Competition and Generic Drugs

An Analysis of Cost on Competition through Medicaid Data

David Zhang

Abstract

First, I analyze the reported average retail prices of drugs and percentage markup from manufacturing costs from Mark Cuban's Cost Plus Drugs website, with the manufacturing costs of these drugs through a simple linear regression. I find that both relationships are highly inelastic, as a 1% increase change in manufacturing costs are associated with only a 0.7% increase in the retail price on average, and a 1% increase in the manufacturing cost of a drug is only associated with a 0.3% decreasing in markup, on average. Secondly, I analyze Medicaid data from 2017-2021 inclusive and utilize fixed effects regression to control for time and entity effects as well as also controlling for demand, and attempt to estimate and quantify the causal effect of the number of manufacturers for a drug on various cost measures found in the data. I find that an increase in one manufacturer for a drug causes a decrease of 12.4% on total spending for a Medicaid drug. Finally, I walk through a qualitative analysis of NADAC data and show that the introduction of generic drugs is associated with a drastic decrease in the NADAC for that particular compound, and also that the distributions of NADAC for generic and brand name drugs are inherently different in a multitude of ways.

1 Introduction

According to Rajkumar (2020), global spending on prescription drugs is estimated to be approximately $1.3 trillion, with the United States accounting for $350 billion. Rajkumar (2020) also discusses possible causal links with high prescription drug costs, claiming that the most relevant cause of high prescription drug costs is the existence of monopolies, although does not directly investigate this claim empirically. Kesselheim et al (2016) also conclude that ”market exclusivity” is likely the primary contributor to high drug prices. Additionally, the FDA find that drug prices decline to approximately 55% of brand-name drug prices with 2 generic manufacturers making the product, 33% with 5 manufacturers, and 13% with 15 manufacturers, according to a paper released in August 2022 which attempts to quantify the savings due to the release of generic drugs.

There is a decent amount of literature which examines the drug development process in the U.S., and the costs associated with it. Lipsky and Sharp (2001) briefly outline the FDA approval process for prescription drugs, starting from preclinical testing to phases 1-4 of the FDA approval process, which takes on average around 8 to 12 years. Additionally, Moore et al (2018) provide statistics for the costs of this process by looking 59 novel therapeutic drugs approved by the FDA during 2015-2016: these costs range from $5 million to $346.8 million, depending on the circumstances of the clinical trials and the particular drug being tested. J.A. Dimasi et al (2016) also provide estimates of total R&D costs, pre-tax, for drugs during pre-approval, including the costs of drugs that are abandoned during testing, although these estimates have been criticized for lack of transparency and replicability. For the pre-human period, defined as discovery research and pre-clinical development, is estimated at around $1098 million capitalized (2013 dollars, $430 million out-of-pocket), and they estimate total costs as the costs of everything up until marketing and selling the drug, which they estimate as $2558 million capitalized (2013 dollars, $1395 million out of pocket). Mulcahy and Kareddy (2022) also provide a brief outline of a general supply chain model for brand-name drugs at retail pharmacies:

