Medicaid Prescriptions and Drug Manufacturer Payments

Introduction

Drug and device manufacturers make payments to health providers. Usually in the form of catered lunches, but they also pay for travel, consulting, and some health providers earn royalties from drug and device manufacturers. In 2014, drug and device manufacturers paid more than $2.6 billion dollars to healthcare providers. But do these payments have any affect on the rate at which providers are prescribing their drugs? To find out, two models will be developed on prescription data, one with information only about the specialty of the provider and the name of the drug that was prescribed, the other with additional information about payments that provider received. If the model with payments data has more predictive power then that is evidence of a potential influence. First we'll look at the data, then how it was shaped, finally what modeling was used and the results.

Data Exploration

The payments are reported to the Center for Medicare and Medicaid Service's Open Payments program. Drug and device manufacturers are required by law to report any payments made to healthcare providers and teaching hospitals to this transparency program, and the CMS releases this information to the public. This data is released yearly and contains a row for every payment made during that release year. The important fields are the physician ID, which uniquely identifies each healthcare provider that receives a payment, the total payment amount, and the manufacturer that is making the payment. Here is a snapshot of what that looks like:

	Physician_Profile_ID	Total_Amount_of_Payment_USDollars	Applicable_Manufacturer_or_Applicable_GPO_Making_Payment_Name
0	269814	17.14	FOREST PHARMACEUTICALS, INC.
1	180504	11.27	FOREST PHARMACEUTICALS, INC.
2	180504	2.20	FOREST PHARMACEUTICALS, INC.
3	86548	47.70	FOREST PHARMACEUTICALS, INC.
4	307584	12.01	FOREST PHARMACEUTICALS, INC.

There are many other fields of interest in this dataset, including the nature and form of the payments issued, but for the purposes of this study these are the three that will be most focused on.

The CMS also tracks every prescription claim filed on behalf of those insured under Medicare or Medicaid. Particularly though, the Center for Medicare and Medicaid Services publicly releases the claims information for those insured under the Medicaid Part D program. This data is released yearly and contains a row for every provider-drug pair. For each of these pairs, the CMS releases the number of claims that were filed from that provider for that drug, and the total day supply of drug that provider issued across all of the claims. As with the payments dataset, there are many other interesting fields, but the ones that will be focused on are the national provider identifier, which uniquely identifies each provider, the total day supply, drug name, and the specialty of the provider. Here is a snapshot of what that data looks like:

	NPI	SPECIALTY_DESC	DRUG_NAME	TOTAL_DAY_SUPPLY
0	1821285826	Urology	TAMSULOSIN HCL	360
1	1093969024	Internal Medicine	PANTOPRAZOLE SODIUM	360
2	1518048750	Pediatric Medicine	VENLAFAXINE HCL ER	360
3	1952310666	Psychiatry	ABILIFY	420
4	1952310666	Psychiatry	ALENDRONATE SODIUM	480

