Healthcare Claims Analysis

Josh Singer

Created for: [redacted]

Covered in this deck

- Background, Objectives & Methodology
- Patient Stratification
- Analysis, Takeaways & Next Steps

Background & Methodology

Background

The dataset provided contained 1M+ healthcare claim line items, where each row corresponded to one rendered service. Some of the relevant features (columns) in the dataset are included below:

Feature	Description
LINE_NO	Unique identifier corresponding to one rendered service
CLAIM_ID	Claim number corresponding to the billed claim containing the line item (service)
CLAIM_TYPE	Type of claim billed: Professional (PR) or Institutional (IN). PR charges are billed via CMS-1500 form, or 837-P, while IN charges are billed on a UB-04, or 837-I
PROV_ZIP	Provider zip code
MEMBER_ID	Identifier of the member (patient) for whom services were rendered
MEMBER_ZIP	Member zip code
PROV_SPECIALTY	Provider specialty
PLACE_OF_SERVICE	Place of service code. See full code set here.
PROC_CD	Procedure code, which identifies the service(s) rendered for a patient. Full list of CPT procedure codes here.
UB_BILL_TYPE	Three-digit alphanumeric code that provides three specific pieces of information (actually a 4-digit code, but CMS ignores a leading 0) - see code list here. - First Digit: Type of facility - Second Digit: Type of care - Third Digit: Sequence of this bill in this episode of care. Referred to as a "frequency" code
DRG_CODE	Diagnosis-related group (DRG) code classifies inpatient (hospital) cases according to certain groups that are expected to have similar hospital resource use (cost). See reference here.
DIAG1_CD	Diagnosis code, which identifies diseases, illnesses, injuries, and other reasons for patient encounters.. Full list of ICD-10-CM diagnosis codes here.
Eligible Charges	The billed charge, or maximum amount a plan will pay for the rendered service.

Objectives & Methodology

Purpose & Objectives:

- The goal of this project was to stratify patients by the quantity and value of healthcare services they received and identify trends within each patient group, for purposes of helping a primary care group and health benefits provider in strategic planning efforts (e.g. lowering their members' total cost of care).

Methodology:

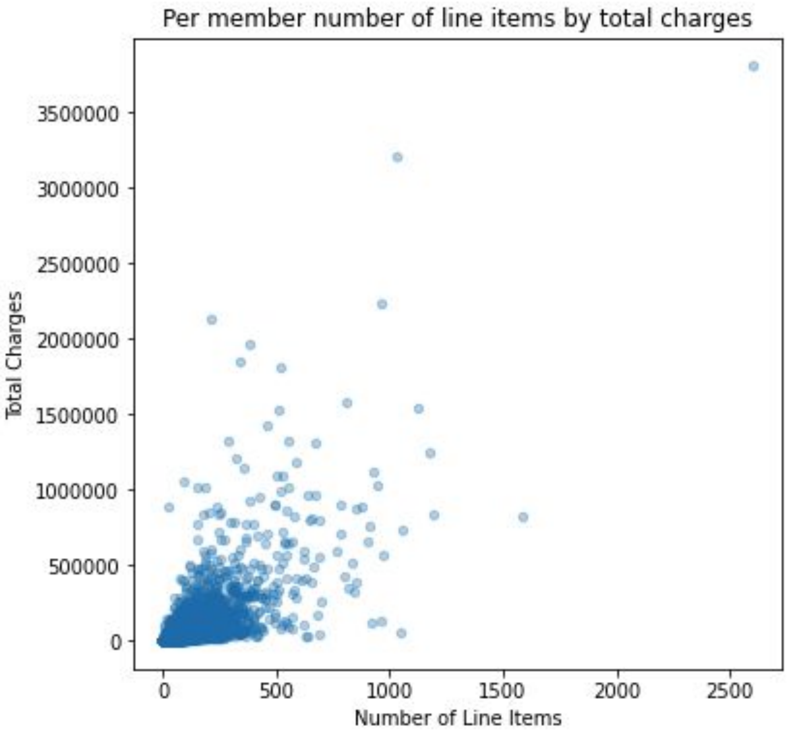
- Dataset contained 1M+ healthcare claim line items, where each row corresponded to one rendered service
- Dataset was received as an Excel spreadsheet, containing two sheets of data with similar features. Both sheets were combined to form one comprehensive dataset
- Stratified groups of patients by the quantity and value of services received using two methods:
    - K-means clustering, where patient clusters (groups) were determined by the machine learning algorithm
    - Manual stratification, where patient groups were determined manually
- Used the patient groups from stratification to identify trends within each group