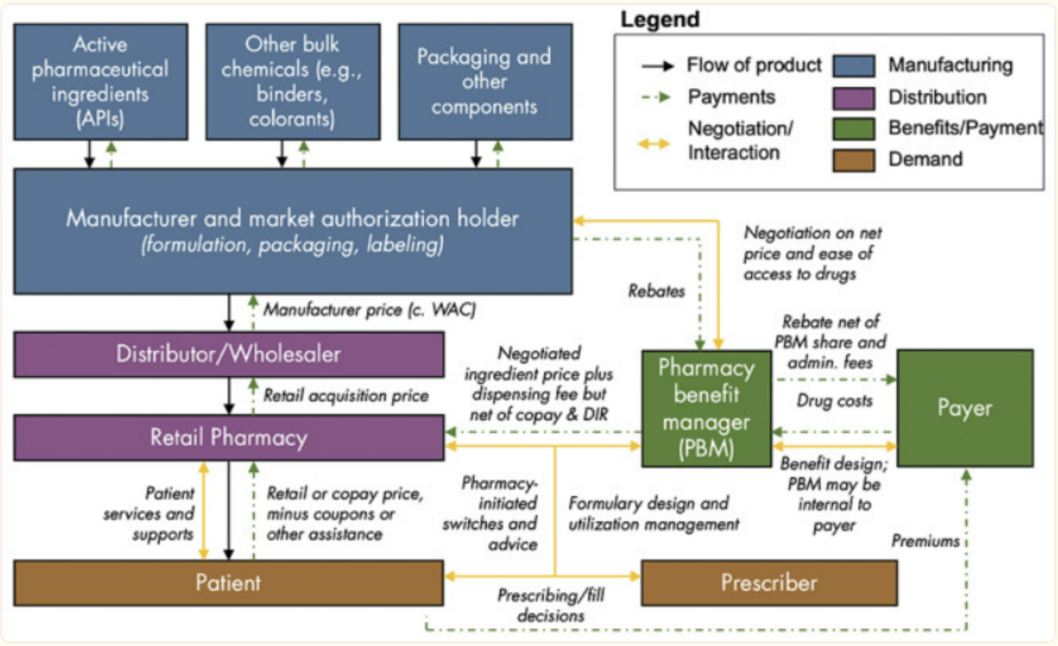

Active pharmaceutical ingredients (APIs)
Other bulk chemicals (e.g., binders, colorants)
Packaging and other components
Legend
Flow of product
Payments
Negotiation/ Interaction
Manufacturing
Distribution
Benefits/Payment
Demand
Manufacturer and market authorization holder
(formulation, packaging, labeling)
Manufacturer price (c. WAC)
Negotiation on net price and ease of access to drugs
Rebates
Distributor/Wholesaler
Retail acquisition price
Negotiated ingredient price plus dispensing fee but net of copay & DIR
Pharmacy benefit manager (PBM)
Rebate net of PBM share and admin. fees
Drug costs
Payer
Retail Pharmacy
Benefit design; PBM may be internal to payer
Patient services and supports
Retail or copay price, minus coupons or other assistance
Pharmacy-initiated switches and advice
Formulary design and utilization management
Premiums
Patient
Prescriber
Prescribing/fill decisions

Figure 1: Mulcahy & Kareddy (2022)

On the question of what goes into pharmaceutical pricing, there is also a decent amount of literature which examines the question, but the literature mainly focuses on the value of intellectual property, risks associated with drug development, and research and development costs, such as from Kesselheim et al (2016). Interestingly, Keyhani et al (2006) also argue that there is little correlation between research and development costs and increasing pharmaceutical prices, based on analyses regarding pharmaceutical development times. However, there is little literature on attempting to create a concrete and objective outline of the factors which could go into pharmaceutical pricing, and along the same lines, there is little literature to attempt to explain the causes of price variation in regions around the U.S. One explanation for price variation is price discrimination, but there is little literature which goes into causes of price discrimination. Even generic alternatives to brand-names in some cases cannot serve as a baseline price: Kesselheim et al. (2016) note several instances where generic drug manufacturers raised prices suddenly and significantly. As a result, it is very hard for policymakers and purchasers to have a good understanding of what is a “fair” price for medications. Daalen et al. (2021) attempts to centralize information on determinants of drug pricing through conducting a massive literature review in order to look for papers which investigate various factors that could affect drug pricing. To paraphrase, they find a lack of quality insight into the things which could factor into drug pricing, and the lack of quality evidence around price discrepancies for drugs does not enable one to make accurate conclusions about price discrepancies. They also find that retail drug prices internationally and nationally (United States) have a particularly high variance which cannot be entirely explained by the factors examined in the papers examined. However, Daalen et al. (2016) do find that determinants which are associated with lower drug prices include higher market share of generics, government purchasing, and pricing regulation. Determinants associated with higher drug prices include higher "originators" market share, which indicates original developers of a particular drug, and markets where drug markups are more common. In short, many of the problems lie with the fact that a lot of data related to determinants of drug pricing is either hard to obtain, or the data is insufficiently clear to draw conclusions about pharmaceutical pricing.

One thing that is common with all of these papers is that none of them attempt to rigorously draw out causal relationships in regards to competition and pharmaceutical prices, as well as other factors. As such, I will be directly investigating causal relationships between various pharmaceutical cost metrics and other variables within this paper.