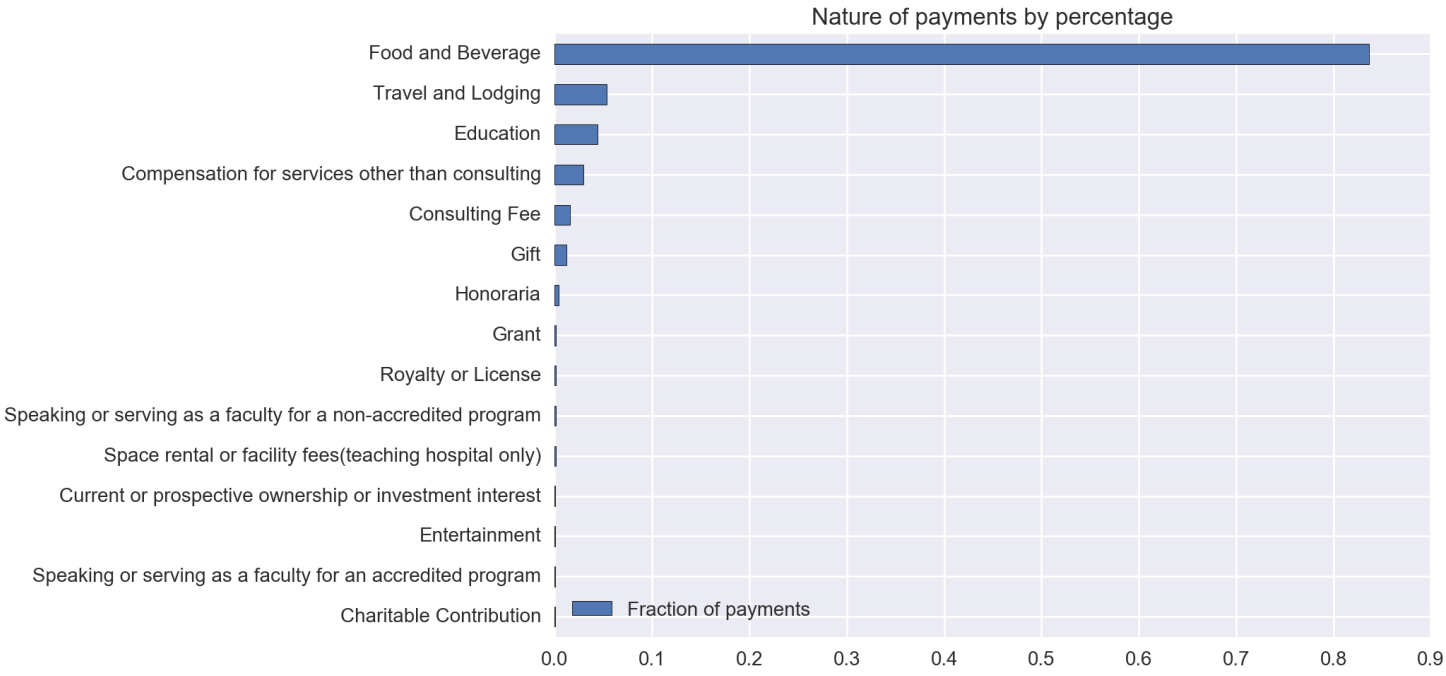
The two datasets used are from the 2014 program year for both the Medicaid Part D data as well as the payments. The prescription data consists of over 23 million provider-drug pairs from over 800,000 providers and more than 2,700 different drugs.

