to check if two or more diagnoses are performed at the same time.
2. Sequential Rule Mining – sequence of procedures performed on the patients. This is used to identify a frequently occurring sequence or subsequence which can be bundled for a group of patients.
3. Process Mining – To evaluate the complete trajectory of every patient, which can help Medicare to observe the sequence of diagnoses or procedures performed, identify common trails.

Data

Data files used for the project are publicly available **synthetic** datasets on the website - https://www.cms.gov/Research-Statistics-Data-and-Systems/Downloadable-Public-Use-Files/SynPUFs/DE_Syn_PUF.html. (CMS, n.d.)

The dataset contains multiple files per year for the years 2008 to 2010. For this project, Beneficiary Summary (for 2008, 2009 and 2010), Inpatient Claims and Outpatient Claims files are used for the analysis and bundling and are obtained from samples 1 to 5 from the above mentioned website. A brief summary of the different data files used is given below (ResDAC)

File	Description	Remark
Beneficiary Summary DE-SynPUF	Beneficiary-level file, contains information on	9,63,230 unique patients (combined from samples 1 to 5)
	1. Personal Information 2. Demographics	
	3. Chronic conditions	
	4. Total Reimbursement data across inpatient, outpatient and carrier files	
Inpatient Claims DE-SynPUF	Information on synthetic institutional claims for hospital inpatient services provided to beneficiaries. The file contains data majorly on	332284 unique records (combined from samples 1 to 5)
	1. Dates of Service 2. Hospital Provider Number	
	3. Claim Payment Amount 4. Physician NPI number	
	5. Medicare Claim Diagnosis Related Group 6. Utilization day count	
	7. Diagnosis codes (ICD-9 diagnosis) 8. Procedure codes (ICD-9 procedure)	
	9. HCPCS codes	
Outpatient Claims DE-SynPUF	Information on synthetic institutional claims for outpatient services provided to beneficiaries. The file contains data majorly on	3955790 (combined from samples 1 to 5)
	1. Dates of Service 2. Hospital Provider Number	
	3. Claim Payment Amount 4. Physician NPI number	
	5. Diagnosis codes (ICD-9 diagnosis) 6. Procedure codes (ICD-9 procedure)	
	7. HCPCS codes	

Note: The CMS Data Users Document mentions,

"*The files preserve the detailed data structure and metadata of key variables at both the beneficiary and claim levels. However, the data are fully "synthetic," meaning no beneficiary in the DE-SynPUF is an actual Medicare beneficiary. They are all synthetic beneficiaries meant to represent actual beneficiaries. In order to protect the privacy of beneficiaries and to greatly reduce the risk of re-identification, a significant amount of interdependence and co-variation among variables has been altered in the synthetic process. The synthetic process used significantly diminishes the analytic utility of the file to produce reliable inferences about the actual Medicare beneficiary population (i.e., univariate statistics and regression coefficients produced with the DE-SynPUF will be biased). Although the DE-SynPUF has limited empirical research utility, it does have the same data and file structure as the actual 5% Medicare beneficiary file and similar number of beneficiaries; it just has a smaller number of claims types and number of variables. Because the structure of the data is maintained, the DE-SynPUF is useful for building data tools that could be used with the actual data."* (CMS, n.d.)

It means that the insights from this project, based on the synthetic data files, cannot be used directly. However, the process and techniques used could be re-used on data having the same structure and parameters.

The complete description for each field can be obtained from this codebook - https://www.cms.gov/Research-Statistics-Data-and-Systems/Downloadable-Public-Use-Files/SynPUFs/Downloads/SynPUF_Codebook.pdf

ICD Diagnosis and Procedure Codes

ICD-9 (International Classification of Diseases, 9th edition) is a list of diagnoses and procedure codes for communicating hospital care in US. These codes are found in the claims data used for billing and reporting purpose. Diagnosis codes are mainly 4 or 5 digit code and have a hierarchy of major, sub-chapter and chapter. ICD-9 procedure codes are 4 digit codes

Package 'icd' in R

CRAN package 'icd' (Jack O. Wasey, 2017) is used for ICD-9 diagnosis codes. 'icd9cm_hierarchy' is a dataframe containing the full ICD-9 classification for each diagnosis. The dataframe provides information Code, short_desc, long_desc, three_digit, major, sub_chapter, chapter.

CPT Procedure Codes and HCPCS Procedure Codes

Procedures are also billed using CPT (Current Procedural Terminology) /HCPCS (Healthcare Common Procedure Coding System) codes, rather than ICD. CPT or Level 1 HCPCS codes are 5 digits numeric codes copyrighted by AMA (American Medical Association). Level 2 HCPCS codes are 5 digit alphanumeric codes.

Scope and Assumptions

1. Based on understanding with the project sponsor: The project is not a standalone application and no user specific inputs are required.
2. Data files used for the project are synthetic as described above. Any insights derived are purely for example to highlight how the approach to bundle can be used. The models and analysis should be applied on real claims data to understand the full worth of the project and make the results meaningful.

3. The data provided is assumed to be correct for the purpose of creating techniques to bundle and this project does not extend to checking the correctness of the data in the files. Due to the heavy imputation and synthetization, it has been an immense effort deriving the results in the Ischemic Bundle example.
4. Analysis and techniques are limited to Beneficiary, Inpatient and Outpatient files.
5. This project can be used as guideline for approach on newer datasets from CMS, the models and statistical tests, however, might not suit datasets where columns and parameters are different.

Challenges

1. Missing information regarding procedure codes. Both ICD-9 and HCPCS codes were missing to a great extent. Also, some of available ICD-9 procedure code were actually found to be ICD-9 diagnosis codes.
2. Missing information regarding the cost for each procedure conducted. Since there are multiple procedures conducted during every visit and claims data consists of one final claim amount, it was challenging to gauge individual procedure cost.
3. Since this is synthetic file, the project only attempts to provide a framework and not an exact solution.

Tools

1. **R Statistical Software -** R is majorly used for data cleaning, pre-processing, data crunching, model building and data visualization. Some of the major packages used for the implementation of this project are
    - data.table, dplyr –Extremely powerful and efficient for analysing large dataset
    - ggplot2 – Used for data visualization
    - bupaR - Business Process Analysis in R, for observing traces within the patients
    - arules – Implementing Association Mining in R
    - arulesSequences – Implementing Sequential pattern and rule mining in R

2. **Tableau-** Tableau 10 is used for analysing datasets and creating constructive, detailed visuals for insights

Analytical Methods

1. Association Mining

Association rule mining is an analytical technique which finds frequent patterns, correlations and associations in a given dataset. Given as set of transactions, association mining finds rules to predict the occurrence of items in a dataset depending on the occurrence of other items.

In healthcare, this can be useful to find co-occurrence of diagnoses or procedures performed at each visit to the facility.

For example in a rule {X} → {Y},

Support is how many times X and Y occur in the transactions; $s = \frac{\text{No. of records with X and Y}}{\text{Total no. of transactions}}$

Confidence is the no. of times Y appears in transactions that contain X; $c = \frac{\text{No. of records with X and Y}}{\text{No. of records with X}}$

Lift is the ratio of the observed support to that if X and Y were independent; $\text{lift}(X \rightarrow Y) = \frac{s(X \cup Y)}{s(X) * s(Y)}$

The package used in R is 'arules' which contains mining algorithms like Apriori. While interpreting association mining results, high confidence, high support and lift >1 makes a good rule. Lift should be greater than 1, implying a proportional impact. Lift = 1 implies X and Y are independent and not associated.

2. Sequential Rule Mining

Sequential Pattern Mining is an analytical technique used to discover patterns in items over a sequence of time. More precisely, it consists of discovering interesting sub-sequences in a set of sequences.

In healthcare, sequential rule mining can be helpful to identify the sequence or flow of diagnoses or procedures performed over a period of time (i.e. over multiple visits). The results of this could be used to determine add-on packages for bundles or combination of bundles.

Sequential models are built using package 'arulesSequences' in R. read_baskets() method is used to convert dataset into transactional data and convert into "sequenceID","eventID","SIZE" format. The cSPADE (Sequential PAttern Discovery using Equivalence classes) is an algorithm in this package. Method 'ruleInduction()' takes the rules calculated by cSPADE algorithm with certain support and then based on desired confidence evaluates the rules for their independence. Higher the confidence of the rules, the better is the probability that one item of the sequence is followed by another item, and the occurrence of both is not independent (i.e. they depend on each other).

3. Process Mining

Process Mining is an analytical technique used to extract information from event logs, in order to analyze the underlying processes. An event log is the set of activities or events (defined in terms of diagnosis or procedures) executed at different times. In healthcare, process mining can be used to identify the complete end-to-end medical path of a patient. The event log, here, is the set of diagnoses or procedures performed on every visit to the hospital or the doctor.

'bupaR' is the package in R, which accepts 'eventlog' object as its input which consists of:

- Case ID - e.g. Unique Patient ID)
- Activity ID – e.g. Different activities in terms of procedure codes or diagnosis codes (one procedure or one diagnosis code or one description which summarises the activity)
- Activity Instance ID - Unique ID to each activity, per case ID. There can be many activity instance ID per case ID
- Lifecycle ID – e.g. status of each activity instance (complete, in progress), or Type of record like inpatient or outpatient
- Timestamp – e.g. Date of hospital visit
- Resource ID – Additional information for activity, e.g. the taxonomy of physician who attended the patient

4. Logistic Regression

Logistic Regression is a statistical method which enables analysing a dataset where one or more variables are independent. The variable that is being regressed (the resultant column) is binary with only two possible outcomes and dependent on the other variables. In Healthcare, the resultant variable might be finding if patient is getting readmitted or not. (Refer Appendix for more on Performance of logistic regression)

Data Cleaning and Pre-processing

Beneficiary files for 2008, 2009 and 2010, inpatient files and outpatient files from samples 1 to 5 are read in R environment. Data cleaning and pre-processing step involves changing the class format of few columns, creating new columns and sub-setting data frames with only necessary columns.

For each Beneficiary files (2008, 2009 and 2010)

- Changing class of column 'BENE_BIRTH_DT' to Date
- Adding new column AGE (whole number)
- Adding new column BENE_IS_DEAD (factor, YES or NO)
- Changing class of column BENE_SEX_IDENT_CD (factor, M or F)
- Changing class of column BENE_ESRD_IND (factor, N or Y)
- Changing class of 11 chronic columns starting with SP_ (factor, N or Y)
- Adding new column 'CountOfDiseases', row sum of 11 chronic columns where value = Y

For Inpatient files

- Changing class of column 'CLM_ADMSN_DT' to Date
- Changing class of column 'NCH_BENE_DSCHRG_DT' to Date
- Adding new column 'Days_Before_Readmission' which is difference of days between NCH_BENE_DSCHRG_DT of row 1 and CLM_ADMSN_DT of row 2. Hence Days_Before_Readmission = CLM_ADMSN_DT (row 2) - NCH_BENE_DSCHRG_DT (row 1)

For Outpatient files

- Changing class of column 'CLM_ADMSN_DT' to Date
- Changing class of column 'NCH_BENE_DSCHRG_DT' to Date

Missing Data

A large set of analysis is conducted on diagnosis and procedure codes. However, lot of procedures (both ICD-9 and CPT/HCPCS codes) were found missing from both inpatient and outpatient.

For each of the following summaries, the first row of data frame shows the percentage of records in the respective column which are NA. The second row shows the percentage of records in the respective column which are blank.

For Inpatient File, summary of missing data:

1. Diagnosis codes missing - The percentage of blank codes increase above 10 beyond diagnosis code 5

	ICD9_DGNS_CD_1	ICD9_DGNS_CD_2	ICD9_DGNS_CD_3	ICD9_DGNS_CD_4	ICD9_DGNS_CD_5	ICD9_DGNS_CD_6	ICD9_DGNS_CD_7	ICD9_DGNS_CD_8	ICD9_DGNS_CD_9	ICD9_DGNS_CD_10
1	0	0	0	0	0	0	0	0	0	0
2	0	1	2	4	8	12	18	25	32	92

2. ICD-9 procedure codes missing - Large percentage of procedure codes are blank

	ICD9_PRCDR_CD_1	ICD9_PRCDR_CD_2	ICD9_PRCDR_CD_3	ICD9_PRCDR_CD_4	ICD9_PRCDR_CD_5	ICD9_PRCDR_CD_6
1	43	0	0	0	0	0
2	NA	66	78	86	90	93

3. HCPCS procedure codes missing - HCPCS codes are completely missing from the dataset

	HCPCS_CD_1	HCPCS_CD_2	HCPCS_CD_3	HCPCS_CD_4	HCPCS_CD_5	HCPCS_CD_6	HCPCS_CD_7	HCPCS_CD_8	HCPCS_CD_9	HCPCS_CD_10	HCPCS_CD_11	HCPCS_CD_12	HCPCS_CD_13	HCPCS_CD_14
1	100	100	100	100	100	100	100	100	100	100	100	100	100	100
2	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
	HCPCS_CD_15	HCPCS_CD_16	HCPCS_CD_17	HCPCS_CD_18	HCPCS_CD_19	HCPCS_CD_20	HCPCS_CD_21	HCPCS_CD_22	HCPCS_CD_23	HCPCS_CD_24	HCPCS_CD_25	HCPCS_CD_26	HCPCS_CD_27	
1	100	100	100	100	100	100	100	100	100	100	100	100	100	
2	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	
	HCPCS_CD_28	HCPCS_CD_29	HCPCS_CD_30	HCPCS_CD_31	HCPCS_CD_32	HCPCS_CD_33	HCPCS_CD_34	HCPCS_CD_35	HCPCS_CD_36	HCPCS_CD_37	HCPCS_CD_38	HCPCS_CD_39	HCPCS_CD_40	
1	100	100	100	100	100	100	100	100	100	100	100	100	100	
2	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	
	HCPCS_CD_41	HCPCS_CD_42	HCPCS_CD_43	HCPCS_CD_44	HCPCS_CD_45									
1	100	100	100	100	100									
2	NA	NA	NA	NA	NA									

For Outpatient File, similar to inpatient files, summary of missing data:

1. Diagnosis codes missing - The percentage of blank codes increase above 10 beyond diagnosis code 5
2. ICD-9 procedure codes missing- Large percentage of procedure codes are blank
3. HCPCS procedure codes missing - HCPCS codes are completely missing from the dataset

Results and Analysis

1. Exploratory Data Analysis

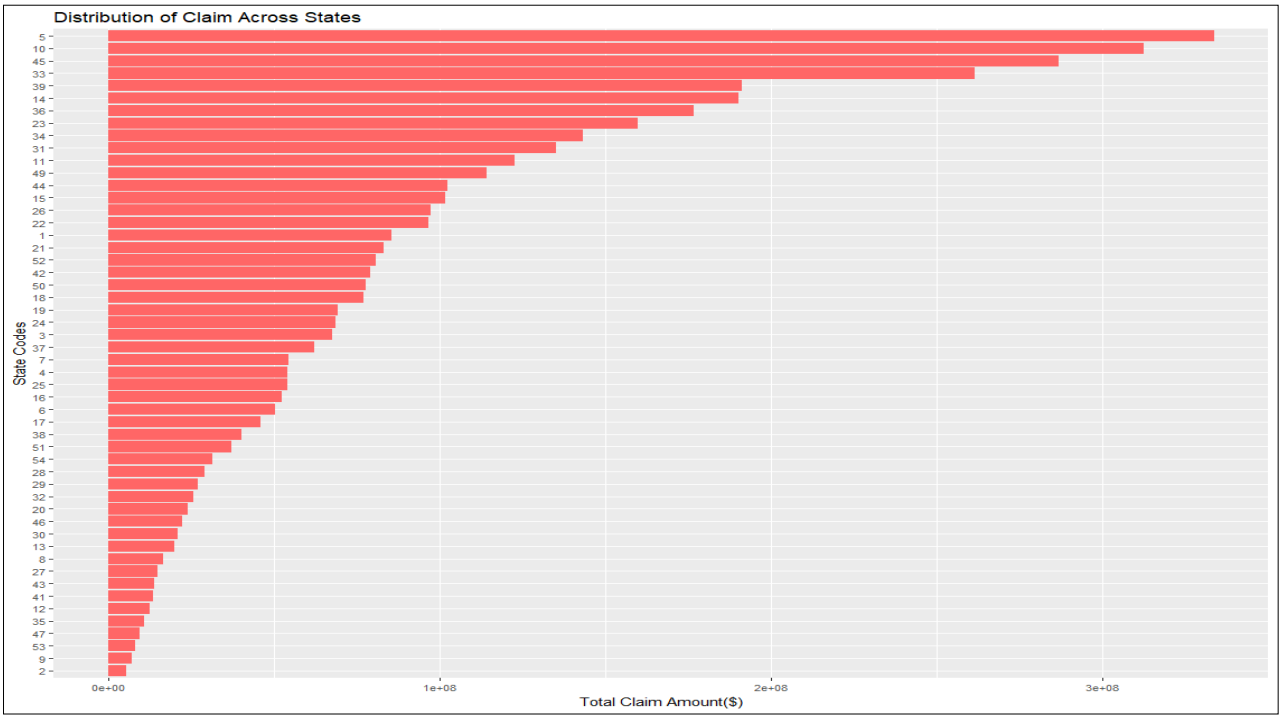
A. Medicare Reimbursement Amount

Claim payment amount is an important parameter for defining and pricing the bundles. Medicare incurs different amount for inpatient and outpatient visits. The total amount reimbursed by Medicare is the sum of these amounts for each patient. A new metric, 'Total Cost for reimbursement by Medicare' per patient is defined which takes into consideration all the amount paid by the Medicare for that patient. Hence