2 Data Overview

2.1 Medicaid - Spending By Drug

The first data source that I will discuss is Medicaid - Spending by Drug. This data contains Medicaid spending by drug from 2017 to 2021 with a variety of other related features within the data set. It represents national-level drug spending for Medicaid covered outpatient drugs, which are paid for by state Medicaid agencies. Over-the-counter drugs in the Medicaid

State Drug Utilization data are not included, and drugs with fewer than 11 claims in the most recent year are also not included. Overall, this includes 16,146 observations across 5 time periods, each of which is a particular drug/compound, with some duplicates in the case of drugs with different dosage forms, strengths, and methods of administration. For each drug, there is an entry which calculates the same features (when applicable) for all of the manufactures of that drug. In total, there are 80,730 observations when accounting for time.

In addition to the 16,146 total observations, each observation has a total of 36 features. Four of the features are the brand name of the drug, the generic equivalent of that drug, the number of manufacturers for the drug, and the name of the manufacturer of the drug: these features are time-invariant during this period. 30 of the features can be divided into 6 distinct spending/financial/counting features for each year in 2017-2021, inclusive. After transforming the data, the data consists of 10 features. For the analyses in this paper, the data is filtered to consider only the entries where the statistics are calculated by aggregating all of the manufacturers for a particular drug, the outlier flag is 0, and all the other statistics are greater than 0. The 10 distinct features are described in Table 6; summary statistics are given in Table 7, and Table 8 for the filtered data.

Outlier_Flag is an indicator variable which represents if the drug is an outlier or not. A drug is labeled an outlier by the standard definition using the IQR, in addition to if it meets two other conditions. First, if there are fewer than 30 records associated with the drug from the initial data source, then it is flagged as an outlier. Second, average spending per dosage unit is calculated with and without outlier records. If these calculated amounts differ by 10 percent and $1, then it is flagged as an outlier. The data used in this paper only includes observations for which Outlier_Flag is 0 i.e. the observations are not outliers.

The last two features are percent change in average total spending per dosage unit from 2020-2021, and the annual growth rate in average total spending per dosage unit throughout 2017-2021, calculated using the compound annual growth rate. These are not used in the analyses.

2.2 Cost Plus Drugs

The second data source that I will discuss is the data from Mark Cuban's Cost Plus Drugs. To obtain the data, I web scraped the website page (legally) containing a list of all medications, using Selenium and BeautifulSoup packages in Python. The features are summarized by the following tables:

This data has 442 observations, with two duplicate rows for two medications that each have two different dosage forms. This data has seven features: the medication name, brand name, dosage form, retail price, the Cost Plus price, the manufacturing cost, and the markup. Cost Plus sells only generic medication, so the medication name is the name for the generic. The retail price is the brand name price as reported by Cost Plus Drugs. According the website, this is calculated as "an average of retail prices across other pharmacies that is sourced from third-party data. This information is published at the time we add the medication so may not reflect current market prices." Therefore, some of the interpretation of this retail price must be taken with a grain of salt, but I choose to give the benefit of the doubt in this paper.

The Cost Plus price is the price that Cost Plus is selling the medication for: it is calculated by factoring in manufacturing/labor/shipping costs in addition to a 15 percent markup. In particular, the Cost Plus price can be decomposed as

$$CP\_price = 1.15(manuf\_cost) + 3$$

where 3 is the pharmacy labor cost. Tax and shipping costs are not included. Features are describe in Table 9.

2.3 NADAC

The third data source is the NADAC data, which stands for the National Average Drug Acquisition Cost. This data contains 1,298,197 observations with 12 features. For the sake of avoiding the difficulties of interpreting the data during and soon after the COVID-19 pandemic, I specifically choose to use the NADAC data from 2019. Two of the features uniquely identify the drug: these are the National Drug Code description, which includes the drug name, strength, and dosage form; and the NDC, which is an 11 digit code. The features are summarized in Table 11, and summary statistics in Table 12.

3 Manufacturing Costs: Mark Cuban's Cost Plus Drugs

I will first investigate the relationship between manufacturing costs and pricing by looking at the Cost Plus data.