The payments dataset from 2014 consists of over 11 million payments made to more than 600,000 physicians from over 1,500 different manufacturers.

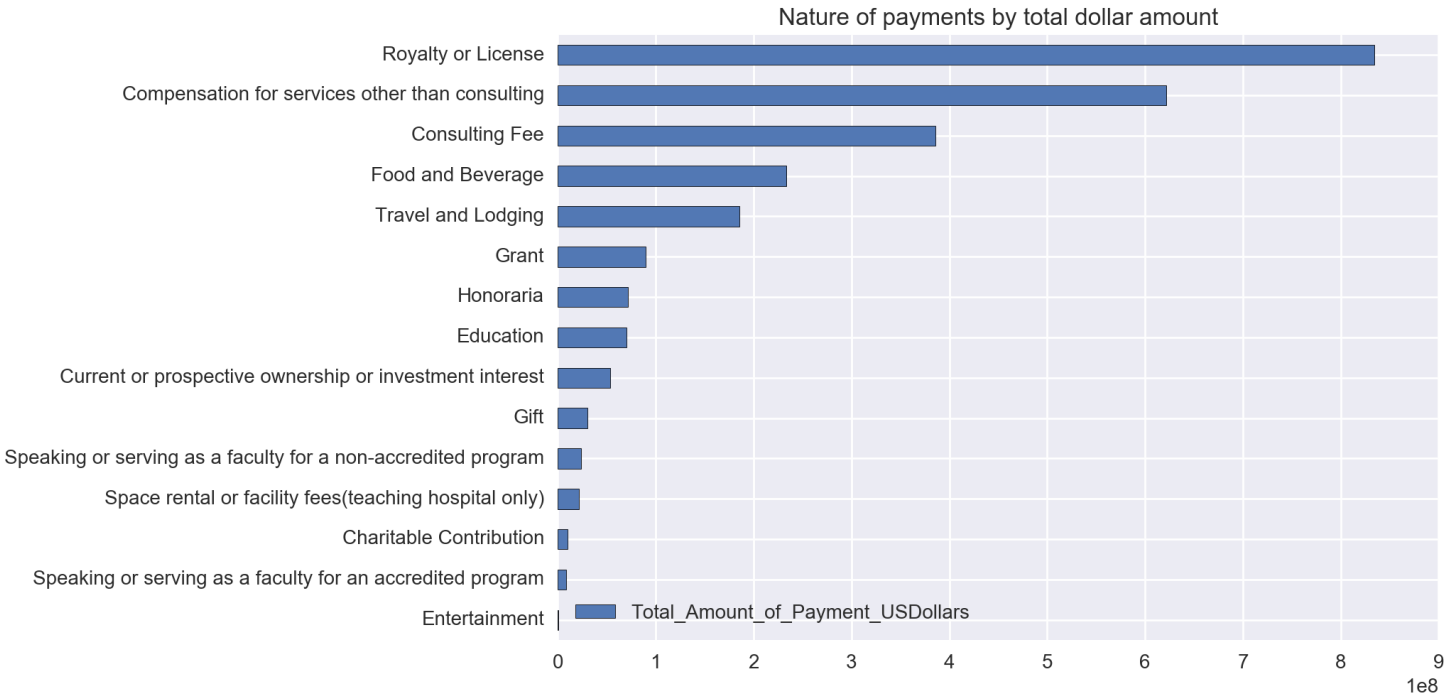
Now let’s take a deeper dive into each of these datasets, starting with the payments.