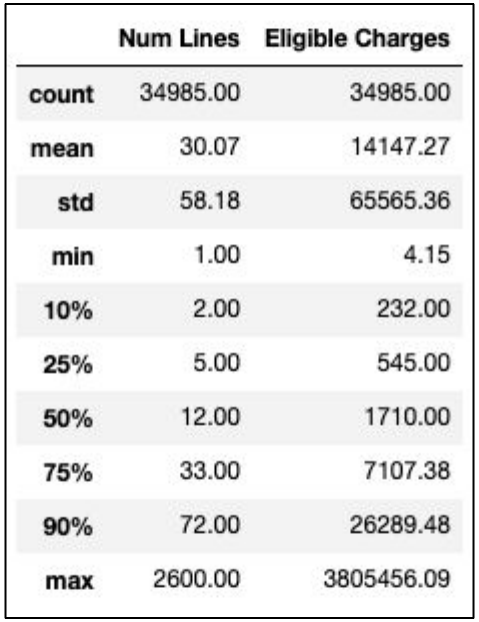
Assumptions & Caveats:

- Assumed Eligible Charges to be a good proxy for the actual cost of services
- Assumed the dataset did not contain duplicate charges, as we did not have a good method for de-duping line items
- No service dates were provided in the data

Patient Stratification

Understanding the Patient Population



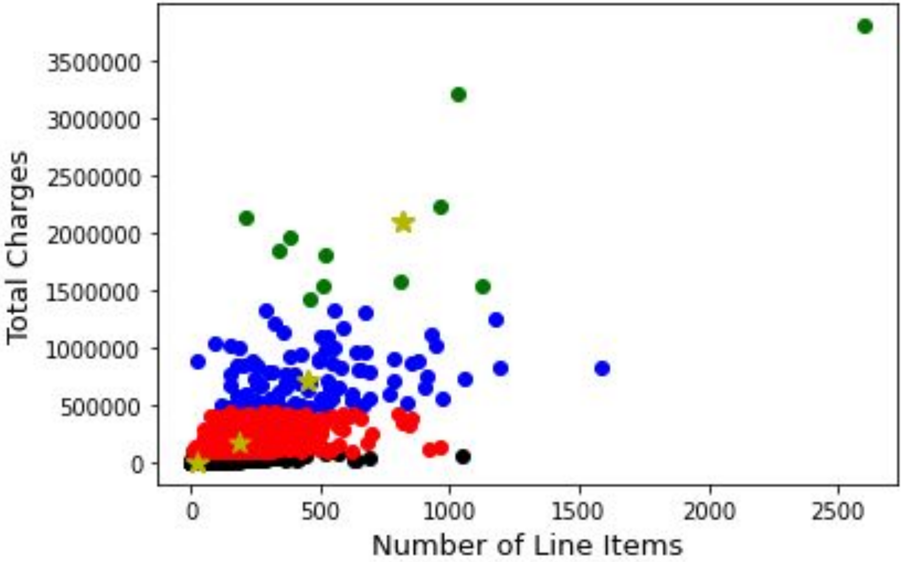


	Num Lines	Eligible Charges
count	34985.00	34985.00
mean	30.07	14147.27
std	58.18	65565.36
min	1.00	4.15
10%	2.00	232.00
25%	5.00	545.00
50%	12.00	1710.00
75%	33.00	7107.38
90%	72.00	26289.48
max	2600.00	3805456.09

- The average patient received 12 services for a total cost of $1,710
- 90% of patients accumulated less than 72 services and $26,289 in charges

Method 1: K-Means Clustering

After training and comparing multiple K-means algorithms and given the healthcare-specific context of interest, we settled on using a K-means model with 4 clusters. The image below shows each cluster output by the model, represented by a different color and with a star marking its centroid (the mean of all the data points in that cluster).