I begin by attempting to isolate an estimate of the manufacturing cost from the CP_price. As alluded to earlier, this is given by

$$ManufacturingCosts = \frac{CP\_price - 3}{1.15}$$

.

Referring back to the literature review, I assume that this calculation of manufacturing costs accounts for the costs reflected in the manufacturing cells (colored blue) in Figure 1.

One place to start with regards to looking at the relationship between manufacturing costs and pricing is to look at the extent of a linear relationship between the manufacturing costs and the retail price given in the data. We can first start by looking at a scatterplot of the data:

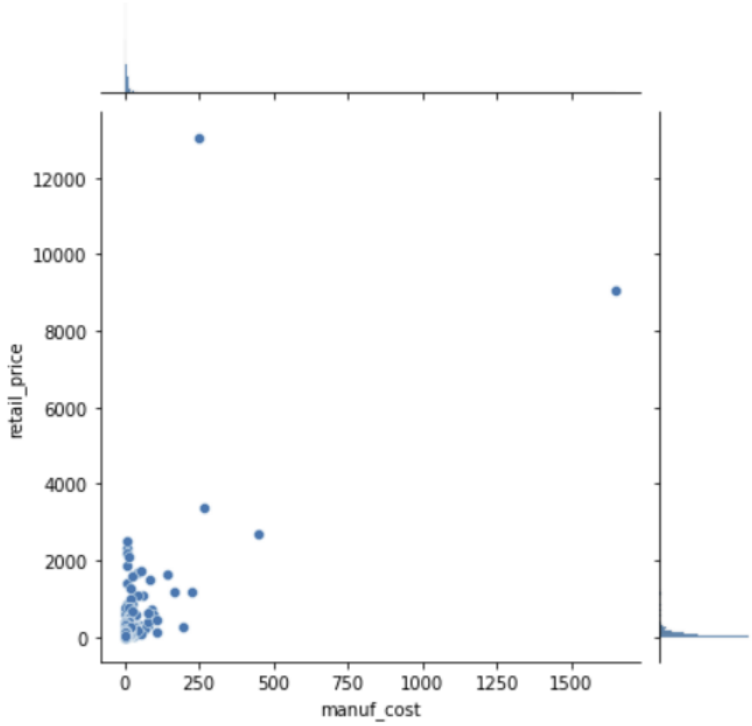


Figure 2: Jointplot of manufacturing costs (x-axis) and retail price (y-axis). Both variables are highly right-skewed.

The highly right skewed nature of both of the variables suggests that taking the natural logarithm of both of the variables may better reveal a linear relationship[1]:

[1] Almost all of the variables in this paper have this relationship, so similar plots for other variables are omitted for brevity.

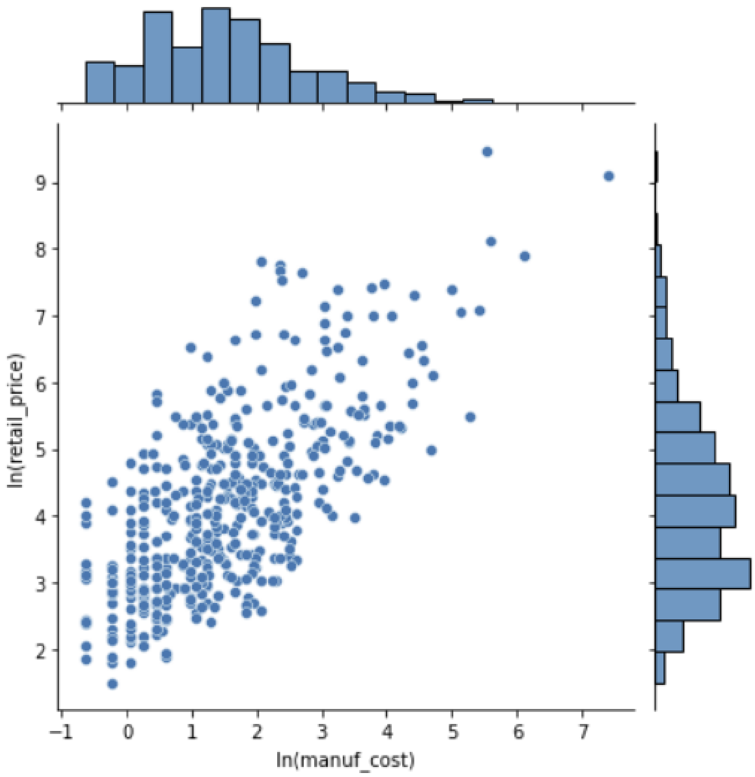


Figure 3: Jointplot of manufacturing costs (x-axis) and retail price (y-axis) after taking the natural logarithm.