Payments

One of the interesting fields in the dataset mentioned was the nature of the payments. As was stated, the most common nature of payment was catered lunches. Below is a full breakdown of the relative frequency of each nature of payment from the 2014 payments dataset. All other natures of payment pale in comparison to Food and Beverage.

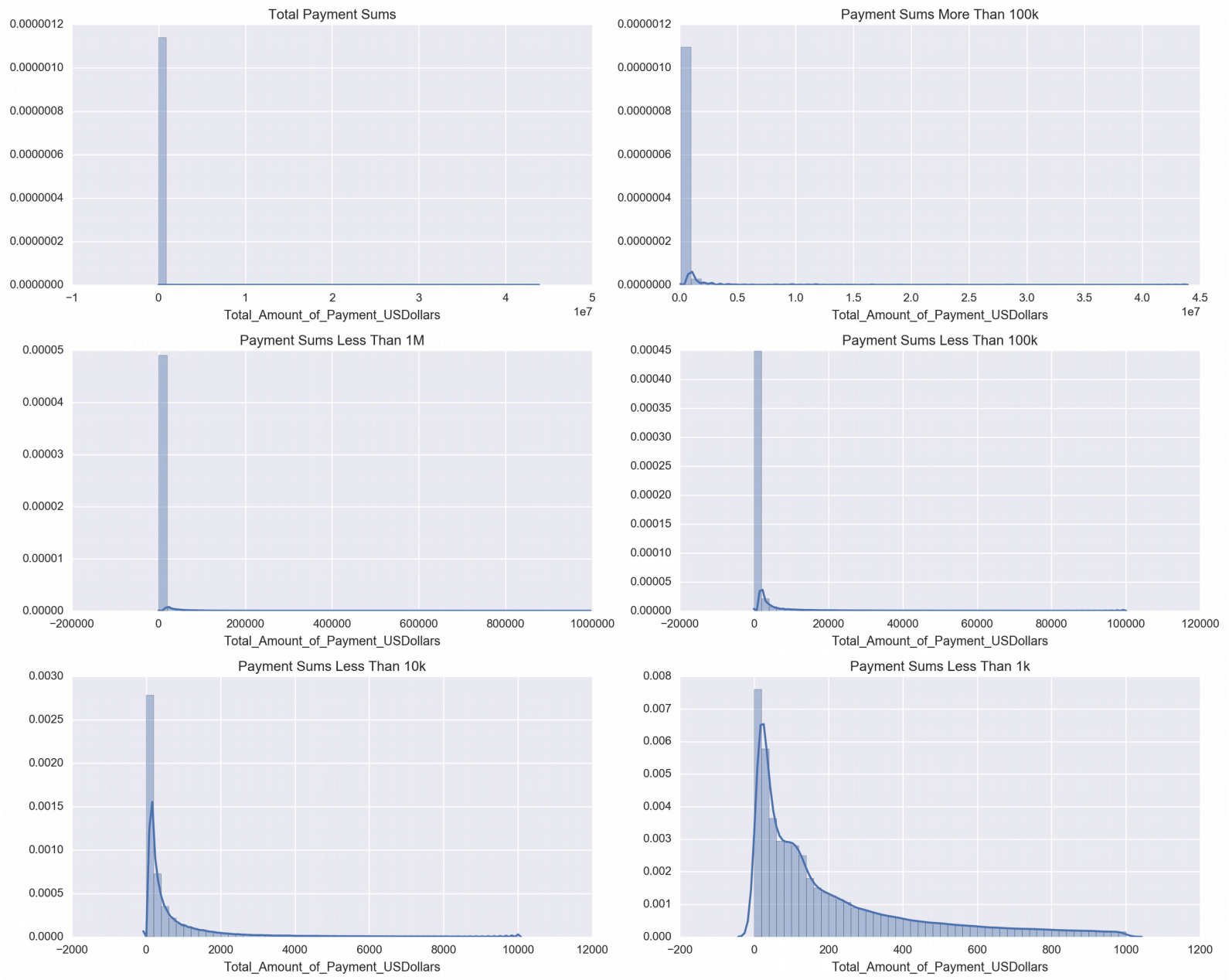


But how do these payments stack up in terms of the total dollar value?



So we can see that by dollar value, royalties and consulting pay out the most to healthcare providers. In general, most payments are of small dollar value for food and beverage expenses, but some healthcare providers receive large sums of money for royalties and consulting. This is also exhibited when we look at the distribution of the total sum of money each physician received over the 2014 year. This distribution is highly tailed to the right, which is to say most physicians received less

than $800 from drug and device manufacturers in 2014, but some received millions of dollars. To visualize this distribution it has been broken down into six categories. The first shows the whole distribution, the second, only those that received payments totaling more than $100,000 from drug and device manufacturers in 2014, and the others show those who received payments of a continually smaller sum of money down to those who received strictly less than $1,000 in 2014.