Total Cost for Reimbursement for Medicare (Per Patient) = Inpatient Medicare Reimbursement Amount + Outpatient Medicare Reimbursement Amount

(Addition of columns **MEDREIMB_IP** and **MEDREIMB_OP** per beneficiary found in Beneficiary files)

Distribution of Medicare Reimbursement Amount across States



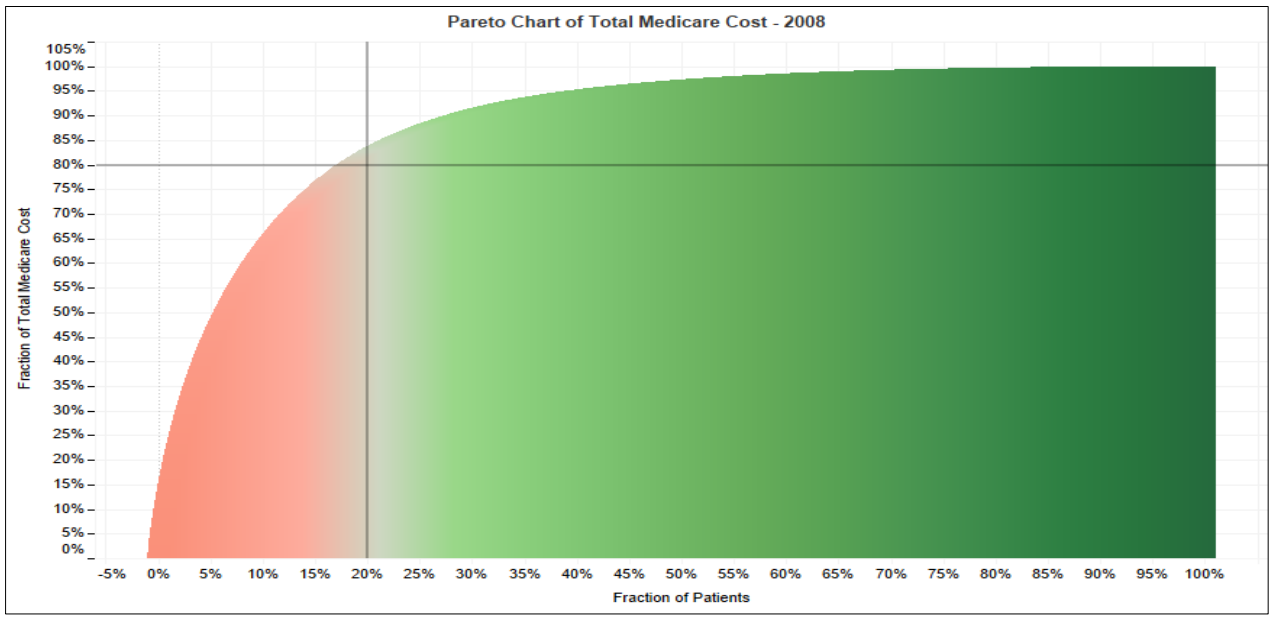
Insights

⇒ *Medicare Reimbursements Amount is found highest for state code- 05 which belongs to California followed by 10 which is Florida.*

B. Cost Bucketing

The range of cost and medical needs from patient to patient is significant. In the given dataset, some patients had $10 in total cost for reimbursement, while others had as high $229500 expenses. Moreover 80% of the overall cost of the population originated from only 20% of patients. The following figures shows the cost characteristics of the population in the given dataset. (Each image concept - (Dimitris Bertsimas, 2017))

Pareto Chart of total cost for 2008



A similar trend was observed in 2009 and 2010. (Refer Appendix for detailed charts)

Insights	*20% or less contributed to 80% of the overall cost of all patients*

In order to reduce noise in the data and at the same time reduce the effects of extremely expensive patients, the patients' cost is partitioned into five cost buckets in such a way that the sum of all patients' costs in each bucket are approximately equal (Approx. 330 Mn for 2008, 337 Mn for 2009 and 187 Mn for 2010). Each cost bucket contains patients whose total cost is equal to 20% of the entire cost, for each year. Each cost bucket, moving from low to high, is an indication for the risk of medical complication with the patient.

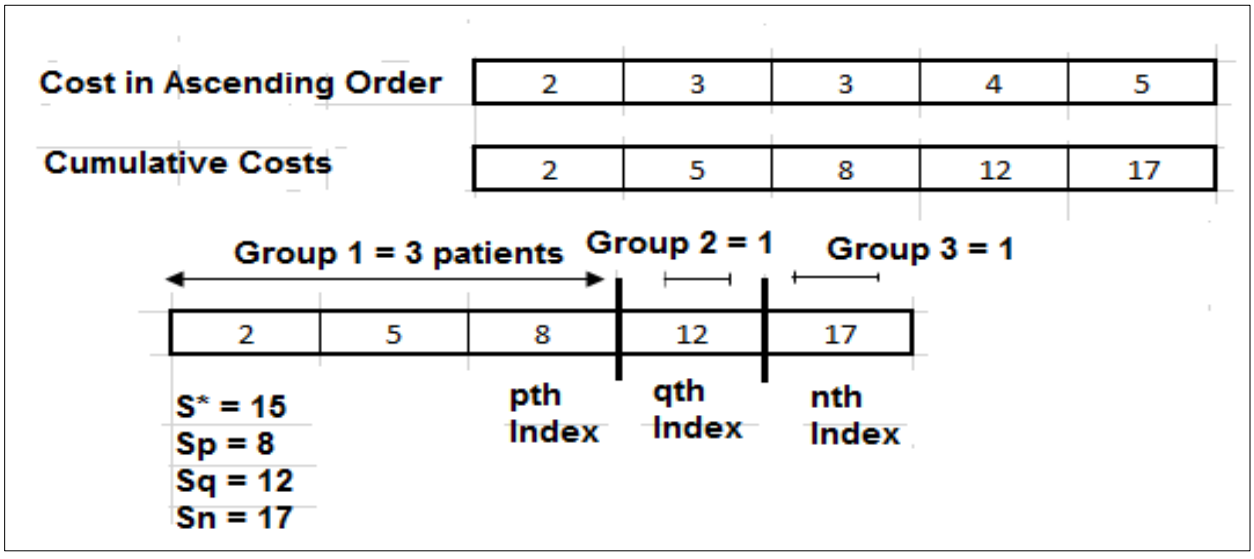
For each year, the individual cost buckets are calculated, using an optimization algorithm (Privé, 2017) as follows

1. Arrange the costs in ascending order (e.g. 2, 3, 3, 4, 5)
2. Take cumulative sum of cost (e.g. 2, 5, 8, 12, 17), Total Cumulative Sum = 44
3. Decide the number of buckets to be created (e.g. 3)
4. Calculate Average cost of all patients per bucket, S*- (Total Cumulative sum)/ (No. of buckets) (e.g. 44/3~ 15)
5. Define the optimization problem such that

If *p*, *q*, *n* are index in the cumulative vector, *Sp*, *Sq*, *Sn* are the respective cumulative sum values

By changing – *p* and *q*
Minimize – The Sum of Squared Difference of total cost per group and S*

Minimize	$D_{p,q} = (Sp - S^*)^2 + (Sq - Sp - S^*)^2 + (Sn - Sq - S^*)^2$



The above optimization is run for each of the years, and the range for cost buckets is calculated for each year. The final bucket ranges is the average of each range. Below figure gives the summary of the ranges for the year 2009

```
2009:
Bucket 1:   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
              10     170     560    1156    1580    6000

Bucket 2:   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
            6000    7180    8810    8974   10630   12930

Bucket 3:   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
           12930   14390   16330   16730   18870   22160

Bucket 4:   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
           22170   24800   28290   29010   32760   39270

Bucket 5:   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
           39270   44390   52840   55720   61290  185500
```

The five buckets, with the cost range are as follows (computed after taking average across all years)

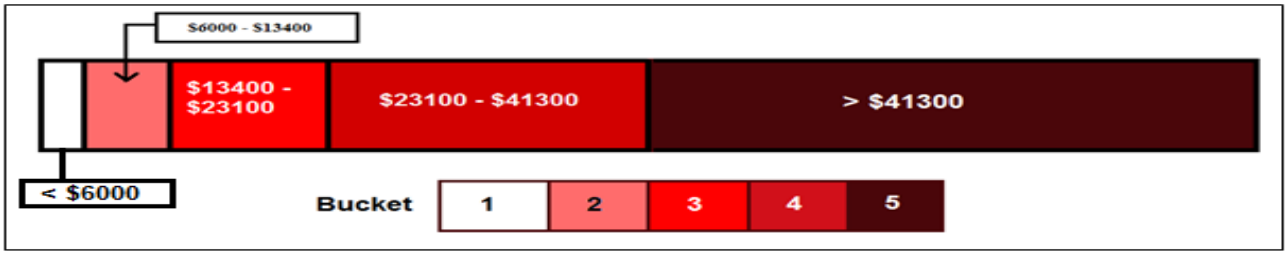


Image Concept-(Dimitris Bertsimas, 2017)

The distribution of patients for each bucket is shown below (for year 2009)

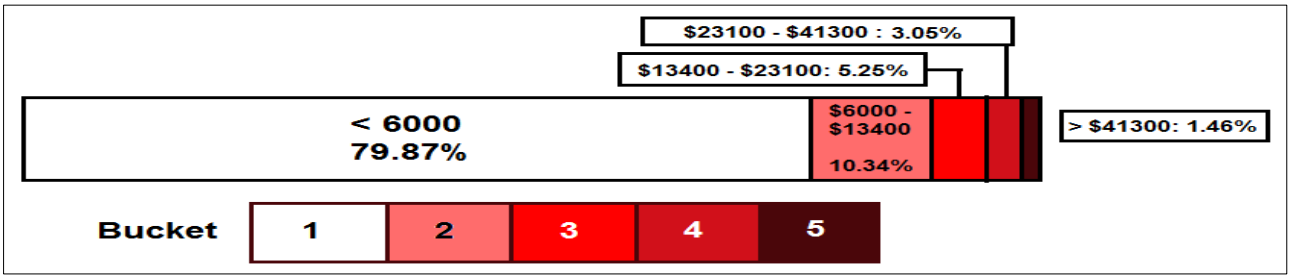


Image Concept-(Dimitris Bertsimas, 2017)

In terms of risk of medical complication, these buckets are defined as

- < 6000 - CB_1 (LOW)
- 6001 – 13400 - CB_2 (EMERGING)
- 13401 – 23100 - CB_3 (MODERATE)
- 23101 – 41300 - CB_4 (HIGH)
- 41300 - CB_5 (VERY HIGH)

It is observed that cost bucket 5 has very less patient count (as expected). For this analysis, cost bucket 4 and 5 are merged together. The final distribution of patients across different buckets for each year is shown below

Cost Bucket Frequency Per year

Year	Cb 1	Cb 2	Cb 3	Cb 4
2008	243,140.00	28,184.00	16,668.00	19,178.00
2008 (Percent)	79.15	9.18	5.43	6.24
2009	293,106.00	37,969.00	19,292.00	16,574.00
2009 (Percent)	79.88	10.35	5.26	4.52
2010	246,132.00	25,014.00	10,649.00	7,324.00
2010 (Percent)	85.13	8.65	3.68	2.53

C. Chronic Condition and their distribution across different cost buckets

The following plot shows the distribution of patients with different chronic conditions across different cost buckets from 2008 to 2010

By Percentage

Distribution of Diseases

Cost Bucket / Year1

Disease1	CB_1			CB_2			CB_3			CB_4		
	2008	2009	2010	2008	2009	2010	2008	2009	2010	2008	2009	2010
SP_ISCHMCHT	0.5922	0.6224	0.5125	0.8474	0.8148	0.6858	0.9000	0.8600	0.7211	0.9345	0.8915	0.7730
SP_DIABETES	0.5408	0.5436	0.4051	0.7724	0.7573	0.5989	0.8285	0.8050	0.6295	0.8714	0.8388	0.6464
SP_CHF	0.3750	0.4299	0.3484	0.7125	0.6657	0.5452	0.7954	0.7176	0.5733	0.8485	0.7730	0.6189
SP_CHRNKIDN	0.1790	0.2298	0.1705	0.5302	0.5275	0.4644	0.6662	0.6296	0.5406	0.7628	0.7160	0.6166
SP_DEPRESSN	0.3022	0.3183	0.2498	0.4762	0.4597	0.3355	0.5100	0.4844	0.3348	0.5446	0.5002	0.3436
SP_ALZHDMTA	0.2546	0.2918	0.2290	0.4838	0.4513	0.3433	0.5365	0.4862	0.3451	0.5843	0.5054	0.3461
SP_COPD	0.1567	0.1788	0.1172	0.4554	0.4061	0.2843	0.5412	0.4527	0.2966	0.5972	0.5040	0.3294
SP_OSTEOPRS	0.2548	0.2546	0.1833	0.3232	0.3318	0.2241	0.3459	0.3480	0.2332	0.3753	0.3489	0.2343
SP_RA_OA	0.2172	0.2282	0.1358	0.3400	0.3287	0.1830	0.4027	0.3645	0.1980	0.4079	0.3575	0.1891
SP_CNCR	0.0801	0.0985	0.0693	0.1785	0.1793	0.1209	0.2083	0.1999	0.1316	0.2458	0.2171	0.1459
SP_STRKETIA	0.0449	0.0572	0.0325	0.1659	0.1286	0.0810	0.2041	0.1550	0.0885	0.2460	0.1846	0.0976

By Absolute Count- (Refer Appendix for Absolute Count)

Insights

⇒ *Ischemic Heart Disease, Diabetes and Chronic Heart Failure are the chronic conditions found highest amongst the patients*

⇒ *Very high percent of people from cost bucket 2, 3 and 4 have the above mentioned chronic conditions as compared to patients from bucket 1.*

⇒ *For each bucket, the percentage of patients suffering from any condition increase from 2008 to 2009 but decrease from 2009 to 2010*

D. Cost Bucket Matrix

The cost bucket matrix describes the flow of patients from one bucket to another over the years.
From the below figure,

- 108221 patients fell into cost bucket 1 for all the three years. It can be said that these were the least risky patients in terms of their medical complications
- 553 patients moved from cost bucket 1 in 2008 to cost bucket 2 in 2009 to cost bucket 3 in 2010. The increase in expense over the years signal high risky patients.
- Similarly 345 patients moved from cost bucket 1 in 2008 to cost bucket 2 in 2009 to cost bucket 4 in 2010. These are the patients who are moving from least to high complications.