Running an OLS regression on the variables above, we get

	Dependent variable: log(retail_price)
	(1)
const	3.068^{***}
	(0.066)
log(manuf_cost)	0.721^{***}
	(0.034)
Observations	442
R^2	0.483
Adjusted R^2	0.482
Residual Std. Error	0.974(df = 440)
F Statistic	446.666^{***} (df = 1.0; 440.0)
Note: Standard Errors are heteroscedasticity robust (HC3).	*p<0.1; **p<0.05; ***p<0.01

With an R-squared value of .483 and highly significant coefficients, it appears that manufacturing costs explain a significant proportion of the variance of retail prices of the drugs in Cost Plus. However, one thing to note is that if medicine was priced with a lot of the costs in mind, then we should expect a coefficient that is greater than or equal to 1. However, this is not the case, and the 95% confidence interval is far from it as well. Since the coefficient is less than 1, we can characterize this relationship as quite inelastic. In other words, the retail price doesn't change much with a change in manufacturing cost. Also, while the manufacturing costs account for almost half of the variance of the retail price, there are other highly relevant factors other than manufacturing-related costs which account for the final retail price.

In an attempt to generalize this, let's assume that the price of every drug can be decomposed into manufacturing costs, a percentage markup from those manufacturing costs, and a fixed labor cost. In the case of Cost Plus, we know the associated manufacturing costs and the labor cost, and so we can estimate the percent markup to the retail price for each drug. I assume that the labor cost is also \$3 for the retail pricing. This is calculated as

$$markup = 100(\frac{retail\_price - 3}{manuf\_cost} - 1)$$

.

Due to right-skewness, the next regression will be log(markup) on log(manuf_cost):

	Dependent variable: log(markup)
	(1)
const	7.411***
	(0.083)
log(manuf_cost)	-0.275***
	(0.044)
Observations	442
R^2	0.084
Adjusted R^2	0.082
Residual Std. Error	1.186(df = 440)
F Statistic	39.883*** (df = 1.0; 440.0)
Note: Standard Errors are heteroscedasticity robust (HC3).	*p<0.1; **p<0.05; ***p<0.01

Interestingly, the coefficient on log(manuf_cost) is negative. In particular, this indicates that for a 1% change in manuf_cost, markup decreases by .275%, on average. Similarly to the first regression, the markup is quite inelastic to changes in manufacturing cost, to an even higher degree compared to the retail price. On first glance, the sign on the coefficient doesn't make sense, since we would expect that an increase in manufacturing cost should increase the markup, since the manufacturer would want to recoup more of the manufacturing cost.

However, the first regression suggests that as the manufacturing cost increases, the retail price increases as well, although not as much as the manufacturing cost. Therefore, the manufacturing cost increases faster than the retail price on average, and so an increase in the manufacturing costs results in a decrease the markup. However, the R^2 is quite low for the second regression, which indicates that manufacturing costs explain little of the variance in the markup.

Overall, these results suggest that manufacturing costs have a small effect on the high retail prices. In particular, manufacturing costs explain very little of the variance in the markup. For the markup and retail prices, they are both relatively unresponsive to changes in the manufacturing cost. One weakness of this approach is that it does not account for the costs incurred to discover the chemical formula for the drug, which are a real part of the manufacturing costs, especially for the companies which discover and charge the retail prices. One way to interpret the markup is the extra price charged for the intellectual property rights in discovering the drug, and this is not captured solely by the costs to produce the medicine, which are significantly lower once the drug has been discovered. This is far more likely to explain the markup, but this data is unavailable.