The clusters generated by the k-means model don't appear to align with our expectations of the following groups of patients:

- Low cost, low utilizers
- Low cost, high utilizers
- High cost, low utilizers
- High cost, high utilizers

So instead, we'll manually stratify our patient population by setting bounds we think are more appropriate.

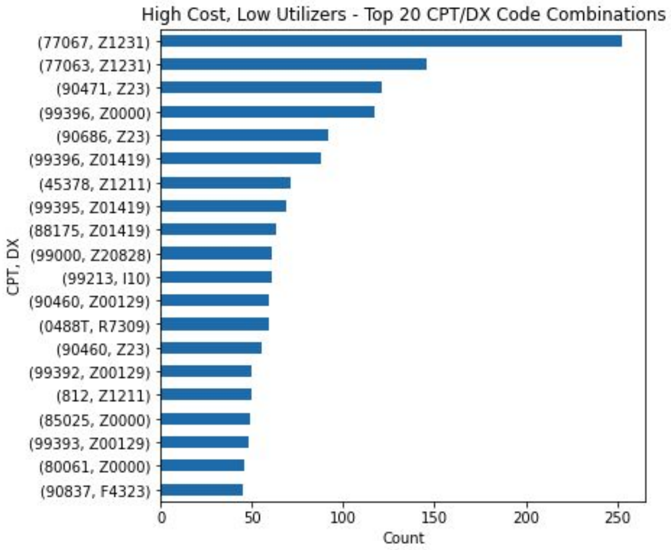
Method 2: Manual Stratification

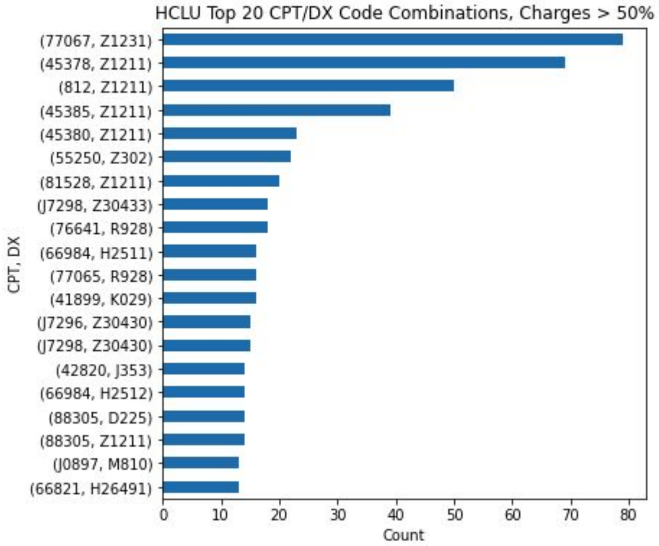
Defining “low” and “high” to be less than and greater than average, we’ll set the bounds for our patient groups at the 50 percentile, where number of services rendered was 12 and total cost of services was $1,710.

Group	Description	Criteria	# (%) of Total Patients	# (%) of Total Services	# (%) of Total Charges
1	Low cost, low utilizers	● # services < 12 ● Charges < $1,710	15,803 (45.17%)	77,203 (7.3%)	$9.2M (1.9%)
2	High cost, low utilizers	● # services < 12 ● Charges > $1,710	1,873 (5.35%)	17,089 (1.6%)	$7.7M (1.6%)
3	Low cost, high-utilizers	● # services > 12 ● Charges < $1,710	1,692 (4.84%)	28,660 (2.7%)	$2.2M (0.4%)
4	High cost, high-utilizers	● # services > 12 ● Charges > $1,710	15,617 (44.64%)	929,198 (88.3%)	$475.9M (96.2%)

Analysis & Takeaways

Top CPT/DX Code Combinations - Group 2 (High Cost, Low Utilizers)