Cost Bucket Distribution of Patients Over Years

	CB_1_2009				CB_2_2009				CB_3_2009				CB_4_2009			
	CB_1	CB_2	CB_3	CB_4	CB_1	CB_2	CB_3	CB_4	CB_1	CB_2	CB_3	CB_4	CB_1	CB_2	CB_3	CB_4
CB_1_2008	1,08,221	11,768	4,649	2,960	14,627	1,253	553	345	6,891	580	230	173	4,910	444	187	171
CB_2_2008	10,832	978	410	274	2,209	362	162	130	1,154	222	123	116	995	238	186	141
CB_3_2008	6,101	394	145	165	1,222	198	113	115	728	172	137	149	825	284	252	180
CB_4_2008	6,330	306	159	140	1,366	245	140	105	960	304	180	141	1,193	494	298	272

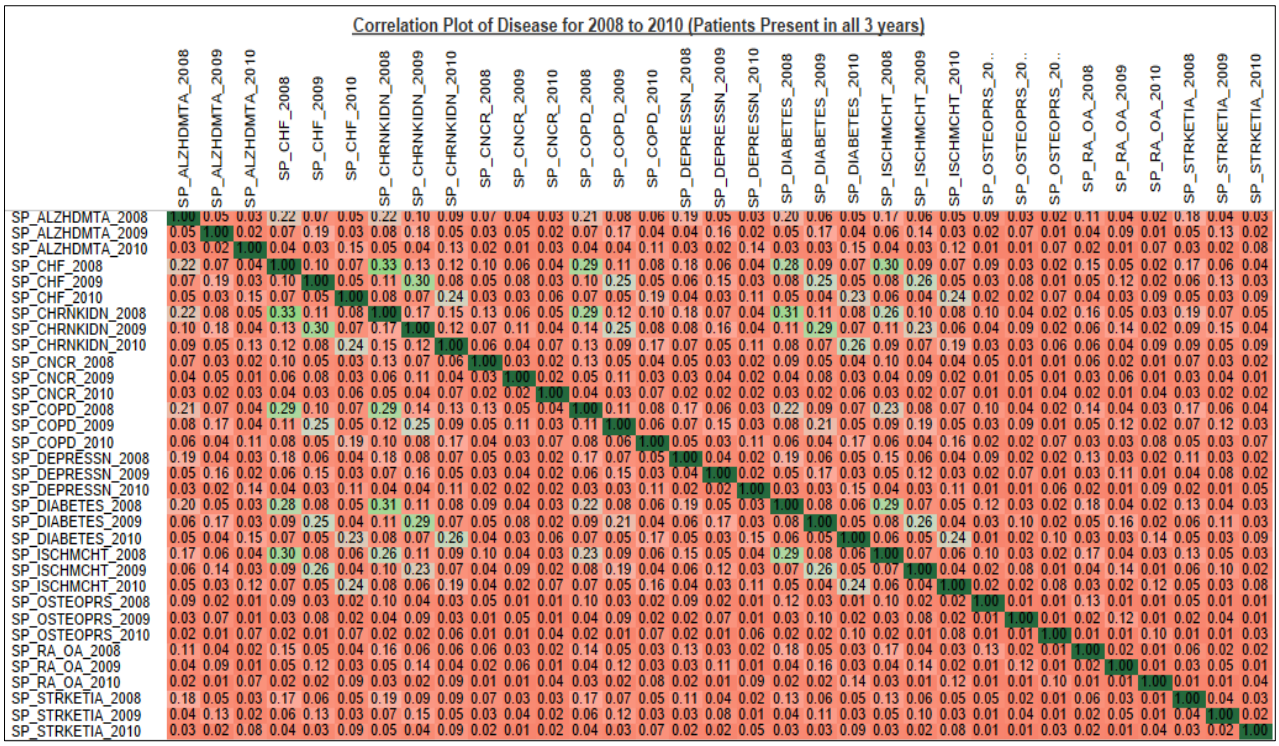
Insights

⇒ *Group of high risk patients can be identified and bundles based on their complication can be formulated.*

⇒ *Based on cost bucket in 2008 and given their chronic conditions and other medical characteristic, models which can predict that which cost bucket the patient will fall in 2009 can be built. Hence utmost care can be provided right from the early stage of complication*

E. Correlation between Chronic Conditions

The chronic condition for each patient is defined in terms of 11 different conditions for each year.
The below correlation plot attempts to check if any two chronic conditions are correlated (positively or negatively) in the same or across different years



Insights

- ⇒ *No strong positive or negative correlation between any of the two chronic conditions, within year or across different years.*
- ⇒ *Weak positive correlation of the order of 0.33 is found between Chronic Heart Failure and Chronic Kidney both in 2008.*
- ⇒ *Weak positive correlation of the order of 0.31 is found between Chronic Kidney and Diabetes*

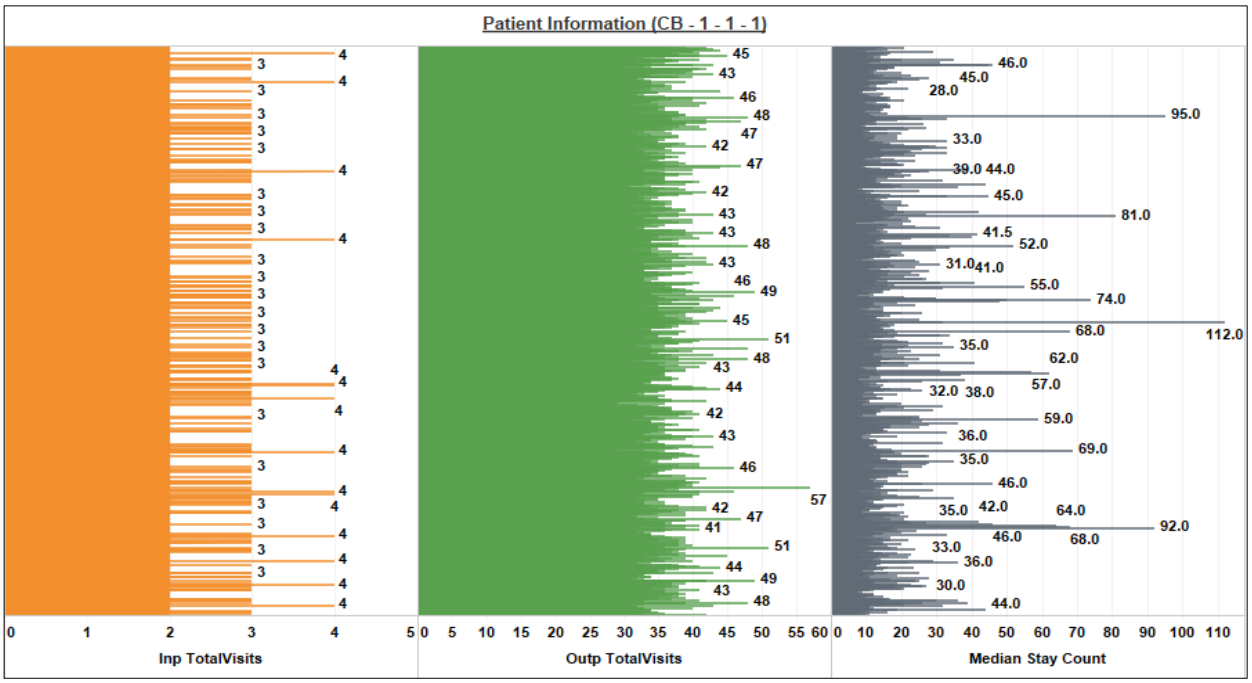
F. Inpatient and Outpatient Visits for different cost buckets

Patients belonging to different cost bucket show different characteristic in their inpatient and outpatient visits. The variation is also observed in their length of stay within hospital, which can also impact their chances of readmission based on complexity. Every time a patient revisits an inpatient or outpatient facility, the cost for Medicare increases. Cost for Medicare also increases if the number of days the patient stays in the inpatient facility is high. This number is represented by the Median Stay Count here.

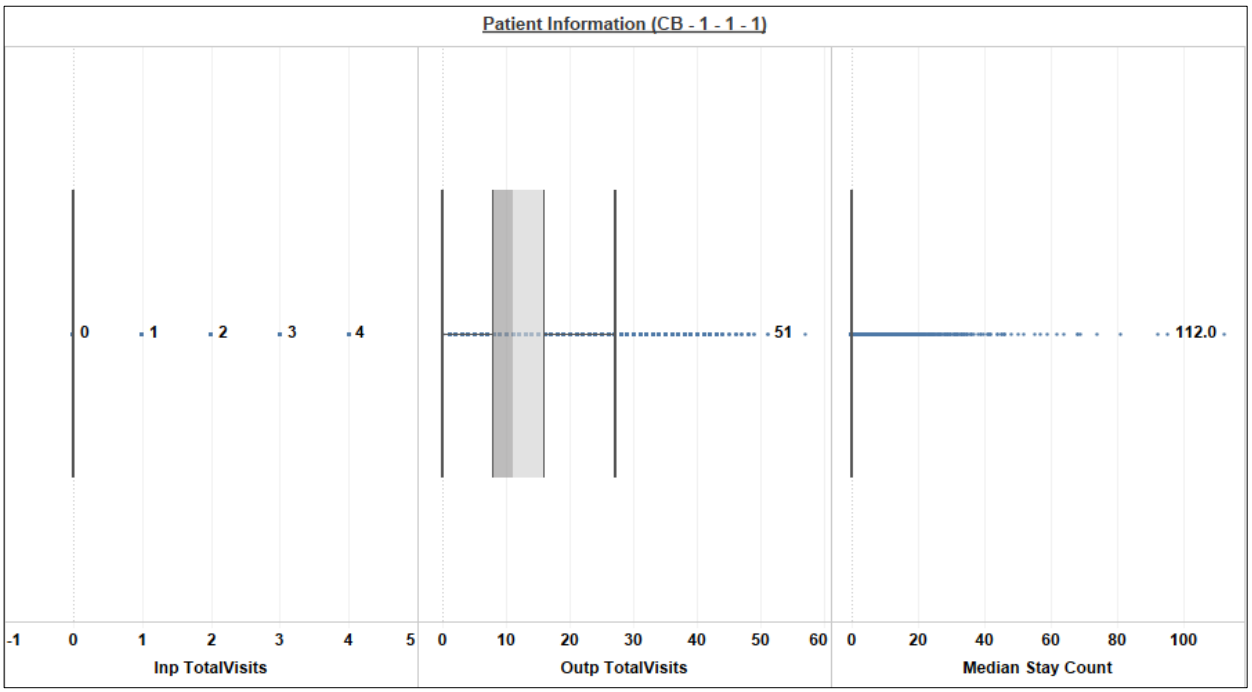
The plots below demonstrate how the number of inpatient visits, outpatient visits and median stay count are factors for patients to have high cost profiles.

Following plot shows the inpatient and outpatient visits, and the median stay length for patients belonging in CB-1, CB-1, CB-1 bucket (least risky patients, count = 108221)

- Majority of the patients have 0 inpatient visits i.e. no overnight hospital stay
- Few patients have 1, 2 or 3 visits and even fewer number with 4 visits.
- Outpatient visits lie in the range of 8 – 18

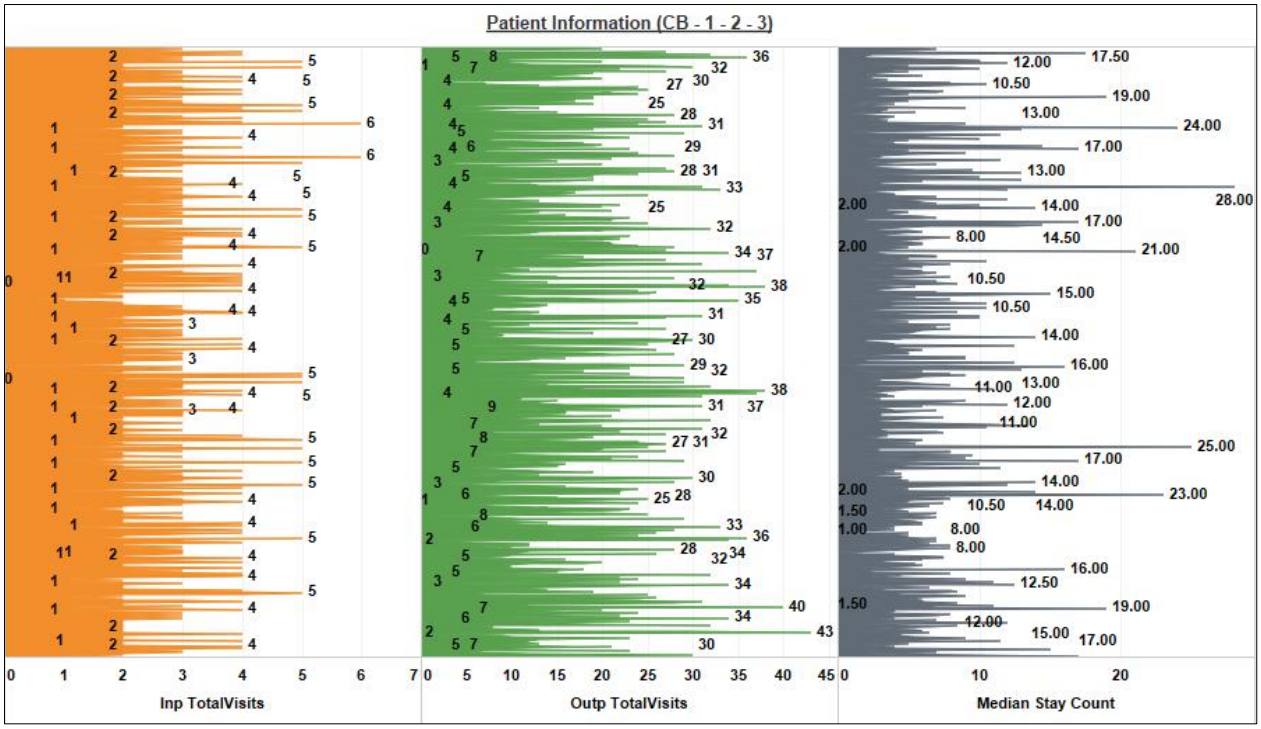


As seen below, the median stay count, if any, is less than 40 reaching 112 max

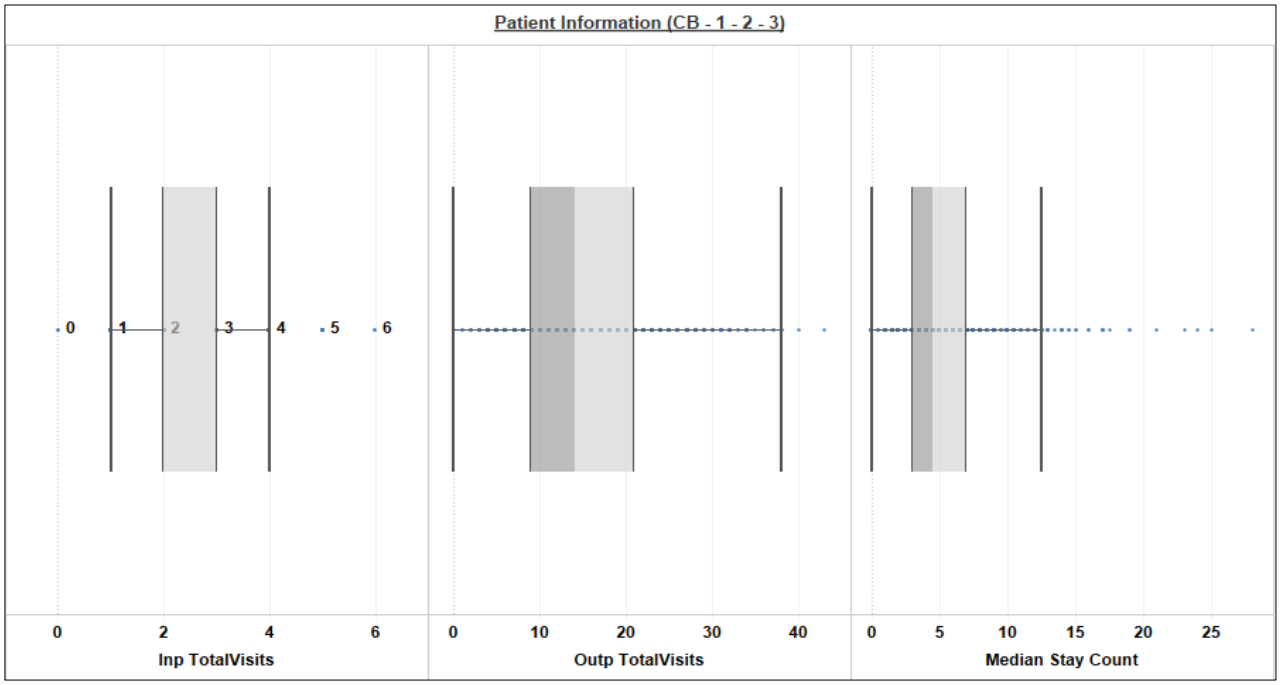


Following plot shows the inpatient and outpatient visits, and the median stay length for patients belonging in CB-1, CB-2, CB-3 bucket (High Risky patients, count = 553)

- Majority of the patients have 2 or 3 inpatient visits
- Outpatient visits lie in the range of 9 – 21