4 Competition and Costs

As alluded to in the literature review, several scientists have suggested that monopoly power is a primary cause of high pharmaceutical costs. In this section, I outline the results of empirically investigating these causal claims from looking at the Medicaid - Spending by Drug data.

4.1 Theoretical Background

First, I will give a brief background on the relevant microeconomic theory [2], specifically the case of a monopoly. Let $y = D(p)$ be the demand curve which is a function of price p. In the case of a monopoly, the firm can choose p, in which case the quantity y is determined by the demand. Alternatively, the quantity y can be set by the firm, and so the price is determined as $p(y) = D^{-1}(y)$. The monopoly firm solves the maximization problem

$$\max_{y} \quad p(y)y - c(y). \tag{1}$$

The first order condition is

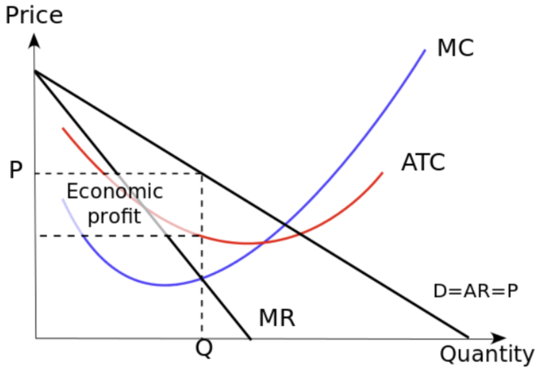
$$p'(y)y + p(y) = c'_y(y)$$

where the $p'(y)y + p(y)$ is the marginal revenue, and $c'_y(y)$ is the marginal cost. In other words, the maximizing y is at the intersection of the marginal revenue and marginal cost.

[2] Sourced from Prof. Cecile Gaubert's notes from ECON 101A, Fall 2021.

Interestingly, the marginal revenue curve is significantly different than the marginal revenue in perfect competition, as marginal revenue is now a function of the quantity produced. In the case of perfect competition (and cases with more than one firm in general), an individual firm cannot totally control the market price: the market price is determined in part by strategic interactions with other competing firms. As the number of firms goes to infinity, the market price cannot be manipulated at all by any single firm: every firm is a price taker, and so the marginal revenue is a constant i.e. the market price. However, a monopoly can now set the market price by manipulating the quantity produced, and so has the freedom to choose the price such that its own economic profit is positive and maximized above all else. This manifests as "unnaturally" high prices.

One thing to note is that the marginal revenue curve lies below the demand curve, since it is the demand curve plus a negative number. Also, marginal revenue is a function of the quantity: in perfect competition, the price is given and is treated as a constant. The monopoly chooses the quantity which is at the intersection of the marginal revenue and marginal cost curve, but since the marginal revenue curve lies below the demand curve, the monopoly charges a price which is higher than the price in perfect competition, where the price in perfect competition is given by the intersection of marginal cost (the supply curve in perfect competition) and demand. Thus, this implies that as the number of firms increases, the market price charged for a good decreases.



Monopoly Production: Monopolies produce at the point where marginal revenue equals marginal costs, but charge the price expressed on the market demand curve for that quantity of production.

Figure 4: The monopoly chooses the price on the demand curve which correponds to the quantity at the intersection of MR = MC (LibreTexts).

4.2 Empirical Analysis

We now see if these theoretical conclusions hold in an empirical setting through looking at the filtered Medicaid data. To do this, we can look at the effect of the total number of manufacturers on various cost metrics within the data through two forms of linear regression: first through a naive regression with just the single feature of manufacturers and then controlling for time and entity effects, as well as demand through accounting for total claims.

The cost metrics we will investigate are Avg_Spnd_Per_Dsg_Unt_Wghtd, Avg_Spnd_Per_Clm, and Tot_Spndng (average spending per dosage unit, average spending per claim, and total spending, respectively). For all of these analyses, I work with the natural logarithm of each of these variables. The total manufacturers for the drug is referred to as Tot_Mftr.

In general for the fixed effects regressions, I estimate the equation

$$Y_{it} = \beta X_{it} + \alpha_i + \lambda_t + u_{it}$$

where α_i represents the individual fixed effects, and λ_t represents the time fixed effects. This is equivalent to the within estimator, which is used for the computations.