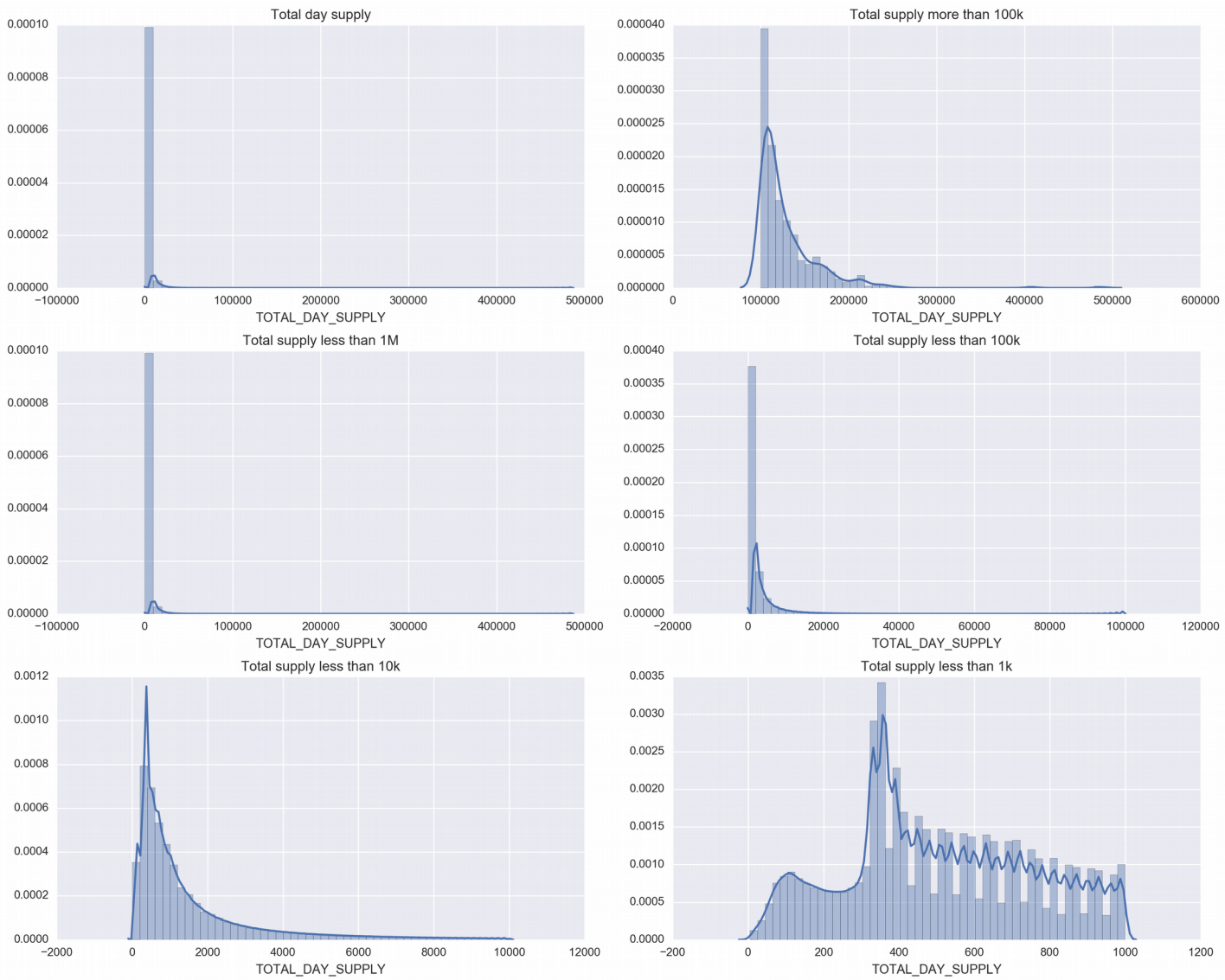
These charts tell the same story, most physicians took in small dollar amounts less than $1,000 in 2014 from drug and device manufacturers, and some made quite a lot. In fact, the physician who received the most in 2014 earned over $4 million dollars. There were 118 physicians who earned over a million.

Prescriptions

The field of interest in the prescriptions dataset is the total day supply. There is another field, total claim count, that can be used to measure the rate of prescription, but intuitively, the total day supply offers a better picture of just how much of a certain drug a provider is prescribing. Also, the total day supply of a drug is under more discretion to the prescriber than a single claim. When a healthcare provider prescribes a drug, one claim is always filed to that patient's insurer, but the number of day's supply is determined by the physician and will vary from patient to patient and drug to drug.

What is important to consider about this dataset is that it consists only of claims filed for patients who are insured by Medicaid Part D. Most healthcare providers serve more than just those who are insured under Medicaid Part D so this often will not provide a picture of a provider's full practice and patient population. Medicaid Part D generally serves an older patient population with different medical problems than other patient populations.

Total day supply is distributed my like the payment sums are. It is highly tailed to the right, with a large concentration less than 5,000. This distribution was broken down into multiple categories just like the payment sums were to illustrate this.



So this tells a story of how physicians prescribe drugs to those insured by Medicaid Part D. Which is that most providers prescribed any particular drug in volumes less than a total 1,000 day supply in 2014. However, some providers prescribed total day supplies of some drugs to Medicaid Part D patients in volumes exceeding 100,000.

Here’s a snapshot of the top provider-drug pairs by total day supply:

	NPPES_PROVIDER_FIRST_NAME	DRUG_NAME	SPECIALTY_DESCRIPTION	TOTAL_CLAIM_COUNT	TOTAL_DAY_SUPPLY
9346913	MARC	WARFARIN SODIUM	Internal Medicine	11604	487092
23426079	WILLIAM	WARFARIN SODIUM	Cardiology	6029	407617
8110871	KATHLEEN	WARFARIN SODIUM	Internal Medicine	3256	256530
11824087	SHIU	NAPHAZOLINE HCL	Ophthalmology	7215	243610
898054	ANALISA	LUMIGAN	Ophthalmology	8938	238990
7271784	BUTCHAIAH	HYDROCODONE-ACETAMINOPHEN	Physical Medicine and Rehabilitation	7979	237610
6014442	DARLENE	WARFARIN SODIUM	Pulmonary Disease	3468	233009
17944358	EDUARDO	CLONAZEPAM	Psychiatry	7675	227163
21340383	VIKRAM	WARFARIN SODIUM	Internal Medicine	4515	218895
23176031	RICHARD	WARFARIN SODIUM	Internal Medicine	4908	217042
11235486	KAREN	ALENDRONATE SODIUM	Physician Assistant	2552	216213
13269379	LAXMAIAH	HYDROCODONE-ACETAMINOPHEN	Interventional Pain Management	7218	216212
4796571	MICHAEL	WARFARIN SODIUM	Cardiology	6544	215266
17480486	ANDREW	LATANOPROST	Ophthalmology	3331	212622
18198153	JESUS	CLONAZEPAM	Neuropsychiatry	7240	211539

There appears to be some medications which are generally prescribed in larger amounts, particularly Warfarin, which is a blood thinner commonly prescribed to treat many heart conditions. It is likely that these providers serve patient populations that are much more likely to have these conditions and use Medicaid Part D.

Prediction

With this data we are going to build a model to predict the total day supply of a provider-drug pair given information on that provider's specialty and the name of the drug. Also we are going to build a model to predict the same output given the same information with additional information on the payments those providers received during that year. If the latter model proves to be more predictive then this is evidence that the payments a provider receives influences the total day supply of the drugs they prescribe. If this is true, then we can use the coefficients of the model to understand which drug manufacturers payments most influence these rates. But first we have to shape the data into a form we can use for prediction.

One problem the reader may have already noticed is that these two datasets do not have a common identifier. This is not the fault of the Center for Medicaid and Medicare services, but is actually dictated by the law. Legally, the CMS keeps track of the National Provider Identifier for each

provider in the open payments dataset, but also legally, they are not allowed to release that identifier to the public. However, they do release other identifying information about each of those providers, including their first, middle, and last names, as well as their business addresses at the time of reporting. This information can be used to look up a providers NPI with the CMS's National Plan and Provider Enumeration System. This entire database is publicly available and it was used to map the physician profile id from the payments dataset to the national provider identifier from the prescriptions dataset.

Unfortunately, this process was not perfect as many providers do not keep their information updated or have moved between reporting times. Also, there are many providers who do not receive any payments from drug and device manufacturers despite serving Medicaid Part D patients. However, a mapping was found for ~36% of the prescription dataset. This amounts to more than 8.7 million provider-drug pairs that can be used for prediction.

The next remaining hurdle was to shape the data for prediction. To do this, the payments dataset was grouped by each physician and manufacturer then summed. What this amounted to was a row for every physician with columns for each manufacturer, where the value under each manufacturer indicated the dollar amount that physician was paid by that manufacturer during 2014. Most of these values were zero. Here is a snapshot of what that looks like:

ABIOMED	ABL Medical, LLC	...	iCAD, Inc	iRhythm Technologies, Inc.	iScreen Vision Inc.	integrated dental systems	nContact Surgical, Inc	optos plc	rEVO Biologics, Inc.	sanofi-aventis U.S. LLC	Physician_Profile_ID	NPI
0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	100000	1215900089
0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1000001	1356429617
0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1000014	1366438970
0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	100002	1215928759
0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	100003	1215928916

This data was then joined to the prescriptions data set on the NPI field. What this produces is a dataset that has a row for every provider-drug pair, with the data on drug manufacturer payments attached. What was kept from the prescription dataset were the specialty and drug name fields. These fields were one-hot encoded so there was a column for each possible specialty and drug name with a value of 1 for the drug and specialty for that row, and every other drug and specialty column was zero.

The final product of this transformation was a sparse data matrix that was 8,743,650 x 4122.

The next question was what type of model to use for this prediction. Regression is obviously the task and after trying a couple of linear models it was clear that a linear model would not predict the total day supply from any of these features with any amount of accuracy. What was settled on was a random forest modeling technique. Intuitively, a decision tree makes sense for this type of data, the data consists of many categories and a large number of numerical features, the hypothesis is that splitting on some of these features will garner more information than others in predicting the output variable. The random forest regressor also offers a handy ranking of feature importances, so we can interpret how important each feature was in the regression.