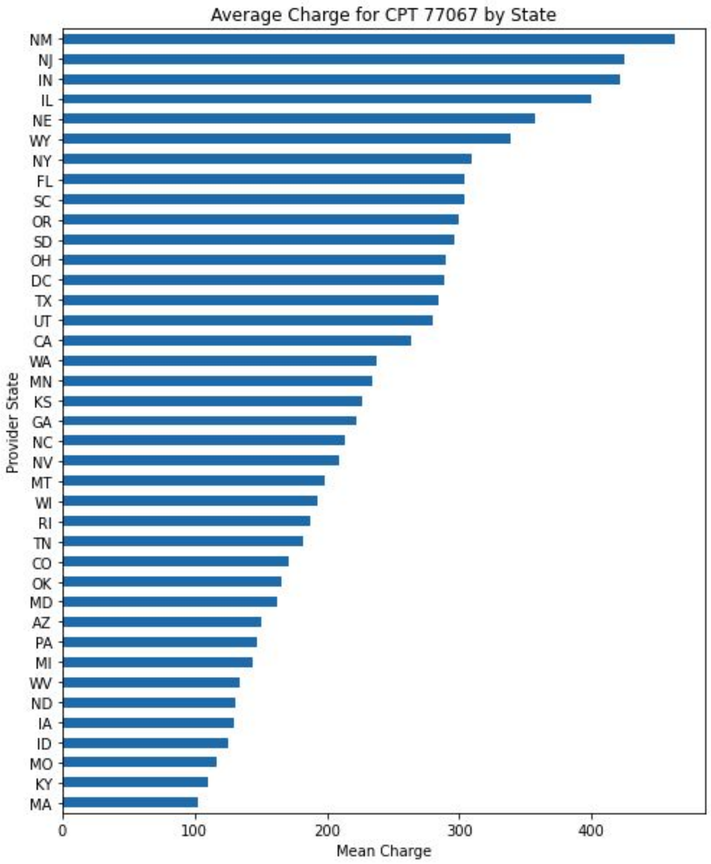


The top two CPT/Dx combinations are both related to mammography screening, and combined, are more than three times as common as all other code combinations.

- CPT 77067: Screening mammography, bilateral
- CPT 77063: Screening Digital Breast Tomosynthesis, bilateral
- Dx Z1231: Encounter for screening mammogram for malignant neoplasm of breast

If we filter the top code combinations for the highest cost services (those with above average charges) we see that mammography screening remains at the top. In other words, mammography screening is the most common high cost service for patients in the high cost, low utilizer group.

Group 2 (High Cost, Low Utilizers) Deep Dive: CPT 77067

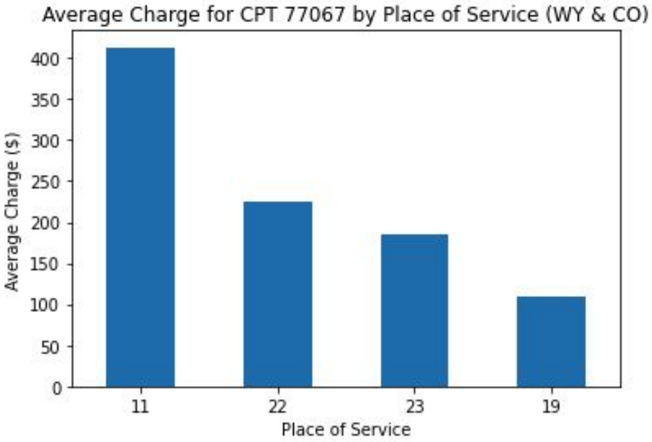


Interestingly, it appears there is high variability with charges associated with CPT 77067, as charges in the dataset range from $38.75 to $1,337.51.

Among the variables available in the data to analyze, it appears that there is a strong correlation between location and cost for mammography screening.

For example, to the left we see more than a four-fold increase in average cost for mammography screening when the service was rendered in New Mexico versus Massachusetts.