Table 1: log(Avg_Spnd_Per_Dsg_Unt) on Tot_Mftr

	Dependent variable: log(Avg_Spnd_Per_Dsg_Unt)		
	OLS	*Fixed Effects* *(Panel)*	
	Naive	Default	Clustered S.E.
	(1)	(2)	(3)
Tot_Mftr	-0.186^{***}	-0.084^{***}	-0.084
	(0.005)	(0.007)	(0.064)
log(Tot_Clms)		-0.036^{***}	-0.036^{***}
		(0.005)	(0.014)
Constant	2.332^{***}		
	(0.028)		
Observations	12,740	12,740	12,740
R^2	0.109	0.031	0.031
Adjusted R^2	0.108	-0.292	-0.292
Residual Std. Error	2.553 (df = 12738)		
F Statistic	$1{,}551^{***}$ (df = 1; 12738)	153^{***} (df = 2; 9550)	153^{***} (df = 2; 9550)
Note: Entity/Time Cluster			*p<0.1; **p<0.05; ***p<0.01

Table 2: log(Tot_Spndng) on Tot_Mftr

	Dependent variable: log(Tot_Spndng)		
	OLS	*Fixed Effects* *(Panel)*	
	Naive	Default	Clustered S.E.
	(1)	(2)	(3)
Tot_Mftr	0.150^{***}	-0.124^{***}	-0.124^{***}
	(0.005)	(0.006)	(0.043)
log(Tot_Clms)		0.973^{***}	0.973^{***}
		(0.004)	(0.011)
Constant	14.047^{***}		
	(0.027)		
Observations	12,740	12,740	12,740
R^2	0.079	0.884	0.884
Adjusted R^2	0.079	0.845	0.845
Residual Std. Error	2.455 (df = 12738)		
F Statistic	$1{,}095^{***}$ (df = 1; 12738)	$36{,}449^{***}$ (df = 2; 9550)	$36{,}449^{***}$ (df = 2; 9550)
Note: Entity/Time Cluster			*p<0.1; **p<0.05; ***p<0.01

Table 3: log(Avg_Spnd_Per_Clm) on Tot_Mftr

	Dependent variable: log(Avg_Spnd_Per_Clm)		
	OLS	*Fixed Effects* *(Panel)*	
	Naive	Default	Clustered S.E.
	(1)	(2)	(3)
Tot_Mftr	-0.170^{***}	-0.124^{***}	-0.124^{***}
	(0.004)	(0.006)	(0.043)
log(Tot_Clms)		-0.027^{***}	-0.027^{**}
		(0.004)	(0.011)
Constant	5.875^{***}		
	(0.021)		
Observations	12,740	12,740	12,740
R^2	0.146	0.070	0.070
Adjusted R^2	0.145	-0.240	-0.240
Residual Std. Error	1.974 (df = 12738)		
F Statistic	$2,169^{***}$ (df = 1; 12738)	362^{***} (df = 2; 9550)	362^{***} (df = 2; 9550)
Note: Entity/Time Cluster			*p<0.1; **p<0.05; ***p<0.01

The first two naive regressions have a negative sign on the coefficient, which is what we expect based on economic theory. In particular, an increase in one manufacturer results in an 18.6% decrease in the average spend per dosage unit, on average; and a 17% decrease in the average spend per claim. Interestingly, an increase in one manufacturer seems to result in a 15% increase in total spending for a drug, as seen in the third naive regression, which seems to conflict with microeconomic theory. However, this is likely due to omitted variable bias.

One possible explanation of the result in the naive regression for total spending is that drugs with a higher number of manufacturers may also have a higher number of people using that drug (higher demand), which would also increase total spending. I account for this by including the natural logarithm of the total number of claims in the regression. In general, it seems plausible that demand causes omitted variable bias in the other regressions as well, so I include it in all of the regressions.

As seen in the regression tables, the inclusion of fixed effects and the natural logarithm of total claims significantly changes the coefficients and R^2 of the regressions.