After some cross-validation on a small percentage of the dataset (using large percentages of this dataset was difficult because of resource limitations), it was determined that a forest of 140 trees that were fully grown and made splitting decisions by looking at all of the features gave the best testing accuracy. This accuracy was determined using R squared as a measurement during cross-validation. The final models were compared using mean squared error, mean absolute error, and median absolute error. Here are the results of both regression models:

Model	Train R^2	Test R^2	Train MSE	Test MSE	Train Mean AE	Test Mean AE	Train Median AE	Test Median AE
No payment information	0.359	0.358	1.09e7	1.08e7	1516.98	1519.26	650.01	652.79
Payment Information	0.907	0.357	1.57e6	1.09e7	566.78	1516.56	222.12	604.1

Discussion of Outcomes

The model trained without payment information did equally poorly on training and testing sets. The success of the payment model on the training set is likely an artifact of having much more uniquely identifiable information. The decision trees were likely able to ‘recognize’ certain rows of data because of the unique attributes of that row, those attributes being the unique string of payments they had. What is interesting though is the reduction in the median testing error on the payment information model.

This amounts to an almost 7.5% reduction in median testing error. The mean squared error and mean absolute error did not change significantly, after investigation this is because of the outliers seen in the prescription dataset. Both models did equally poorly at predicting the large prescription rates shown on page 7. However, since the median absolute error decreased, the model with payment information did increase in predictive power.

With those models we can examine what features were most important during prediction. Here are the top 5 features for each model.

No Payments Model

	feature	importance
1349	LEVOTHYROXINE SODIUM	0.090877
2123	SIMVASTATIN	0.085321
226	AMLODIPINE BESYLATE	0.078578
1368	LISINOPRIL	0.078518
299	ATORVASTATIN CALCIUM	0.076280

Payments Model

	feature	importance
3612	Novo Nordisk Inc	0.045810
3355	Janssen Pharmaceuticals, Inc	0.035798
1349	LEVOTHYROXINE SODIUM	0.032142
2784	AstraZeneca Pharmaceuticals LP	0.030318
2123	SIMVASTATIN	0.029629

The importance metric here is relative, not absolute, so just because LEVOTHYROXINE SODIUM has an importance of 0.09 in the no payments model and 0.03 in the payments model, it does not mean that it is 1/3 as important in the payments model. But what is interesting to note are the manufacturer features that are ranked as highly important in the payments model. Each of the features in all capital letters is a drug name and the mixed case features are manufacturer names. This means that the manufacturer features in the payments model were more important in predicting the total day supply than knowing what drug was being prescribed. It is possible this could just be a strange artifact of the data, but the fact that multiple manufacturer features are more important than drug names in prediction indicate that knowing a provider received payments from these manufacturers offers more predictive power than only knowing the providers specialty and the name of the drug.

The results of this analysis do not reveal any inherently malicious activity, but it is an indicator that the manufacturers whose features are most important in prediction may have some influence on the rate of prescription for certain drugs and providers.

Further Research

Additional research needs to be conducted into the nature of the payments made to providers by each manufacturer in order of feature importance. Also a better model could be built by removing many of the outliers and by training on a larger percentage of the dataset. These models were trained on 15% of the total dataset, this subset was more than 1 million provider-drug pairings, but because of resource limitations a larger proportion of the dataset could not be used for training. Furthermore, additional features including the providers location should be added to the model to see if that can give even more predictive power. Another potential improvement could come from only adding payment information to drugs which a manufacturer actually manufactures. The approach used here blindly attaches payment information to every drug that a provider has prescribed, by only attaching payments to drugs that a provider prescribed and that manufacturer manufactures perhaps a stronger relationship would reveal itself.

Client Recommendations

It is my recommendation for the Department of Health and Human Services, that investigations be launched into the business practices of the top manufacturers by feature importance. I also recommend that the providers whose prescription rates were most accurately predicted by this model have their patients followed up with to determine if the claims filed were appropriate for the patient and in a reasonable amount. Thirdly, I recommend that, pending further investigation, limitations be placed on the types and amount of payments that drug manufacturers can make to healthcare providers.