Group 2 (High Cost, Low Utilizers) Deep Dive: CPT 77067



Place of Service Code	Description
11	Office
22	On Campus-Outpatient Hospital
23	Emergency Room – Hospital
19	Off Campus-Outpatient Hospital

Top 4 Highest Cost Providers (WY & CO)	Avg Charge
Removed due to sensitive information	$1,257.61
	$1,132.96
	$1,070.89
	$1,060.90

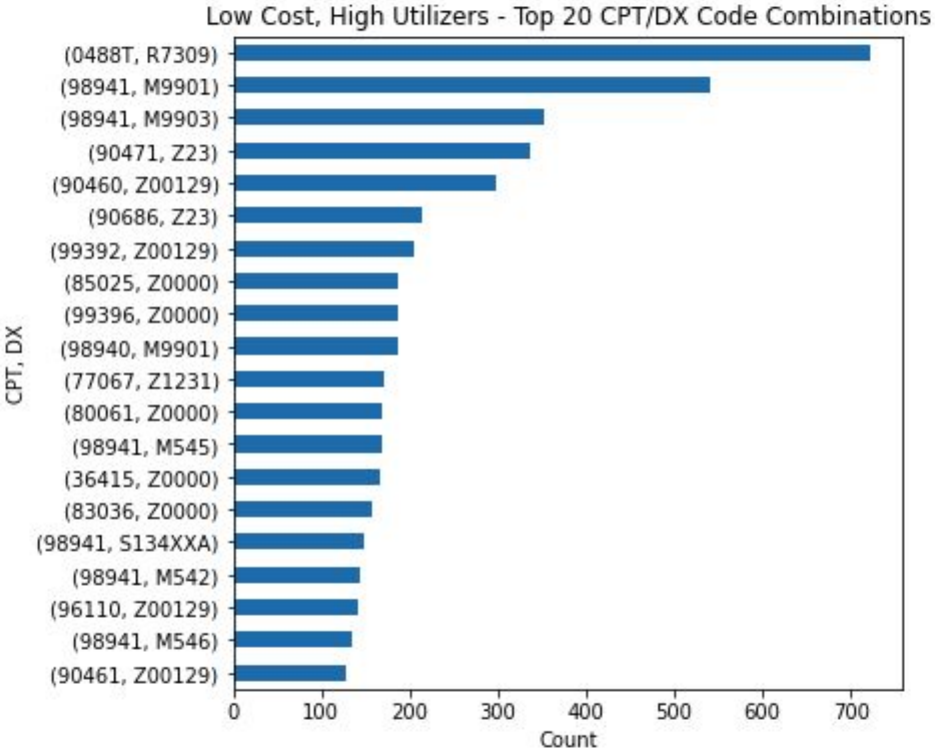
Top 4 Lowest Cost Providers (WY & CO)	Avg Charge
Removed due to sensitive information	$94.43
	$78.50
	$70.00
	$53.00

Further, it appears that provider and place of service greatly influence the cost of mammography screening.

Recommendations / Next Steps:

- Further investigation into the high variability of charges associated with mammography screening, as well as whether all of these screenings are medically necessary (is there an opportunity to reduce the number of member screenings to lower total cost of care)?
- Steer patients to lower cost providers and locations when possible

Top CPT/DX Code Combinations - Group 3 (Low Cost, High Utilizers)



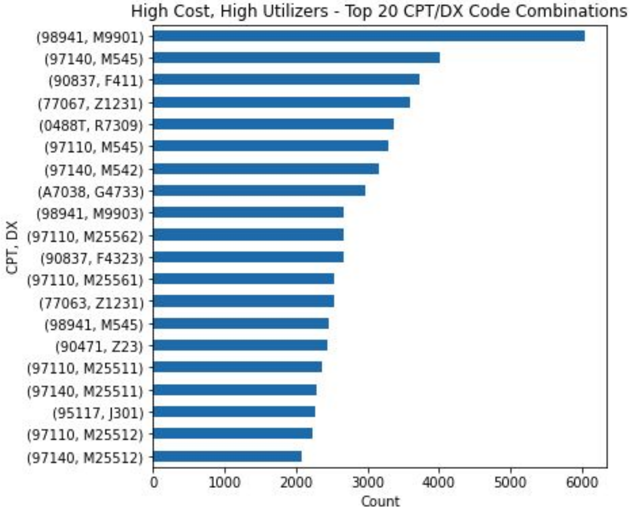
The top CPT/Dx code combination for patients in the low cost, high utilizer group is related to diabetes treatment.