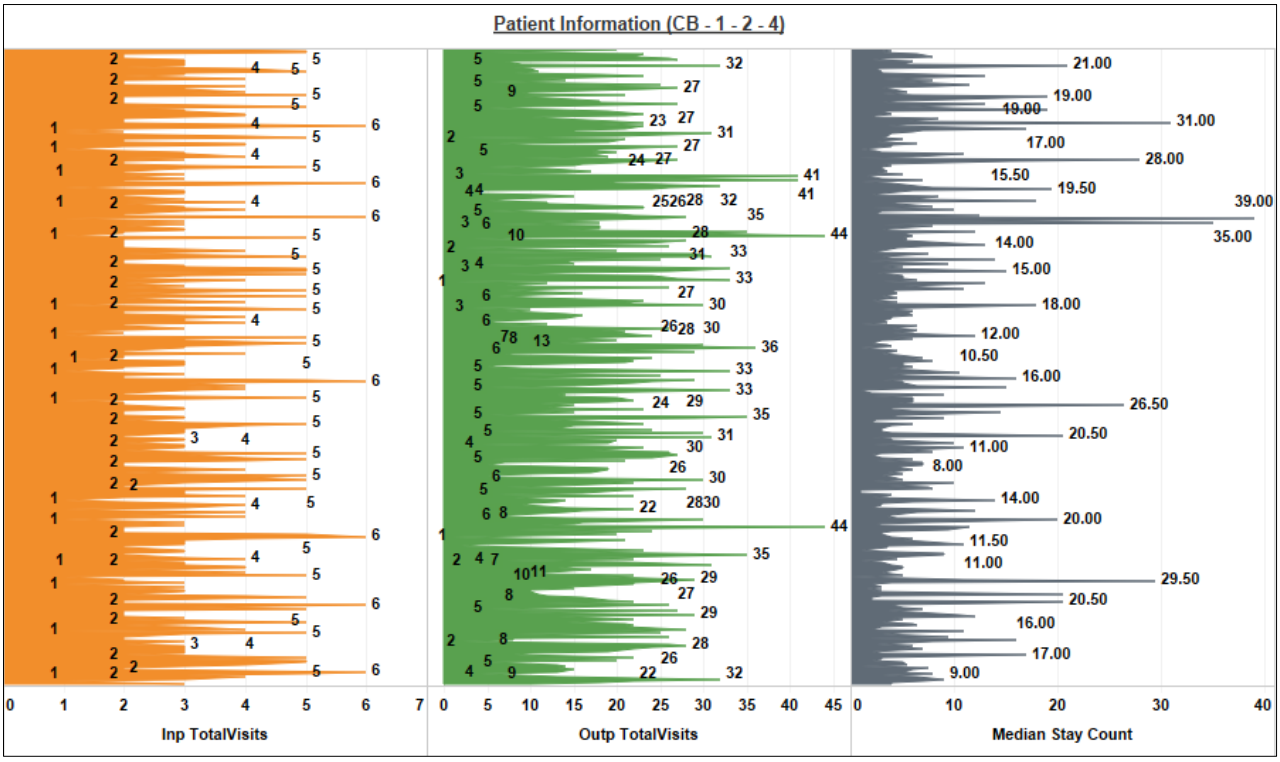


Below it is observed that, the median stay count is in the range 3 to 7, with max stay length as 28

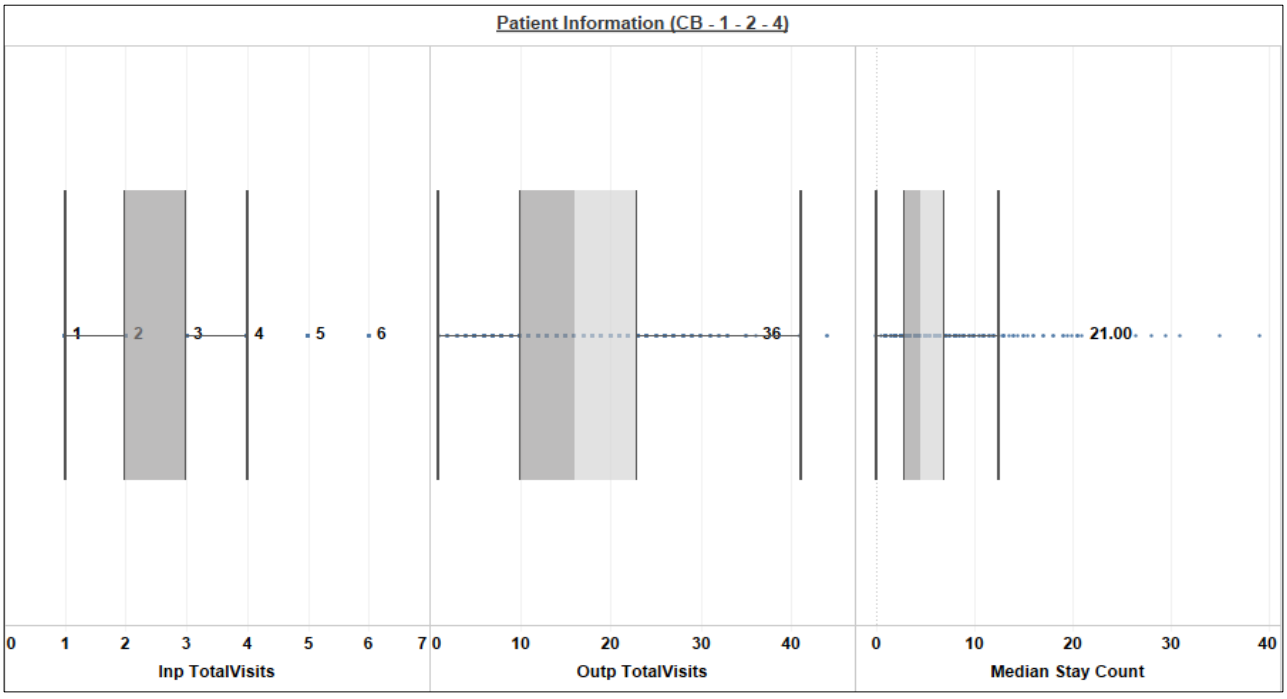


Following plot shows the inpatient and outpatient visits, and the median stay length for patients belonging in CB-1, CB-2, CB-4 bucket (Very High Risky patients, count = 345)

- Majority of the patients have 2 or 3 inpatient visits
- Outpatient visits lie in the range of 10 – 22



As seen below, the median stay count in the hospital is in the range 2 to 7, with max stay length as 39



Insights

⇒ *The number of inpatient (hospital overnight stay) increases as the patient moves from less to high risky patients.*

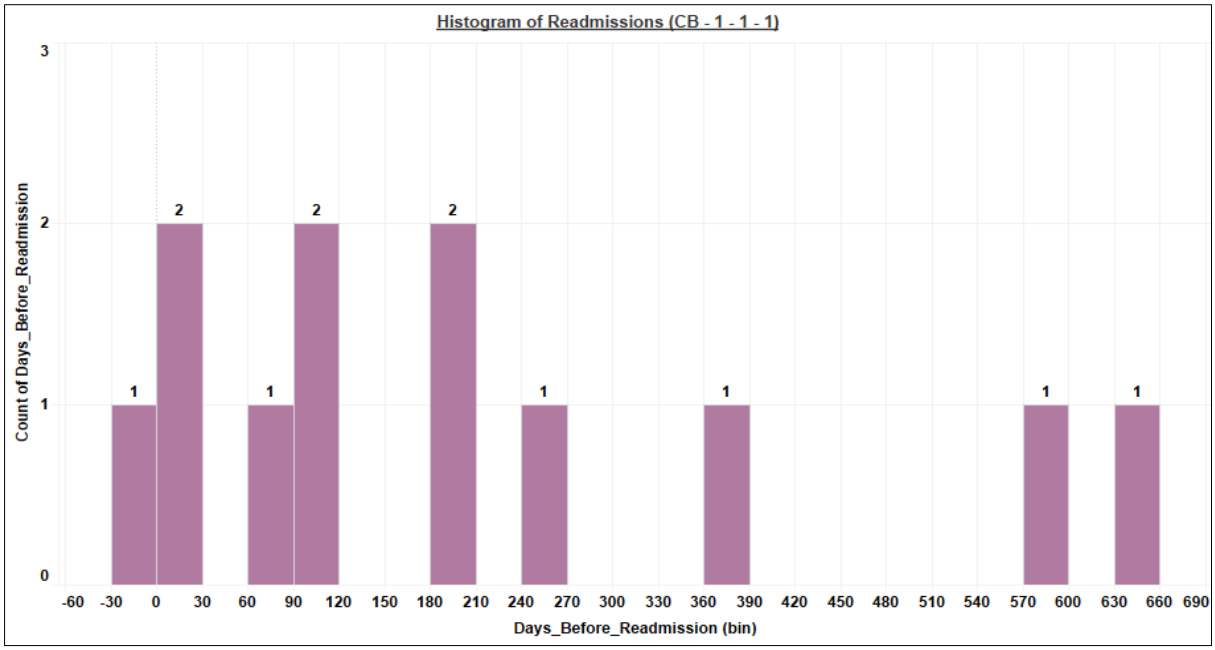
⇒ *Readmitted patients increase Medicare's reimbursement cost*

⇒ *The median stay of length within hospital is higher for less risky patients, compared to high and very high risky patients. There is a possibility that the care provided to the patient by the providers may not be sufficient enough, and which might drive readmissions for such patients.*

G. Readmission Checks for Inpatient Visits for different cost buckets

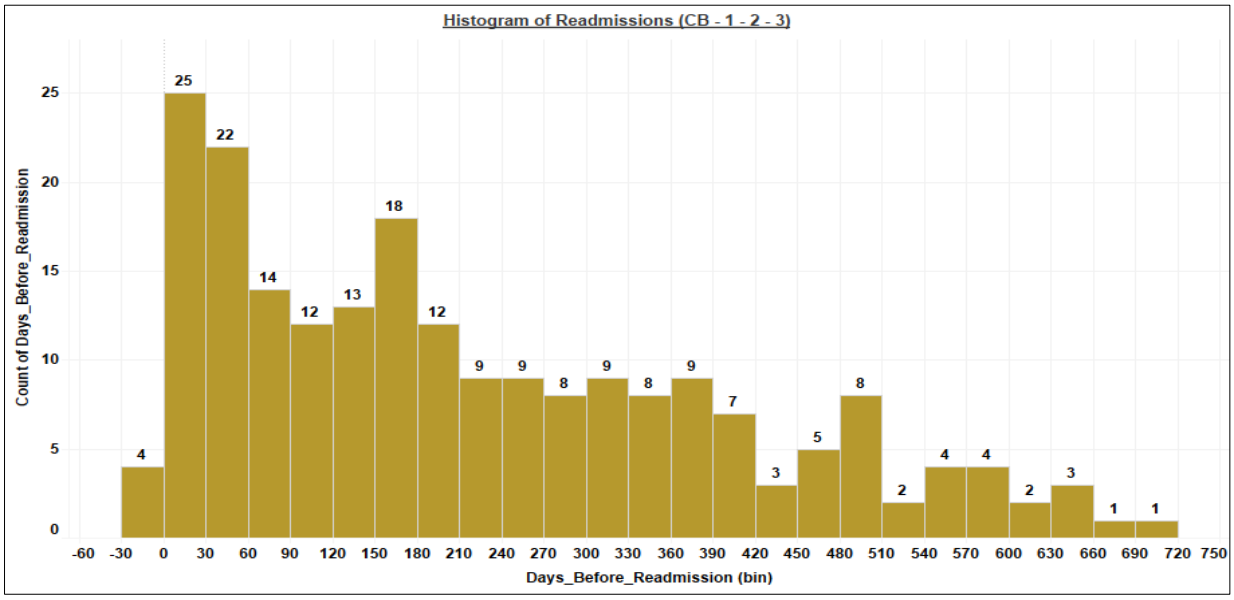
Inpatient visits drive higher cost for Medicare. Hence it is important for the providers that the care provided to the patients is of good quality which in turn will lead to lower rates of readmission. The following plot shows how frequently patients are readmitted to the hospital, across different cost buckets of patients

Readmission for patients belong to CB-1, CB-1 and CB-1 (least Risky)



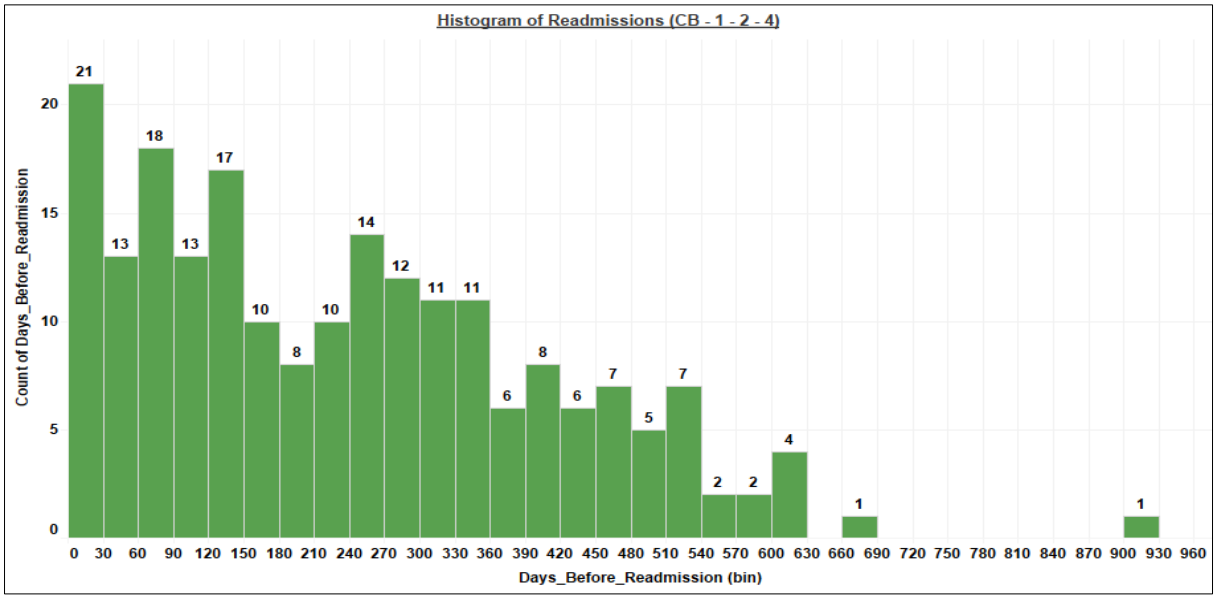
- Very few patients are readmitted within 90 days of their previous visit

Readmission for patients belong to CB-1, CB-2 and CB-3 (High Risk Patients)



- Close to 61 patients are readmitted within 90 days from their previous visit (~ 11% of patients in this group)

Readmission for patients belong to CB-1, CB-2 and CB-4 (very High Risk Patients)



- Close to 52 patients are readmitted within 90 days from their previous visit (~15% of patients in this group)

Insights

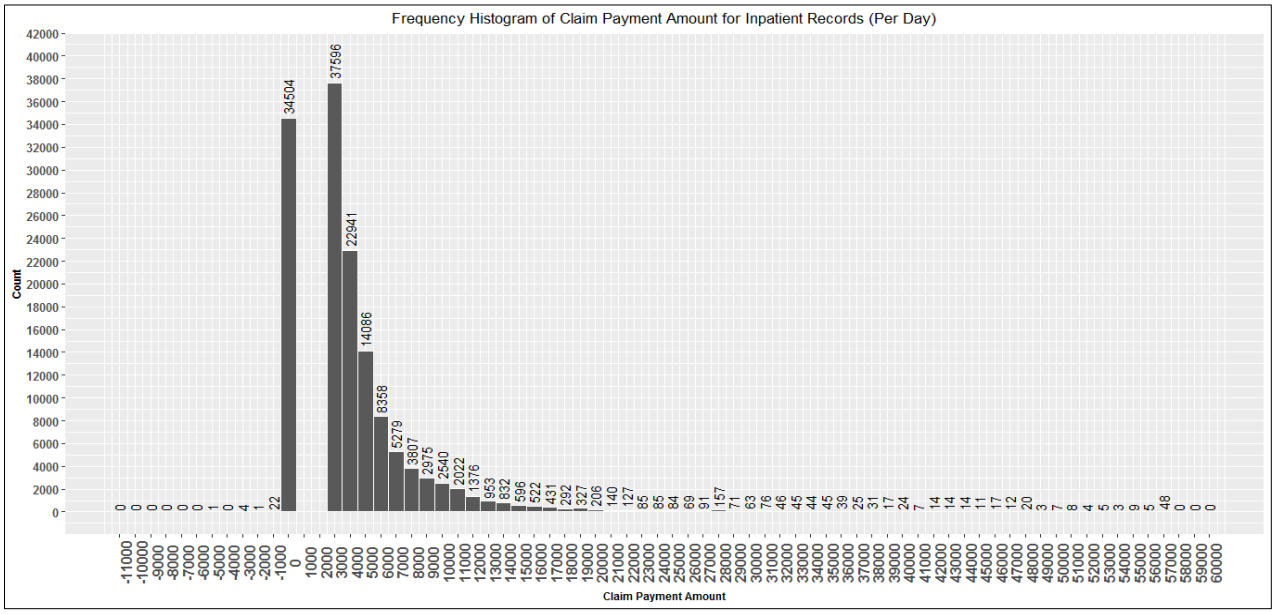
⇒ *Readmission within 90 days is higher for high risk and very high risk patients, compared to low risk patients*

⇒ *Medicare can identify frequently readmitting patients and their characteristic and probably bundle the services*

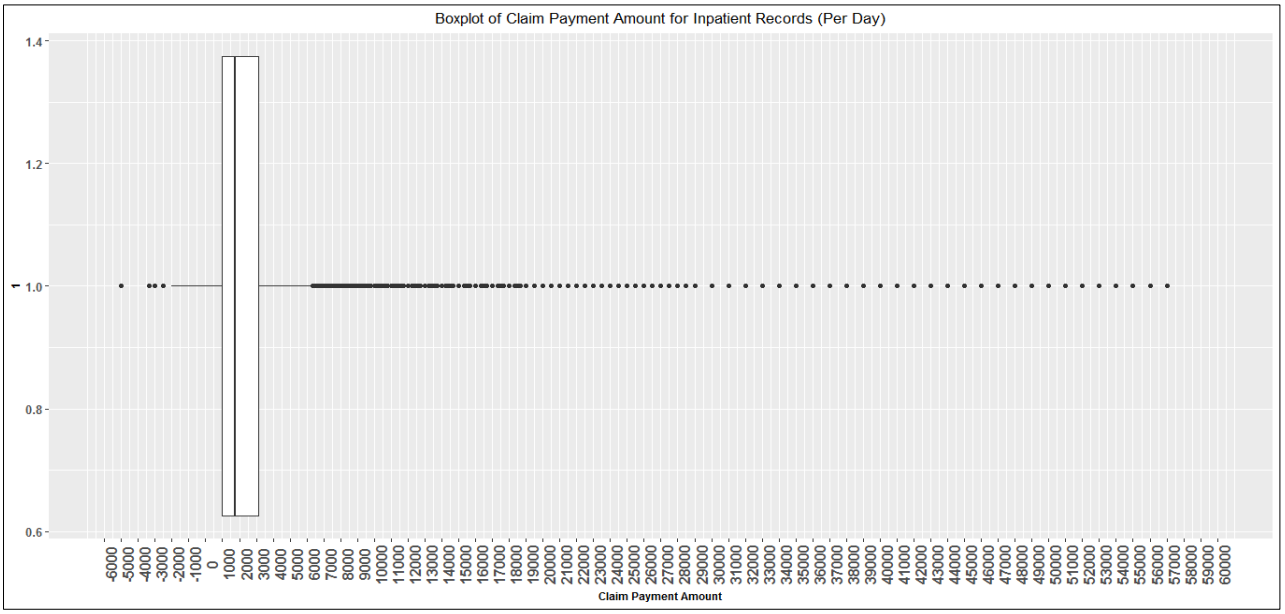
H. Claim Payment Amount for Inpatient and Outpatient visits

The claim payment amount for each inpatient and outpatient visit varies significantly.

