Milestone Report

Medicaid Prescriptions and Drug Manufacturer Payments

The Problem

The problem that I want to solve, or rather the question I want to ask is, do payments made to doctors by drug manufacturers influence the rates at which their drugs are prescribed?

The Client

The client in this case could be from two organizations. The U.S. Department of Health and Human services, or a news organization. Both of these organizations would be interested in knowing the answer to this question, though the news organization for less benevolent motives. A news organization may be more interested if the answer is no, as it may draw more attention. However, the Department of Health and Human services would like to know this answer so it can better server United States citizens and advance their health and well-being.

If it is the case that the payments drug manufacturers make to physicians significantly influence the rates at which those drugs are prescribed, this could mean poorer treatment outcomes, and larger medical expenses for patients. In this case, the HHS may want to put through legislation that prohibits such payments by drug manufacturers in order to benefit United States citizens.

A news organization would want to report these findings to make their readers aware that their doctors could be prescribing them drugs they otherwise would not have, had they not received the payments from drug manufacturers.

The Data

The data I'm using here comes from two different sources, each with their own important fields, and I'm using two supplemental data sets in order to connect the data sources that I have. I'll describe each of these below, the field names are abbreviated where appropriate because some of the source field names are quite long.

Sources:

1) Payments data from the Center for Medicare and Medicaid Services.

   This data has a row for each recorded payment made to a health provider from a drug manufacturer. The important fields are: [Total_Payment_Amount], [Physician_ID], [Payment_Form], [Manufacturer_Making_Payment], and

[Nature_of_Payment]. There are some other fields in here that I have found useful for identifier reasons, and those are the physicians, first_name, last_name, middle_name, and their business address, city, and state. There also is a date of payment field which I used during my exploratory analysis, but I have not yet found it particularly useful in answering the question I am after.

2) Medicare Part-D prescription claims.

This data has a row for every combination of [Drug_Name] and [Provider_Identifier]. For each of those unique combinations, it provides some summary information, [Total_Claim_Count], [Total_Day_Supply], [Total_Drug_Cost]. The [Total_Claim_Count] represents the total number of claims that were filed for that drug by that health provider over that year, the day supply is the total day supply over the year prescribed by that provider for that drug.

Supplemental:

1) Food and Drug Administration

This data has information on a vast set of prescription and over the counter drug labels. Each drug label has the [Drug_Brand_Name] as well as the [Manufacturer_Name], which are the two fields I am after. With these two fields I created a reference table that allows me to tie together the manufacturers making payments, with the drugs that they manufacture.

2) National Plan and Provider Enumeration System

This data has demographic information on every United States registered provider. The important fields in this data set are the [First_Name], [Last_Name], [Middle_Name], [Business_Address], [City], [State] for every registered provider. With this supplemental data set I was able to get the [Provider_Identifier] for the providers in the payments data set from their demographic information. This was a crucial step in analysis, without it I would be unable to tie a provider who was paid by a manufacturer to their prescriptions.

Limitations:

The obvious limitation of the payment dataset is that it does not include the [Provider_Identifier] that is present in the prescription data set. I was able to overcome this with a

supplementary data set, I do not have perfect coverage, but I am able to tie together ~20% of the providers in the prescription dataset to those in the payments dataset. Another limitation is that the payments dataset only covers August through December of 2013. I do have both data sets for 2014, but so far all I have worked with are the 2013 datasets. It is likely I can get more coverage between the 2014 payments and prescription data sets.

Both of these datasets are pretty extensive in terms of the information they provide, but a question I can not answer would be, 'how does the rate of prescription claims filed change after a manufacturer payment?' The prescription data set does not have any time element associated with the prescriptions, just aggregate information from over the year.

Another limitation presented itself with the payments data. Since each payment does not come with a list of drugs that the manufacturer manufactures, I had to use the supplemental data set to find this out. I was able to tie together ~80% of the payments with the drugs that those payers manufacture. But unfortunately, I'm not able to perfectly cover the dataset without significant manual work.

Wrangling:

I had to do a significant amount of data cleaning and wrangling to get to this point. The most arduous of which was parsing the drug manufacturer information from the Food and Drug Administration's drug labels dataset. The downloaded data came in five json files and the data I was after was nested within each individual 'result' included in the files. Ultimately, I had to go through each result in each json file, pull out the 'label' data, remove any that were empty, cast the label data to a dictionary, convert the list of generic names to a comma separated string and then make a dataframe from that, the drug brand name, and the manufacturer name. Once I completed this process I saved the results to a 'csv' file that I now use to map the drug names to their manufacturers.

When tying together the manufacturers in the payments data to the manufacturers I had information for from the Food and Drug Administration's label data I had to manually look up some of the manufacturers so I could align the manufacturer name given in the payments data to the manufacturer name listed on the drug label. One example of this is Forest Pharmaceuticals. Forest Pharmaceuticals was listed as a manufacturer in the payments data set, but on their drug labels they list 'Forest Laboratories' which I confirmed is their parent company.

So after all of that, what I have are three data sets I can use for analysis. One data set on payments, one on prescriptions, and one with drug information. I have identifiers for the providers who received payments in the payments data set, which I can match with ~20% of the providers in the prescription data set, and I have drug brand and generic names that I can match with about ~80% of the

payments made by drug manufacturers.

Initial Findings

I have done some initial exploration with these three datasets from 2013 and done some hypothesis tests to see if there is a difference in the average number of claims filed and average total day supply between providers who were and were not paid by the drug manufacturer. I focused on only prescriptions claims for drugs that were made by the manufacturer and I had ~40k prescription records for those who were paid by Forest, and ~132k prescription records for providers who were not paid by Forest Pharmaceuticals. Both t-tests and Mann-Whitney tests give an incredibly high statistical significance showing a difference between these two groups. With a t-statistic of 33 and 47 for the difference in number of claims and total supply, respectively. Providers who were paid by Forest Pharmaceuticals filed on average more than 1.6 times as many claims and an over 1.7 times greater average total day supply in 2013. These give p-values of less than $1 * 10^{-200}$

In both groups there are significant outliers and there are a lot of factors that can play into this difference outside of the payments alone, but this initial finding suggests that there is something to be investigated here.

Going Forward

Ultimately I want to answer my general question, which is, does any payment from a drug manufacturer influence the rate of prescriptions filed for their drugs? I hypothesize that the answer is yes, and my initial findings are indicative of that conclusion. My biggest concern is that there is a spurious correlation between the payments drug manufacturers are making and the rates of prescriptions for those drugs. I want to eliminate or reduce this concern as much as possible, which is going to require further investigation into the prescribers.

Looking at only those prescribers who have filed claims for drugs made by the manufacturer is a good start, but the types of and number of claims filed will vary greatly depending on the patient population that the provider serves. I am considering using geographic location to control for patient population differences. I can operate under the assumption that geographically close providers should serve roughly the same types of patient populations. So if I can find providers who are geographically close, some of which received payments and others that did not, that may be my best shot at controlling for differences in patient populations.