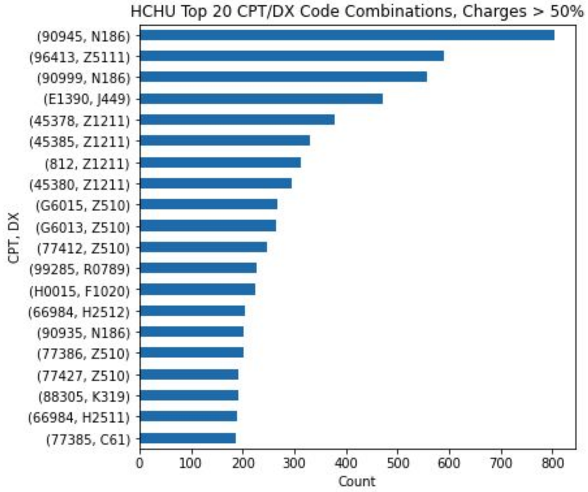
- CPT 0488T: Preventive behavior change, online/electronic structured intensive program of prevention of diabetes using a standardized diabetes prevention program curriculum, provided to an individual, per 30 days
- Dx R7309: Other abnormal glucose

Recommendations / Next Steps:

- Since the progression of diabetes can lead to costly downstream services, it may be worth investigating this patient populations’ current utilization and methods of diabetes care and prevention to identify opportunities to improve outcomes and lower total cost of care. For example, is there an opportunity to leverage new technologies/services in the space that have shown positive outcomes in diabetes treatment (e.g. Virta Health)?

Top CPT/DX Code Combinations - Group 4 (High Cost, High Utilizers)





The top CPT/Dx combination is related to back issues/treatment.

- CPT 98941: Chiropractic manipulative treatment
- Dx M9901: Segmental and somatic dysfunction of cervical region

Recommendations / Next Steps:

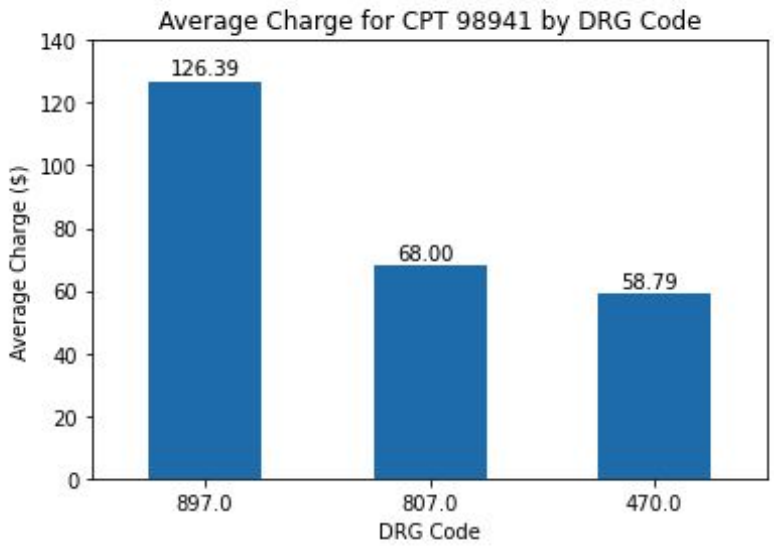
- Are there ways we can help patients proactively address back issues to prevent further complications and high-cost services (e.g. surgery)?
- Are there lower cost alternatives for these services (e.g. tele-rehabilitation v. in-person care)?

If we filter the top code combinations for the highest cost services (those with above average charges), we see that dialysis services for ESRD is the most common high cost service.

Recommendations / Next Steps:

- Further investigation into this patient population's current methods of ESRD management and opportunities for improved management/cost savings (e.g. is there an opportunity to move patients to home dialysis rather than in-person)?
- Further investigation of other top code combinations.

Group 4 (High Cost, High Utilizers) Deep Dive: CPT 98941



DRG Code	Description
897	Alcohol, drug abuse or dependence without rehabilitation therapy without major complication or comorbidity
807	Vaginal delivery without sterilization or d&c without cc/mcc
470	Major Joint Replacements or Reattachment of Lower Extremity

It appears there is high variability with charges associated with CPT 98941, as charges in the dataset range from $5.00 to $989.41. Much of the variability appears to be a result of the DRG code associated with the procedure.