Histogram of Claim payment Amount for Inpatient (Per Day)

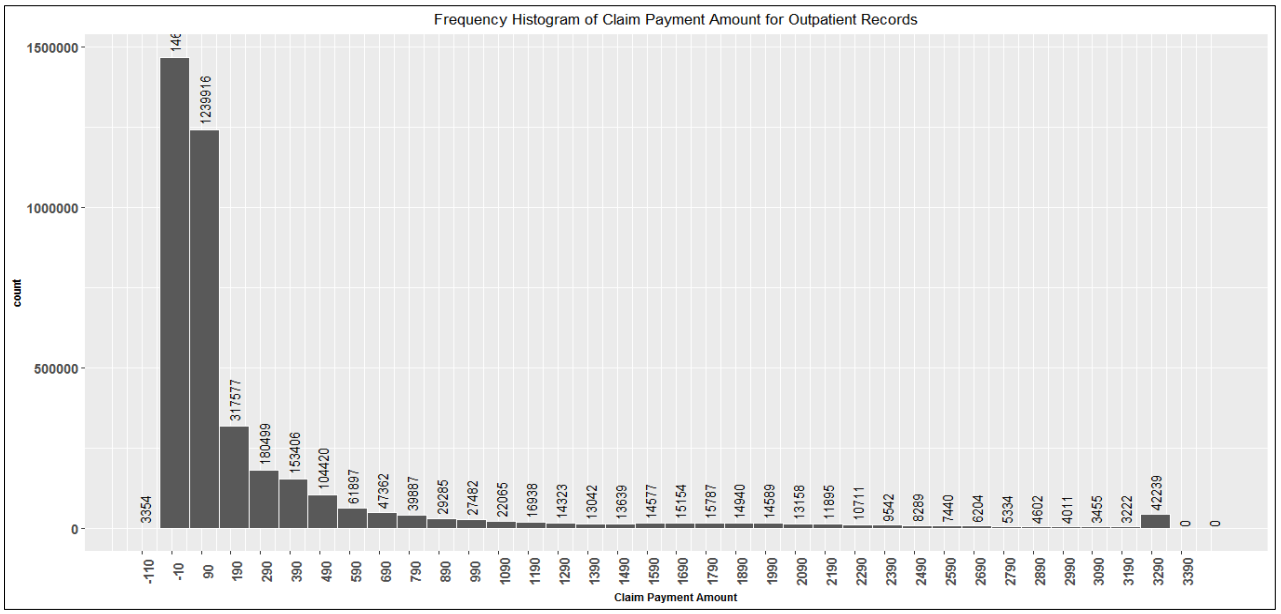


Boxplot of Claim payment Amount for Inpatient (Per Day)



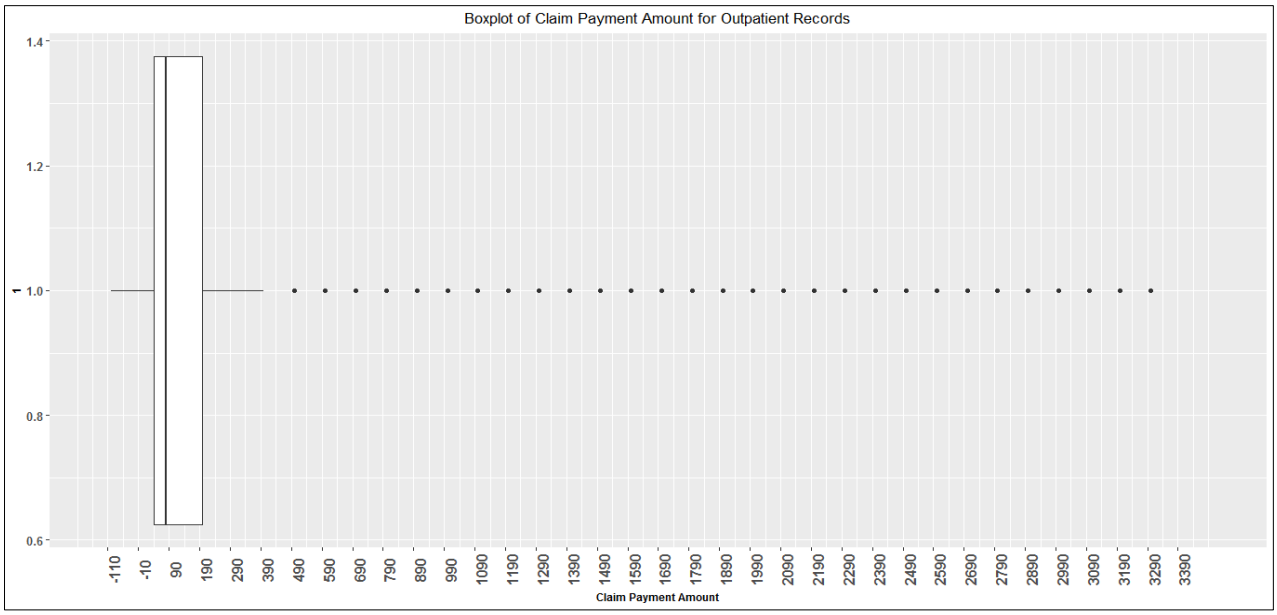
Median cost for each inpatient visit per day ranges between $1000 and $3000

Histogram of Claim payment Amount for Outpatient



It is evident from the graph above that a majority of the patients had claim

Boxplot of Claim payment Amount for Outpatient



Insights

⇒ *The amount that Medicare provides as Claim Payment Amount to providers is significantly higher for inpatient facilities than outpatient facilities per day.*

⇒ *As readmission increase, the cost to Medicare will increase.*

2. Bundling Framework

The following section discusses a possible framework using data mining techniques such as Association rule mining, Sequential rule mining and Process Mining. The process is formulated to provide an end-to-end result, beginning from selecting a particular diagnosis, exploring for possible sequences and subsequence, co-occurrences of diseases, observing patient traces and evaluating the cost of probable bundle formed using these steps. This approach is just ONE of the many to bundle. The results are obtained programmatically and hence the framework can be extended to include all combinations.

The summary of the entire process is

A. Combining Inpatient and Outpatient records with necessary columns
B. Finding co-occurring Diagnosis using Association Mining
C. Running Association Mining for Inpatient records with results of Association Mining for frequently occurring diagnosis (e.g. Ischemic Heart Disease)
D. Sequential Rule Mining on Inpatient records with a particular diagnosis using ICD-9 procedure codes
E. Process Mining using procedure codes for Inpatient records with a particular diagnosis
F. Pricing the Inpatient Bundle for the selected diagnosis
G. Building data for bundling for Outpatient Records
H. Running Sequential Mining and Process Mining on data built in step G for Outpatient Visits.
I. Pricing the Outpatient bundle

A. Combining Inpatient and Outpatient records with necessary columns

Following is the list of columns selected from inpatient and outpatient files for analysis

File Type	Columns				
Inpatient	DESYNPUF_ID	CLM_PMT_AMT	CLM_ADMSN_DT	NCH_BENE_DSCHRG_DT	PRVDR_NUM
	ICD9_DGNS_CD_1 to ICD9_DGNS_CD_5	ICD9_PRCDR_CD_1 to ICD9_PRCDR_CD_5	Days_Before_Readmission		
Outpatient	DESYNPUF_ID	CLM_PMT_AMT	CLM_FROM_DT	CLM_THRU_DT	PRVDR_NUM
	ICD9_DGNS_CD_1 to ICD9_DGNS_CD_5	HCPCS_CD_1 to HCPCS_CD_5			

- Outpatient contains records of only those patient IDs which have records in inpatient file.
- The above two files are combined to create a new data frame.
- Few column names are renamed from inpatient and outpatient files (e.g. all procedures, whether ICD or HCPCS are termed Procedure 1, Procedure 2 etc.) and additional columns are calculated. The final list of columns is as follows

File	Column	Description
Combined	DESYNPUF_ID	Patient ID
	CLM_PMT_AMT	Claim Payment Amount
	CLM_FROM_DT	Claim Start Date
	CLM_THRU_DT	Claim End Date
	PRVDR_NUM	Provider Number
	DGNS_1 to DGNS_5	ICD-9 Diagnosis code
	PRCDR_1 to PRCDR_5	ICD-9 or HCPCS Procedure code
	Days_Before_Readmission	Number of days before next hospital visit
	Type	Record is Inpatient or Outpatient
	Diagnosis1 to Diagnosis5	Sub_Chapter description of the Diagnosis code obtained from icd package in R
	SIZE	Number of Procedure codes present per record
	AVG_AMT	Average Claim amount per procedure, Claim Payment Amount/SIZE
	DGNS_1_3digit to DGNS_5_3digit	3 digit ICD-9 Diagnosis code

- Since there are blank codes between two procedure columns, the columns are shifted left such that the starting procedure codes are no blank

After completing the steps above it was observed that over **4.8%** of the times when patients were hospitalized (inpatient records), they were diagnosed with Ischemic Heart Disease.

	V1	N	Pct
1	Other Forms Of Heart Disease	128446	7.9574614133
2	Other Metabolic And Immunity Disorders	120988	7.4954248593
3	Hypertensive Disease	102275	6.3361207515
4	Symptoms	78002	4.8323646136
5	Ischemic Heart Disease	77522	4.8026277477
6	Diseases Of Other Endocrine Glands	54742	3.3913656532
7	Diseases Of The Blood And Blood-Forming Organs	51291	3.1775699777
8	Nephritis, Nephrotic Syndrome, And Nephrosis	50619	3.1359383654

Similarly while analysing Outpatient diagnoses it was found that over **4.6%** of the times patients were diagnosed with 'Other forms of Heart Disease' like Chronic Heart Failure.

	V1	N	Pct
1	Symptoms	447707	8.509705322
2	Persons Encountering Health Services For Specific Procedur...	409493	7.783360013
3	Hypertensive Disease	339255	6.448324639
4	Other Metabolic And Immunity Disorders	283721	5.392772737
5	Other Forms Of Heart Disease	244086	4.639418042
6	Diseases Of The Blood And Blood-Forming Organs	224929	4.275295022
7	Arthropathies And Related Disorders	198506	3.773064894
8	Diseases Of Other Endocrine Glands	183563	3.489038675
9	Persons With A Condition Influencing Their Health Status	160450	3.049722740

B. Finding co-occurring Diagnosis using Association Mining

Association Mining is run on all Inpatient records using DGNS_1_3digit to DGNS_5_3digit to check for co-occurring diagnosis codes i.e. to check whether a diagnosis belonging to a major or subchapter co-occurs with another major or subchapter very frequently. Steps for association mining:

- Subset dataset keeping only Diagnosis 3digit codes
- Removing duplicate codes, if found on any row
- Storing the records in a text file
- Read the records as 'transactions'
- Run apriori algorithm for association mining with support = 0.01, confidence = 0.1
- Subset rules with Lift > 1

Some of the rules obtained are as follows

	lhs		rhs	support	confidence	lift
[1]	{995}	=>	{038}	0.01152628	0.5025587	10.377989
[2]	{038}	=>	{995}	0.01152628	0.2380213	10.377989
[3]	{584}	=>	{276}	0.01496009	0.2301922	1.518727
[4]	{491}	=>	{428}	0.01052112	0.2093288	1.435815
[5]	{038}	=>	{276}	0.01033453	0.2134112	1.408012
[6]	{585}	=>	{428}	0.01449664	0.1942260	1.332223
[7]	{486}	=>	{428}	0.01262173	0.1882828	1.291458
[8]	{427}	=>	{428}	0.02694984	0.1879762	1.289355
[9]	{428}	=>	{427}	0.02694984	0.1848526	1.289355
[10]	{584}	=>	{428}	0.01210109	0.1862005	1.277175

Though we get some rules with good lift and confidence values, the support for the same is not very strong.

Running association mining on the entire dataset generated many rules. However, it is more meaningful to bundle based on a particular diagnosis. For this analysis, the diagnosis '**Ischemic Heart Disease**' is selected since it was proven to be the fifth most commonly occurring disease among the inpatient records. All further analysis is focused on patients diagnosed with Ischemic Heart Disease.

Data frame DF = Inpatient records diagnosed with Ischemic Heart Disease

Wiki

⇒ *Ischemic Heart Disease is a condition where blood supply to the heart is reduced due to blocked blood vessels.*

C. Running Association Mining for Inpatient records with Ischemic Heart Disease

Association mining was run on 3 digit diagnosis codes with support = 0.01 and confidence = 0.1. Following patterns were observed:

- 1 out of every 100 patients who were already diagnosed with Ischemic Heart Diseases were also diagnosed with Acute kidney failure (584) and Acute myocardial infarction (410)
- 3 out of every 100 patients who already were diagnosed with Ischemic Heart Diseases were also diagnosed with Acute myocardial infarction (410) and Heart failure (428)
- 3 out of every 100 patients who already were diagnosed with Ischemic Heart Diseases were also diagnosed with Cardiac dysrhythmias (427) and Heart failure (428)

It could be concluded that Chronic Ischemic Heart Disease patients might also suffer from diseases like Chronic Kidney Disease or Other Forms of Heart Disease like Chronic Heart Failure. It could also mean the reverse that Chronic Kidney Disease patients or Chronic Heart Failure patients might also be suffering from Chronic Ischemic Heart Disease patients.

	lhs		rhs	support	confidence	lift	count
[1]	{584}	=>	{410}	0.01502935	0.3001499	1.900636	1001
[2]	{518}	=>	{410}	0.01568998	0.2927991	1.854088	1045
[3]	{585}	=>	{410}	0.01575004	0.2363145	1.496411	1049
[4]	{414,585}	=>	{428}	0.01156104	0.2440571	1.396832	770
[5]	{403}	=>	{410}	0.01309250	0.2182728	1.382166	872
[6]	{411}	=>	{272}	0.01078030	0.2089031	1.353592	718
[7]	{585}	=>	{428}	0.01555485	0.2333859	1.335757	1036
[8]	{403}	=>	{428}	0.01243187	0.2072591	1.186223	828
[9]	{584}	=>	{428}	0.01029984	0.2056972	1.177283	686
[10]	{410}	=>	{428}	0.03118478	0.1974710	1.130202	2077
[11]	{428}	=>	{410}	0.03118478	0.1784824	1.130202	2077
[12]	{427}	=>	{428}	0.03142501	0.1929744	1.104466	2093
[13]	{428}	=>	{427}	0.03142501	0.1798574	1.104466	2093

D. Sequential Rule Mining on data frame DF using ICD-9 procedure codes

Sequential Mining using procedures is used to check the presence of frequent sequences or subsequence occurring within the dataset. Steps to generate the sequences are:

- Start with data frame DF
- Remove rows where all the procedure codes are either blank or NA
- Keep records from those patient IDs which have more than 1 record (per patient ID). Sort the data frame by Patient ID, Date of Claim and rank the records starting from 1 up to number of records present per patient ID
- Note the patient IDs obtained at the end of cleaning and processing (used for other analysis)
- Sequential mining is run on this dataset with support = 0.01 to get rules on the procedures performed on Ischemic Heart Disease patients.

The results of the sequential mining showed patients diagnosed with Ischemic Heart Disease have the following inpatient procedures:

	rule	support	confidence	lift
10	<{66,2724}> => <{66}>	0.01535088	0.1671642	0.5699205
8	<{66}> => <{66}>	0.04550439	0.1551402	0.5289265
5	<{4111}> => <{66}>	0.01260965	0.1513158	0.5158879
9	<{66,4019}> => <{66}>	0.01398026	0.1495601	0.5099022
2	<{2724}> => <{66}>	0.02384868	0.1294643	0.4413885
4	<{4019}> => <{66}>	0.02741228	0.1162791	0.3964350
6	<{41401}> => <{66}>	0.01589912	0.1084112	0.3696113
22	<{4019}> => <{4019}>	0.02494518	0.1058140	0.4488480
7	<{4280}> => <{66}>	0.01014254	0.1010929	0.3446606

Wiki	⇒ *Ischemic Heart Disease is a condition where blood supply to the heart is reduced due to blocked blood vessels.* ⇒ *PTCA is a procedure to open up blocked coronary arteries using stents.* ⇒ *Biopsy of the mouth is a diagnostic tool to determine various diseases similar to blood tests.*

- Percutaneous transluminal coronary angioplasty [PTCA] (66) and Biopsy of mouth, unspecified structure (2724) were performed first in the same visit followed by PTCA (66) again. This was observed for 1 in every 100 patients
- For 2 in every 100 patients, Biopsy of mouth (2724) was followed by PTCA (66)
- For 4 in every 100 patients, PTCA was performed in consecutive visits

E. Process Mining on a Procedure for Inpatient Ischemic Heart Disease Patients

Though sequential mining in the previous step generated few frequent sequence or subsequence, before the bundle is concluded it is of advantage to observe an end-to-end trajectory of the patient with respect to procedures performed for Ischemic Heart Disease. Here, through visual analysis, it is possible to explore the entire path of patients, and check whether few things can be added or deleting before finalising the bundle.
It is assumed here that procedure code 1 is the major procedure performed. However this can be replaced with any other procedure code or description which summarizes the major objective of the patient visiting the hospital. This is because process mining accepts only one activity per record – This could be any major procedure code or some description. The objective here is to trace the different activities (major procedure or description) performed on patients and check an end-to-end flow

To perform Process Mining:

- Start with dataframe DF
- Remove rows where all the procedure codes are either blank or NA
- Obtain patient ID, Claim start date, procedure code 1
- Keep records from those patient IDs which have more than 1 record (per patient ID).
- Sort the dataframe by Patient ID, Date of Claim and rank the records starting from 1 up to number of records present per patient ID
- Generate event log with following data which accepts data like:
  - Name of dataset
  - Case ID (patient ID)
  - Activity ID (Procedure Code 1)
  - Activity Instance ID (Rank)
  - Lifecycle ID (Type of record – Inpatient or Outpatient)
  - Timestamp (Date of Claim)
  - Resource ID (Could be the taxonomy of physician who conducted the procedure)

A summary of event log generated on inpatient records with ischemic heart disease on procedure 1 is

Event log consisting of:	
3648 cases	Unique Patient IDs
7782 events	Total Number of records in the event log
2615 traces	Unique end-to-end traces found within patients
600 activities	Unique Procedure Codes
6 activity instances	Highest number of record for any patient ID. For all patients, the rank runs from 1 minimum to 6 maximum

Top 10 traces (paths) found from the above event log

	trace	trace_id	absolute_frequency	relative_frequency
1	66,66	1534	121	0.0331688596
2	66,3722	1437	41	0.0112390351
3	3722,66	444	37	0.0101425439
4	3722,3722	381	25	0.0068530702
5	9904,66	2464	20	0.0054824561
6	3893,66	783	14	0.0038377193
7	3995,66	959	14	0.0038377193
8	66,9904	1624	14	0.0038377193
9	66,3995	1486	13	0.0035635965
10	3722,3995	410	13	0.0035635965

Most of the frequently occurring traces (paths) are length 2. The above plot gives a visual understanding of different traces along with their frequency of occurrence in the data.

Based on the above two analysis – Sequential Mining and Process Mining , for an Ischemic Heart Disease patient visiting inpatient facility, 66 and 2724 can be part of the bundle. Since the frequency of occurrence is high, it would make sense for Medicare to at least bundle these two procedures.

A probable solution that can be suggested is

Inpatient Bundle for Ischemic Heart Disease Patient = **Biopsy of the mouth (Procedure 1)** + **PTCA (Procedure 2)**

F. Pricing the Inpatient Bundle

The next step is to decide the price for the bundle discovered. One of the major challenge of this dataset is that it doesn't mention individual costs of procedures. To hypothetically demonstrate how pricing can be done, following steps explain an approximate solution

- Filter the dataset with IDs where the above bundle (sequence of procedure) exist
- For every patient ID, mark the record where biopsy was performed as cost1 and the record where PTCA was performed as cost2

	DESYNPUF_ID	CLM_PMT_AMT	PRVDR_NUM	PRCDR_1	PRCDR_2	PRCDR_3	PRCDR_4	PRCDR_5	AVG_AMT	Rank
	)F4D76166DAF2A	All	All	All	All	All	All	All	All	All
1	1B9F4D76166DAF2A	8000	1800QU	66	V4582	4019	V171	2724	1600	Cost1
2	1B9F4D76166DAF2A	8000	1801WK	66	V4581	2724	41400	V173	1600	Cost1
3	1B9F4D76166DAF2A	14000	1800QU	66	99672	2768	27800	39891	2800	Cost2

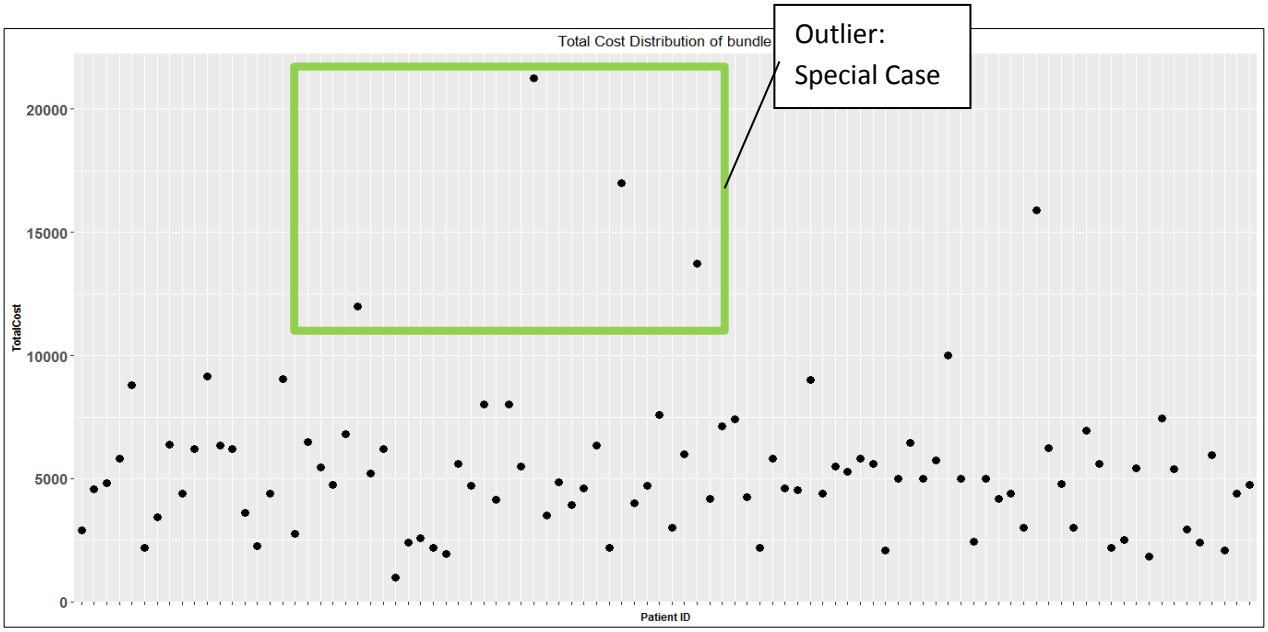
- For every patient ID, calculate the average of cost1 and cost2

	DESYNPUF_ID	Cost1	Cost2
	1B9F4D76166DA	All	All
1	1B9F4D76166DAF2A	1600	2800

- For each patient ID, calculate the total cost of the bundle as cost1 + cost2

	DESYNPUF_ID	Cost1	Cost2	TotalCost
	)F4D76166DAF2A	All	All	All
1	1B9F4D76166DAF2A	1600	2800	4400

The total cost of patients is plotted below. From the graph it is evident that the cost does not cross $10000 for majority of the patients. Hence, for Inpatient Bundle we can set a cost of $10000 as a threshold and average cost as $5400 and median cost as $4900. For outliers, as highlighted below, they will be considered as special cases and Medicare would have to look into these patients on a case by case basis. If there are many outliers, the average cost and threshold will change accordingly.



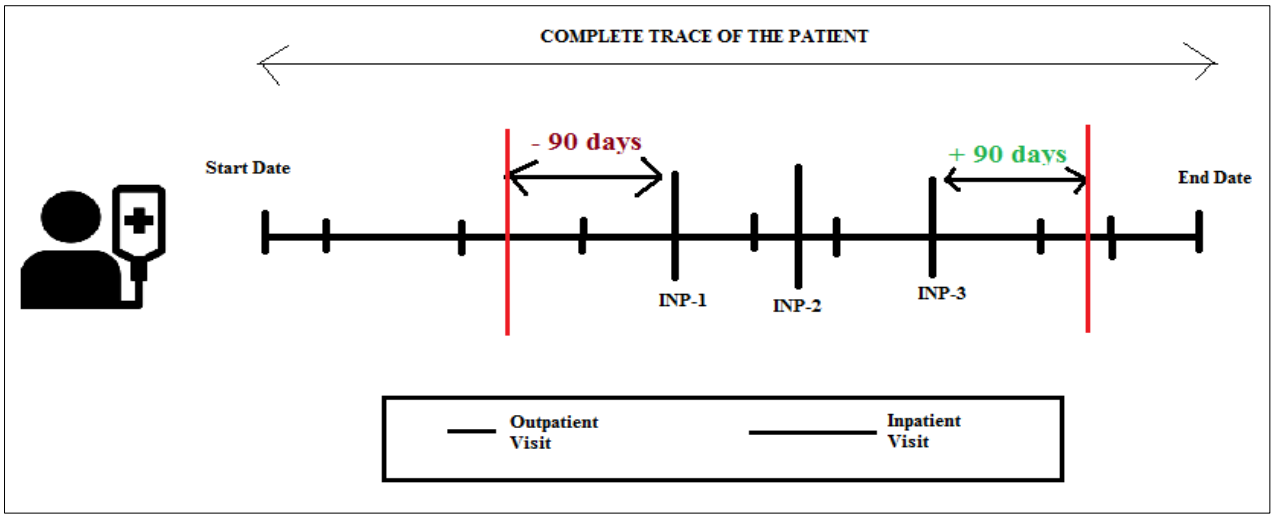
G. Bundling for Outpatient Records

This analysis focuses on running sequential and process mining (as discussed in previous steps) using outpatient records on procedures performed in outpatient facilities. However the focus is on those patients who were diagnosed with Ischemic heart disease during inpatient facility. Following steps summarize bundling for outpatient

- Obtain the patient IDs on whom Sequential Mining was performed in previous step
- Any patient ID will contain many outpatient visits. However the focus is on those records which fall around the dates of inpatient visit where they were diagnosed with ischemic heart disease.
- Define a window of including the admission date of first inpatient visit (for a particular diagnosis) minus 90 days and discharge date of last inpatient visit (same particular diagnosis) plus 90days
- Fetch all the outpatient records falling during this period. Hence these are outpatient records of the patients who were diagnosed with say ischemic heart disease during their each hospital visit. It is believed that outpatient facility visit is linked with inpatient visit and hence procedures performed in outpatient facility belonging to such nature can be used for probable bundling

Consider an example below, where a complete clinical path (trace) of patient is shown for a chosen diagnosis (e.g. ischemic heart disease)

The patient below has multiple outpatient records (where the patient may or may not be diagnosed with ischemic heart disease) however for 3 of its inpatient visits (out of all inpatient visits) was diagnosed with ischemic heart disease. In order to get the outpatient records which may have relation to their inpatient visit, the following window is defined



- Get Admission date of first inpatient visit (INP-1). Go back by x days in the past (say 90 days)
- Get Discharge date of last inpatient visit (INP-3). Add x days to the future (say 90 days).
- Here INP-1 to INP-3 are inpatient visits where the patient was definitely diagnosed with Ischemic Heart Disease. The patient could have more inpatient visits where he/she may not be diagnosed with ischemic heart disease and hence is not considered here.

Further, all the outpatient records are collected between the above defined window (between two red lines)

H. Sequential Mining for Outpatient Visits

Once the outpatient records are collected, the sequential mining steps as show in section D are followed. Results of the sequential mining showed patients diagnosed with Ischemic Heart Disease have the following Outpatient procedures:

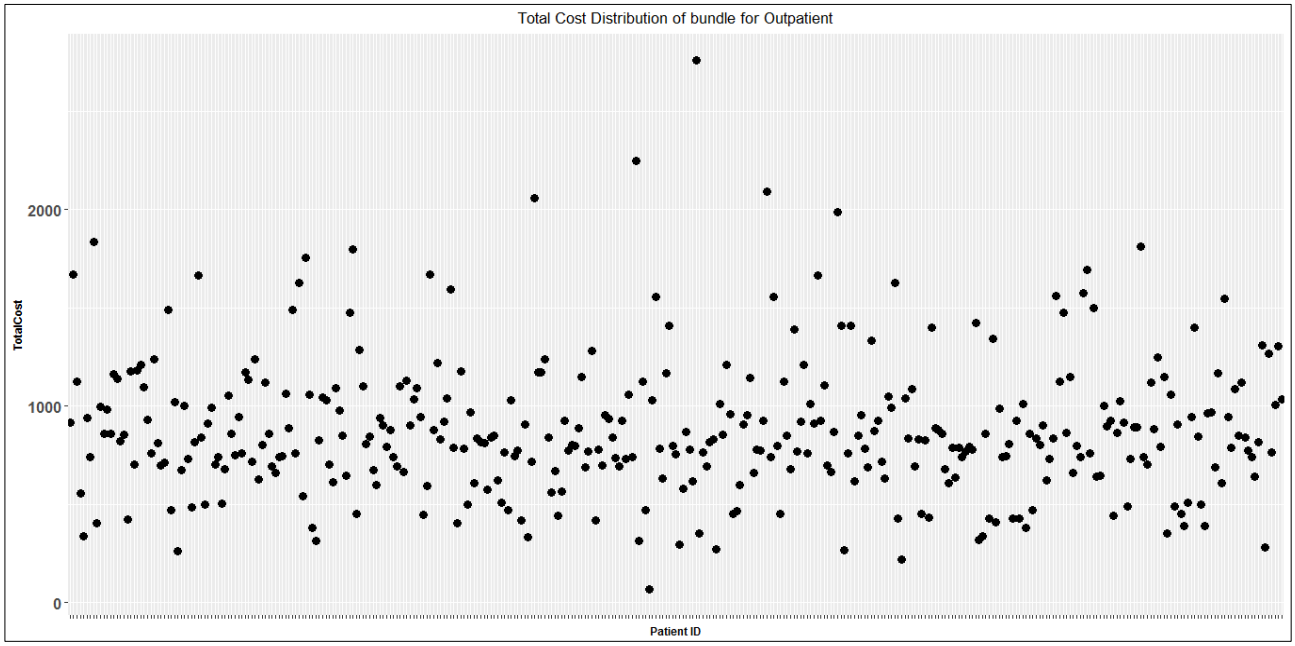
- Multiple Syringe, with or without needle, each (A4657) where in middle Dialysis Services and Procedures (90999) is performed on 10% of patients
- Multiple Syringe, with or without needle, each (A4657) where in middle Epoetin alfa, 100 units (for ESRD on dialysis) (Q4081) is performed on 10% of patients
- A4657 is occurring multiple times suggesting that patients have to undergo this injection multiple times

	rule	support	confidence	lift
28	<{A4657},{90999}> => <{A4657}>	0.1043505	0.9146667	6.4795460
31	<{A4657,Q4081},{A4657}> => <{A4657}>	0.1010040	0.9120879	6.4612779
22	<{Q4081},{A4657}> => <{A4657}>	0.1031336	0.9088472	6.4383205
27	<{A4657},{A4657}> => <{A4657}>	0.1098266	0.9070352	6.4254841
32	<{A4657},{90999,A4657}> => <{A4657}>	0.1010040	0.9021739	6.3910467
30	<{90999},{A4657}> => <{A4657}>	0.1013082	0.8951613	6.3413689
26	<{A4657,J2501}> => <{A4657}>	0.1073928	0.8914141	6.3148239
24	<{A4657},{Q4081}> => <{A4657}>	0.1043505	0.8909091	6.3112461

Outpatient Bundle for Ischemic Heart Disease Patient	**= Dialysis Services and Procedures (90999) (Procedure 1) + Syringe, with or without needle, each (A4657) (Procedure 2)**

The total cost distribution is as follows,

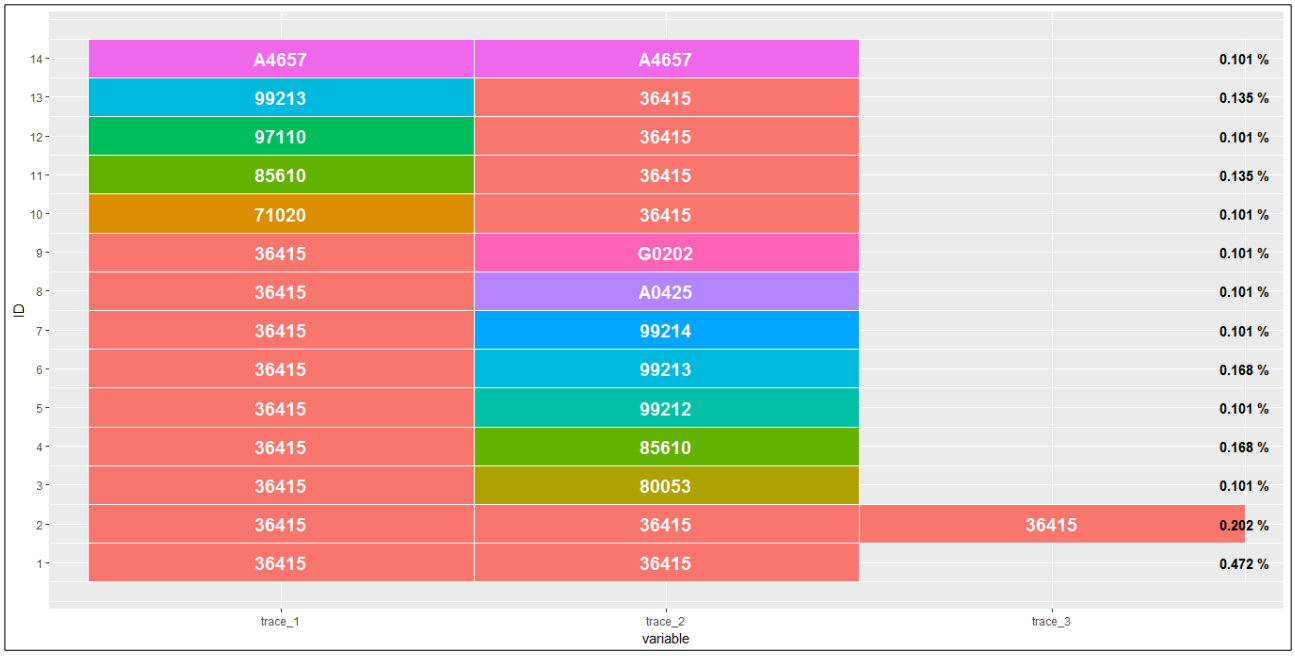
	DESYNPUF_ID	Cost1	Cost2	TotalCost
1	0019EF0547183BF0	424.0000	493.3333	917.3333
2	00291F39917544B1	393.3333	1276.6667	1670.0000
3	002E815B763CB020	316.0000	808.0000	1124.0000
4	009408B0FD1B047F	238.2857	320.0000	558.2857
5	0251A33476BA47D9	339.1111	0.0000	339.1111
6	02A88E1822A73372	280.0000	660.0000	940.0000
7	04AF914517BB6A32	348.5714	390.0000	738.5714
8	05D0C5BD6BF70BC3	345.7143	1490.0000	1835.7143
9	05FC0A83573A10BE	403.7500	0.0000	403.7500
10	07E45533FE2B5C86	497.5000	500.0000	997.5000



As shown above, the total cost does not exceed $2000 for Ischemic Heart Patients. Hence the threshold cost for Outpatient Bundle can be $2000 per patient.

The average cost of the bundle is $880 per patient and median cost is $830

Using process mining, the following trace plot is generated for frequent traces to check end-to-end traces of the patient and if bundles can be improved by adding or deleting any procedure. (Figure on right is relative frequency). The activities generated here confirm the sequential mining above as well.



In totality a bundle for Ischemic Heart Patients can be created:

Bundle for Ischemic Heart Patients = Inpatient Bundle + Outpatient Bundle

Cost of the Ischemic Heart Disease Bundle = Cost of Inpatient Bundle + Cost of Outpatient Bundle

Bundle	**Bundle for Ischemic Heart Patients = PTCA (Inpatient Procedure 1) + Biopsy of the mouth (Inpatient Procedure 2) + Dialysis Services and Procedures (90999) (Outpatient Procedure 1) + Syringe, with or without needle, each (A4657) (Procedure 2)**
Pricing the Bundle	**Threshold Cost of the Ischemic Heart Disease Bundle = $10000 + $2000 = $12000 per patient** **Mean Cost Bundle = $5400 + $880 = $6280** **Media Cost of the Bundle = $4900 + $830 = $5730**

3. Bundling Framework options

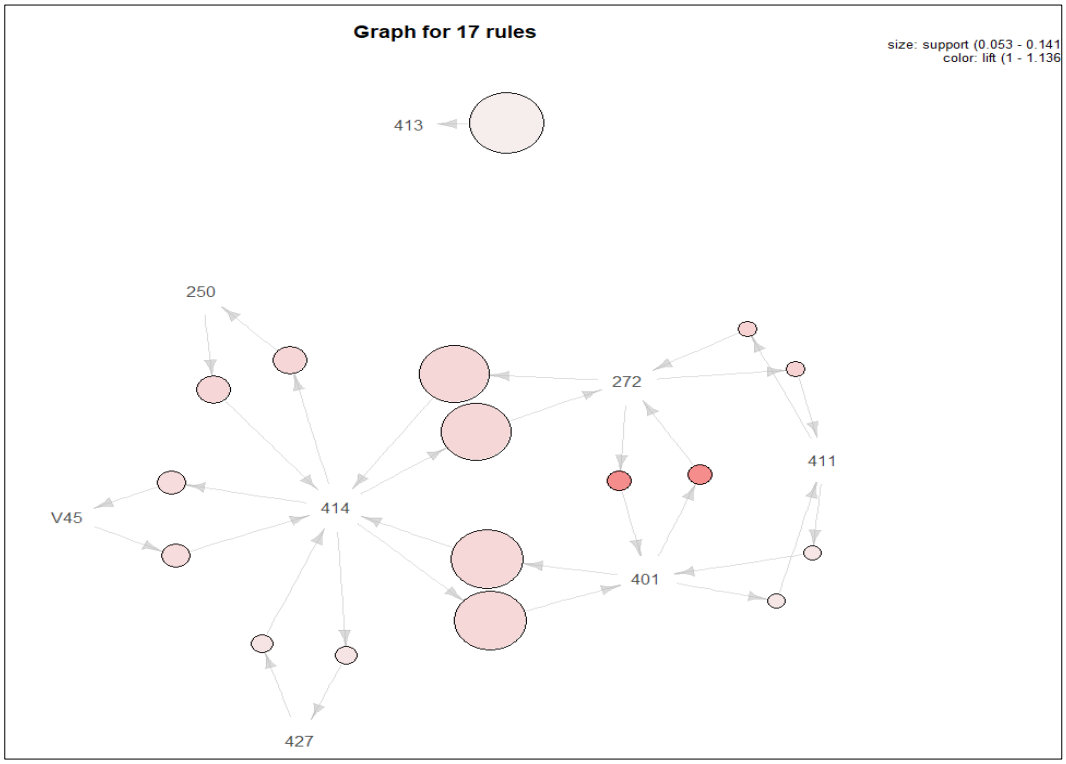
A. <u>Association mining with a primary diagnosis</u>

The above process of bundling could also be done in by taking a primary disease into account. For example, consider the primary disease ICD Diagnosis 1 to be Ischemic Heart Disease.
The steps remain the same:

A. Combining Inpatient and Outpatient records with necessary columns
B. Finding co-occurring Diagnosis using Association Mining with ICD Diagnosis 1 as "Ischemic Heart Disease"

After running association mining, 17 patterns were observed and could also be depicted graphically as shown below. The size of the bubble depicts the support. The bigger the bubble the larger the support. The shade of the colour shows the lift. The darker the bubble the better the lift.

- Ischemic Heart Disease (414) is co-occurring with Hypertensive Disease (401) and has a good support
- Ischemic Heart Disease (414) is also co-occurring with Diseases Of Other Endocrine Glands like Diabetes Mellitus(250)
- Ischemic Heart Disease (414) co-occurring with Hypertensive Disease (401) and Other Metabolic and Immunity Disorders (272) has a good lift.



C. The rest of the steps can be applied the same as Bundling Framework section to get the final bundle and price.

B. Sequential mining to find diagnoses occurring over a period of time

Apart from finding co-occurring diseases, it might also be helpful to find diseases that the patient might contract over the course of the medical journey. Say, a patient was diagnosed with "Ischemic Heart Disease". The diseases/ conditions leading to this diagnosis or afterwards is a useful insight. So using sequential mining it is possible to find the order in which diseases occurred.
Run Sequential Rule Mining on Inpatient records with a primary diagnosis ICD Diagnosis 1 as "Ischemic Heart Disease" and the other four diagnoses as one of the results from A, for example "Diseases Of Other Endocrine Glands". The following rules were formed:

rule	support	confidence	lift
<{414}> => <{250}>	0.3555777	0.6714132	0.6762241
<{V72}> => <{250}>	0.1025896	0.6560510	0.6607518
<{414}> => <{401}>	0.3387877	0.6397098	0.7698426
<{403}> => <{401}>	0.1323278	0.6019417	0.7243915
<{414},{427}> => <{401}>	0.1114115	0.5873968	0.7068879
<{414},{401}> => <{401}>	0.1984917	0.5858883	0.7050724
<{414},{401}> => <{250}>	0.1983495	0.5854683	0.5896634
<{401},{719}> => <{401}>	0.1129767	0.5851142	0.7041409
<{401},{414}> => <{250}>	0.1969266	0.5800503	0.5842066
<{401},{414}> => <{401}>	0.1967843	0.5796312	0.6975425

- It was observed in 35% of patients that Diseases of Other Endocrine Glands like Diabetes Mellitus (250) occurred after Ischemic Heart Disease (414).

This shows that there is a high possibility that an Ischemic Heart Disease patient might become Diabetic. This insight could help Medicare come up with add-on packages for the base Ischemic Heart Disease bundle. Since diseases never occur in silos, the combination of bundles and pricing is an option Medicare could explore.

4. Provider Cost Analysis

Medicare follows a prospective payment system (PPS) for reimbursing hospital operating cost. Most hospitals are paid a fixed amount, determined in advance for this operating cost according to one of the Diagnosis Related Groups (DRGs). The cost is based on number of discharges. A discharge is assigned to a DRG based on diagnosis, surgery, patient age, discharge destination, state and sex. Each DRG has a weight based primarily on Medicare billing and cost data, which differs between providers.

The inpatient files were analysed for years 2008-2010 to investigate how Medicare payments are varying across different providers for similar diagnosis.

1. Filter the Inpatient Files where the diagnosis code description is 'Ischemic Heart Disease'.
2. Filter the most expensive DRG code for the above Ischemic Heart patients. This is done to ensure the comparison is made between providers for the same DRG for the same episode(Ischemic Heart Disease)

3. Create a new dataset with relevant columns – Provider Number, Total Claim, Average Stay Per Patient, Total Patients and Per Day Hospitalization Cost for all the providers.

Calculated variables:

- Total Claim per Provider Reimbursed by Medicare= Sum (Claim Payment Amount)
- Average Stay per Patient per Provider = Total Hospital Stay /Number of Unique Patients
- Per Day Hospitalization Cost per Provider = Total Claim/Total Hospital Stay

Top 5 Providers with Highest per Day hospitalization cost

PRVDR_NUM	totalclaim	Totalhospitalstay	Totalpatients	AverageStayperPatient	PerDayHospitalisationCost
1002DC	44000	1	1	1	44000.0000
4507NJ	38000	1	1	1	38000.0000
0506CA	28000	1	1	1	28000.0000
2900UB	57000	4	1	4	14250.0000
49S0PN	12000	1	1	1	12000.0000

Bottom 5 Providers with lowest per day Hospitalization Cost

24008S	3000	4	1	4	750.0000
0505BP	3000	5	1	5	600.0000
1001NA	3000	5	1	5	600.0000
3901ZS	4000	10	1	10	400.0000
36008A	3000	14	1	14	214.2857

Insights

The above analysis gives us the best glimpse that:

⇒ *There exists a huge variation in the per day hospitalization cost for different providers for the same diagnosis ranging from $200 to $44000.*

⇒ *Even for one day stay, few providers are charging exorbitant amount of money. On the contrary, few providers are charging nominal rates even when the hospital stay is greater than 7 days.*

4. Calculate the average hospitalization cost and stay per patient across all providers to make the comparison.

Average Hospitalization Cost = $4000
Average Stay per Patient across = 5 days

5. Compare the individual provider cost with the average cost to differentiate the providers that charge more for a given DRG and episode compared to others.

The below table lists all such providers that charge less than or equal to the average cost.

PRVDR_NUM	totalclaim	Totalhospitalstay	Totalpatients	AverageStayperPatient	PerDayHospitalisationCost
3301QB	12000	3	1	3	4000.000
3400XN	10000	3	1	3	3333.333
3300YK	16000	5	1	5	3200.000
3301NV	6000	2	1	2	3000.000
07008H	3000	1	1	1	3000.000
0101MC	6000	3	1	3	2000.000
1100SK	2000	1	1	1	2000.000
1200RV	2000	1	1	1	2000.000
01S1YV	5000	3	1	3	1666.667
44006N	5000	3	1	3	1666.667
4900GV	5000	3	1	3	1666.667
4901US	5000	3	1	3	1666.667
5201NR	5000	3	1	3	1666.667
1002SK	6000	4	1	4	1500.000
3302QQ	3000	2	1	2	1500.000
14018H	4000	3	1	3	1333.333
5000YH	4000	3	1	3	1333.333
0504BP	3000	4	1	4	750.000
10S0SU	3000	4	1	4	750.000
24008S	3000	4	1	4	750.000
0505BP	3000	5	1	5	600.000
1001NA	3000	5	1	5	600.000

Insights

After running the above analysis completely, it could be concluded that:-

- ⇒ *Medicare needs to investigate the reason why certain providers are charging more for a given DRG and episode compared to others.*
- ⇒ *When creating bundles for Ischemic Heart Disease, Medicare could consider the above providers who charge less for the same quality of care and service as compared to others.*
- ⇒ *Medicare could set the average cost as the initial payment level to reflect the current cost of care for the episode. This would enable the providers to become more cost efficient at their end and deliver efficient quality of care.*
- ⇒ *The same analysis could be applied to any episode of care and could be improved further if more data points are available.*

5. Readmission Analysis

Once a bundle has been formed for a particular or set of diagnosis, one of the useful analysis would be to predict which patients would be most eligible for the bundle right from initial analysis. Early indication of a future diagnosis or procedure can help the hospital facilities to extend the bundles to the patients at the early stage of medical care. The following section describes how logistic regression model can be one of the methods which can be used in that, given a patient is suffering from a particular diagnosis and is undergoing a particular procedure today (which is part of the bundle), what is the probability that the same patient will return tomorrow to perform the second part of the bundle. Accurate prediction at the start can help the patient avail directly the bundle, right from the initial treatment. The following steps describe the complete process for building the model

- From the inpatient file, select the columns of interest such as diagnosis codes, procedure codes, claim payment amount, utilization count
- From this inpatient file, select only those records which are diagnosed with a diagnosis. In this case the filter used is for Ischemic Heart Disease
- Consider a bundle of procedures – as discussed in the above section, (2724, 66)
- Create a dataframe 1, where all the records have the procedure 2724. Create an additional column of readmitted as N
- Create dataframe 2, where all the records meet the above sequence of 2724 followed by 66. Create an additional column of readmitted as Y (as these are the records where the patient came back and was performed with 66 procedure code)
- From dataframe 1, remove all the IDs belonging in dataframe 2
- Create dataframe 3 – combining records from dataframe 1 and dataframe 2
- Combine all the beneficiary files of all years. Select columns of interest such as sex of the patient, race, age, renal disease (Y or N), chronic conditions (Y or N)
- Merge dataframe 3 with the beneficiary file above. The final columns used for model building are

```
> names(readmission.df)
 [1] "BENE_SEX_IDENT_CD"  "BENE_RACE_CD"      "BENE_ESRD_IND"     "SP_ALZHDMTA"       "SP_CHF"            "SP_CHRNKIDN"
 [7] "SP_CNCR"            "SP_COPD"           "SP_DEPRESSN"       "SP_DIABETES"       "SP_ISCHMCHT"       "SP_OSTEOPRS"
[13] "SP_RA_OA"           "SP_STRKETIA"       "AGE"               "CLM_PMT_AMT"       "CLM_UTLZTN_DAY_CNT" "Readmitted"
```

Base Model

The base model has accuracy of 0.965

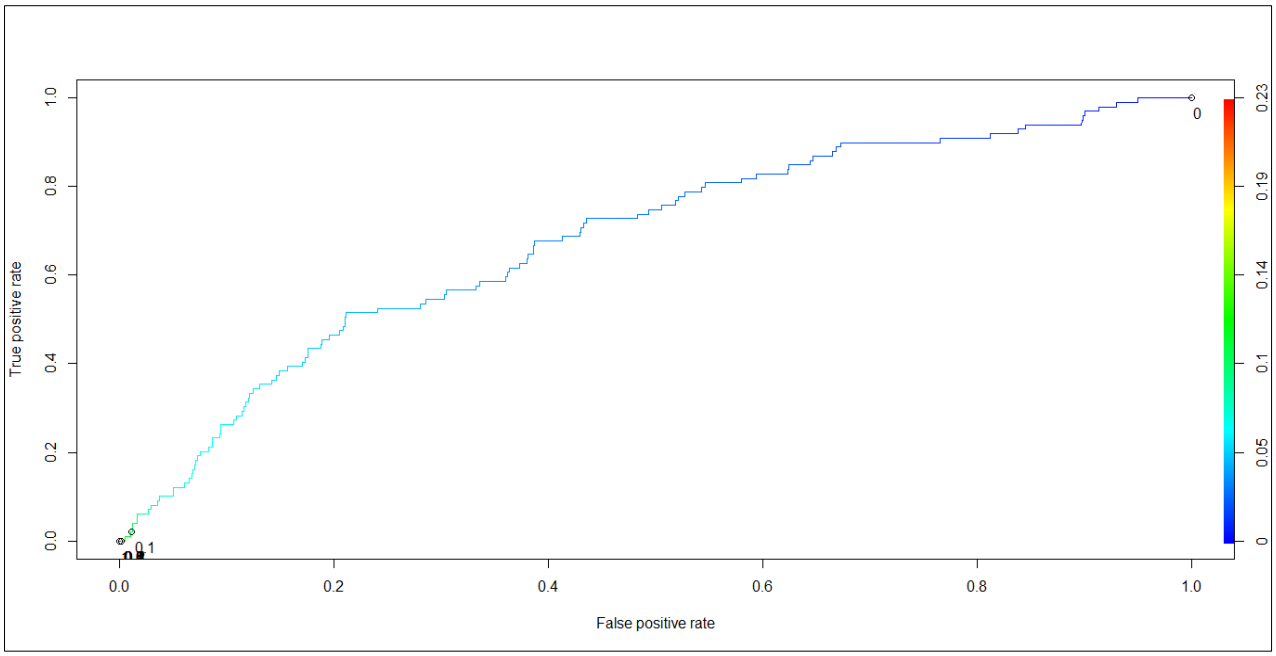
Logistic Regression Model

Logistic regression model is used to predict the probability of patient readmitting for the second procedure of the bundle created. If the probability is high of readmission, the patient can be provided with the bundle at the early stage of care.

The dataset is split into 80-20 ratio. The formula used to fit the model is

Readmitted ~ BENE_SEX_IDENT_CD + BENE_RACE_CD + BENE_ESRD_IND + SP_ALZHDMT+ SP_CHF + SP_CHRNKIDN + SP_CNCR + SP_COPD + SP_DEPRESSN + SP_DIABETES + SP_ISCHMCHT + SP_OSTEOPRS + SP_RA_OA + SP_STRKETIA + AGE + CLM_PMT_AMT + CL M_UTLZTN_DAY_CNT

The following plot shows the ROCR curve (calculated on Training data set)



From the above curve, the best threshold value lies around 0.2

With threshold as 0.17, the test data accuracy is **0.964**

The performance of the logistic regression model is very similar to the base model. Hence although the performance of the model is not high, there is a high probability of improving it with the addition of some more features.

The other model tried was randomForest, however the performance of this model was very poor

ISB

Conclusion

This report summarizes an approach for bundling which provides a more coordinated and integrated methodology for care delivery using analytical techniques. The Bundling Framework demonstrates how cost could be reduced for Medicare by combining procedures within the episode into a bundle, across inpatient and outpatient facilities. In addition, the framework shows the entire trajectory of patients undergoing different activities, which could indicate which areas might need more focus. Provider analysis highlights how Medicare could render the bundle with providers that are most cost-efficient. The process of exploratory analysis by itself could help derive many valuable insights which would be useful to Medicare. The implementation of the framework could prove to be beneficial to Medicare and can be extended to any episode of care or other datasets with similar attributes.

Future work

There is definitely a lot of scope for more experimentation and approaches in the future, given the complexities of the claims data and Healthcare sector.

1. Additional research and knowledge of the domain could further improve the framework prescribed.
2. The above bundled approach is limited to hospital based services which could further be extended to include other care settings or post-acute care.
3. The approach could be made more holistic, based on supplementary data containing information of patients, co-morbidities, Cost of individual services, carriers, prescription drugs and physicians.
4. Further, additional data in terms of geographic variation (say: population characteristics) in the episode of care and how these variations are affecting the application of episode
5. The long term goal should consider how to standardize payments across providers with increased coordination and efficiencies in delivery.
6. Bundling can be tried with combination of diseases.

Appendix

Logistic Regression (Contd...)

The log of the odds ratio (probability of success/ probability of failure) is the result of running logistic regression. If this log of odds is positive, then probability of the success is more than 50%.
Hence the accuracy of a logistic regression model is always filtered with results > 0.5.

Performance of Logistic Regression can be verified using:

Null Deviance and Residual Deviance: Lower the null deviance better the model. It takes into account the model with no other variables except the intercept. Residual deviance is adding all the variables and checking the model. Again this should be low.

Confusion Matrix: Displays the actual vs predicted values

		PREDICTED	
		GOOD	BAD
ACTUAL	GOOD	True Positive :d	False Negative :c
	BAD	False Positive :b	True Negative :a

(Source: Analytics Vidhya)

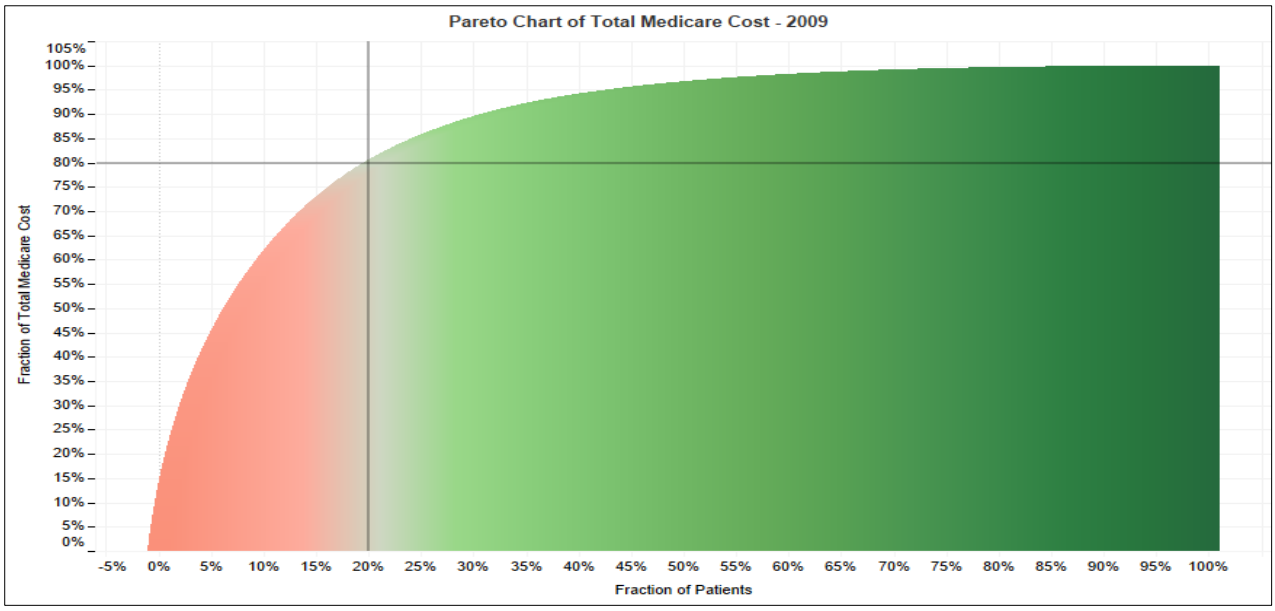
Accuracy of the model: (a + d) / (a +b+ c+ d)

Sensitivity is the proportion of true positives identified as meeting a condition viz. mammography test, showing proportion of patients with cancer who test positive.

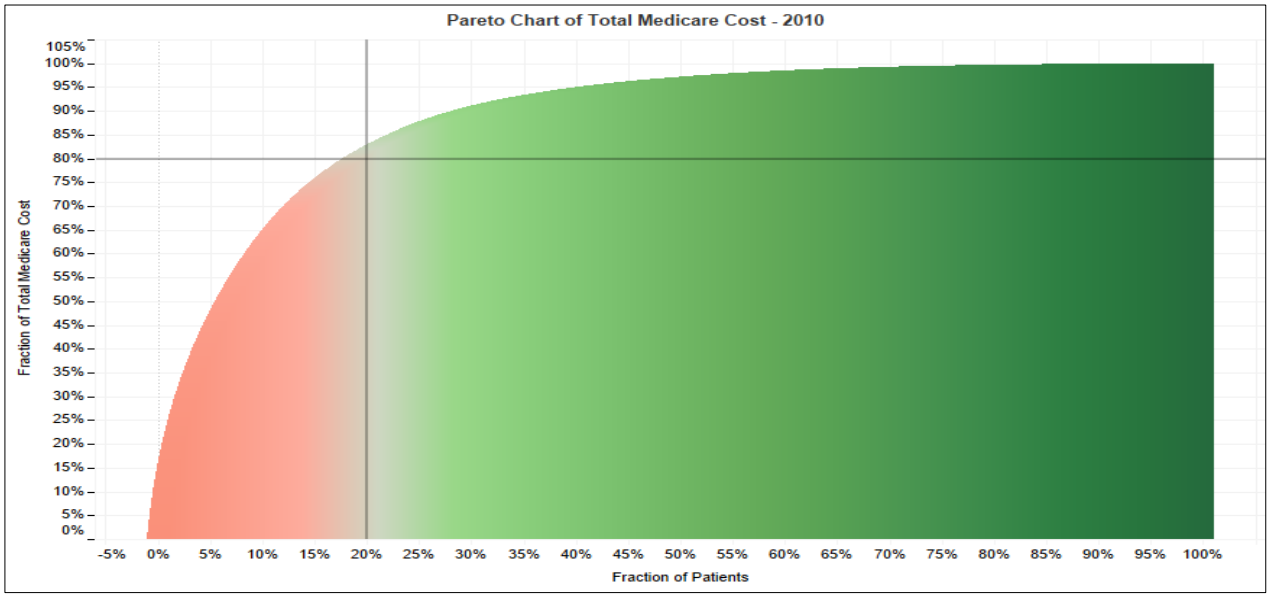
Specificity is the proportion of true negative identified as not meeting a condition viz. mammography test, showing proportion of patients who don't have cancer who test negative.
Area under the ROC (Receiver Operating Characteristic) curve: this is a graph showing the trade-off between false positive (1-specificity) on X axis and true positive (sensitivity or 1-false negative) on Y axis.
The curve closer to the Y axis and the top of the graph is more accurate.

Cost Bucketing (Contd…)

Pareto Chart of total cost for 2009



Pareto Chart of total cost for 2010



Insights	*20% or less contributed to 80% of the overall cost of all patients*

Chronic Condition and their distribution across different cost buckets (Contd…) (By Absolute count)

Distribution of Diseases

Disease1	Cost Bucket / Year1											
	CB_1			CB_2			CB_3			CB_4		
	2008	2009	2010	2008	2009	2010	2008	2009	2010	2008	2009	2010
SP_ISCHMCHT	1,43,976	1,82,428	1,26,151	23,883	30,938	17,155	15,001	16,591	7,679	10,913	9,975	4,038
SP_DIABETES	1,31,479	1,59,318	99,708	21,768	28,754	14,982	13,809	15,531	6,704	10,176	9,385	3,377
SP_CHF	91,172	1,26,015	85,752	20,080	25,276	13,637	13,257	13,844	6,105	9,909	8,649	3,233
SP_DEPRESSN	73,468	93,284	61,473	13,421	17,456	8,392	8,500	9,346	3,565	6,360	5,597	1,795
SP_ALZHDMTA	61,910	85,520	56,370	13,635	17,135	8,588	8,942	9,379	3,675	6,823	5,655	1,808
SP_CHRNKIDN	43,528	67,355	41,957	14,944	20,028	11,617	11,104	12,147	5,757	8,908	8,011	3,221
SP_OSTEOPRS	61,956	74,624	45,114	9,108	12,597	5,606	5,765	6,714	2,483	4,383	3,904	1,224
SP_RA_OA	52,808	66,896	33,425	9,583	12,481	4,577	6,712	7,032	2,108	4,764	4,000	988
SP_COPD	38,088	52,396	28,849	12,836	15,418	7,112	9,020	8,734	3,158	6,974	5,639	1,721
SP_CNCR	19,472	28,863	17,061	5,032	6,809	3,025	3,472	3,857	1,401	2,870	2,429	762
SP_STRKETIA	10,910	16,763	7,988	4,677	4,883	2,025	3,402	2,990	942	2,873	2,066	510

Insights

⇒ *Ischemic Heart Disease, Diabetes and Chronic Heart Failure are the chronic conditions found highest amongst the patients*

⇒ *Very high percent of people from cost bucket 2, 3 and 4 have the above mentioned chronic conditions as compared to patients from bucket 1.*

⇒ *For each bucket, the percentage of patients suffering from any condition increase from 2008 to 2009 but decrease from 2009 to 2010*

References

CMS. (n.d.). *CMS Data*. Retrieved from CMS: https://www.cms.gov/Research-Statistics-Data-and-Systems/Downloadable-Public-Use-Files/SynPUFs/DE_Syn_PUF.html.

Dimitris Bertsimas, A. K. (2017, June). *edx - MITx: 15.071x - The Analytics Edge*. Retrieved from www.edx.org: https://courses.edx.org/courses/course-v1:MITx+15.071x+2T2017/courseware/3372864201764d6d9f63931920e5152e/367a32fc4a7747fc8e9f5ca7ebd72999/?activate_block_id=block-v1%3AMITx%2B15.071x%2B2T2017%2Btype%40sequential%2Bblock%40367a32fc4a7747fc8e9f5ca7ebd72999

Jack O. Wasey, W. M. (2017, October 12). *Package 'icd'*. Retrieved from https://cran.r-project.org/: https://cran.r-project.org/web/packages/icd/icd.pdf

Michael E. Porter, R. S. (2016, July-August). How to Pay for Health Care. *Harvard Business Review*. Retrieved from https://hbr.org/2016/07/how-to-pay-for-health-care

Privé, F. (2017, September). *Split a vector into chunks such that sum of each chunk is approximately constant*. Retrieved from https://stackoverflow.com/: https://stackoverflow.com/questions/46431527/split-a-vector-into-chunks-such-that-sum-of-each-chunk-is-approximately-constant

ResDAC. (n.d.). *Data Entrepreneurs' Synthetic Public Use Files (DE-SynPUF)*. Retrieved from https://www.resdac.org/: https://www.resdac.org/cms-data/files/de-synpuf