In regression (3) for average spend per dosage unit (Table 1), the R^2 decreases to 0.031, and the adjusted R^2 is -0.292. This suggests a poor fit, and that adding log(Tot_Clms) does not add any predictive power. Regression (3) seems to indicate that when clustering standard errors to account for heteroscedasticity and autocorrelation, the number of manufacturers may not have much effect on the average spend per dosage unit. Additionally, the coefficent on Tot_Mftr decreased in magnitude, indicating that we were able to account for some omitted variable bias from time-fixed effects and total claims. The coefficient on Tot_Mftr indicates that the average spend per dosage unit decreases by 8% on average for an increase in 1 manufacturer, holding other variables constant, although it is not significant. In general, after controlling for fixed effects, the regressions suggest that there is little to no relationship between the number of manufacturers and the average spend per dosage unit when also controlling for demand. Additionally, the coefficient on log(Tot_Clms) shows that a 1%

increase in total claims results in a 0.03 decrease in average spend per dosage unit on average holding all else constant, which suggests that the average cost per unit of Medicaid medicines is not strongly influenced by demand.

In regression (3) for total spending (Table 2), the R^2 increases to 0.884, and the adjusted R^2 is 0.845: this is a massive increase. The most surprising result is that the sign of the coefficient on Tot_Mftr changes in magnitude and sign, indicating that the omitted variable bias has been successfully accounted for. In particular, we can now see that the effect of the total manufacturers aligns with what we expect from economic theory, in that an increase in total manufacturers results in a decrease in total spending for a particular drug, on average. Specifically, an increase of 1 manufacturer results in a 12.4% decrease in total spending on average, holding other variables constant. The effect is also significant at the 1% level. Another thing to notice is that the coefficient on log(Tot_Clms) suggests that when holding all else constant, the total spending can be interpreted almost entirely as a function of the total claims, since the coefficient suggests that a 1% increase in total claims results in a 0.973% increase in total spending on average, which is very close to a "$y = x$" line.

Looking at regression (3) for average cost per claim (Table 3), we can see that controlling for fixed effects lowered the R^2 to 0.070, and the adjusted R^2 is -0.240, indicating a poor fit similar to the results in Table 1. We may think that this regression should provide similar results to the regression for total spending with regards to fit, since the average spend per claim is simply an averaged version of total spending. However, this regression indicates that after we average the total spending by the total claims, the manufacturers and total claims provide little predictive power, indicating that there may be more relevant variables to consider. Also, we can notice that while the coefficient on Tot_Mftr is the same as in Table 2, the coefficient on log(Tot_Clms) is very small. Specifically, it indicates that a 1% increase in total claims results in a 0.02 decrease in the average spend per claim on average, holding all else constant. This suggests that the average cost per claim of Medicaid medicines are actually quite stable, and are not strongly influenced by demand. This is a similar result to Table 1.

In summary, the regressions seem to confirm what we know from economic theory and the claims of several scientists from the literature. One thing to note is that the degree of generalizability of the results of the regression are likely not high. This is because the Medicaid data is not a good representation of general medicine usage patterns and costs since Medicaid serves a particular demographic of people, specifically low-income individuals, the elderly, and disabled. In other words, the Medicaid data may not be representative of the population at large, and so the conclusions from the regressions may vary with more general data. Additionally, the R^2 values are quite low with regards to the average cost per dosage and average cost per claim, indicating that the total number of manufacturers does not seem to explain a high degree of the variance in the average cost of drugs, which somewhat conflicts with the claims made by the scientists mentioned in the literature review, specifically Rajkumar and Kesselheim et al (2016). However, I do believe that these results are very applicable to Medicaid specifically.

Since I control for demand and fixed effects, I argue that the coefficients are causal, specifically for these particular Medicaid companies during this time period. The regressions run are essentially population level regressions, where the population is the Medicaid companies during this time period. I argue that it is safe and perhaps even best to disregard the outlier flagged entries, since I argue that the outlier flagged entries do not accurately reflect the Medicaid company population anyway. Additionally, I argue that controlling for demand and fixed effects accounts for most, if not all, ommitted variable bias, since demand is a strong determinant of relevant spending measures, particularly for Medicaid and medicines in general.

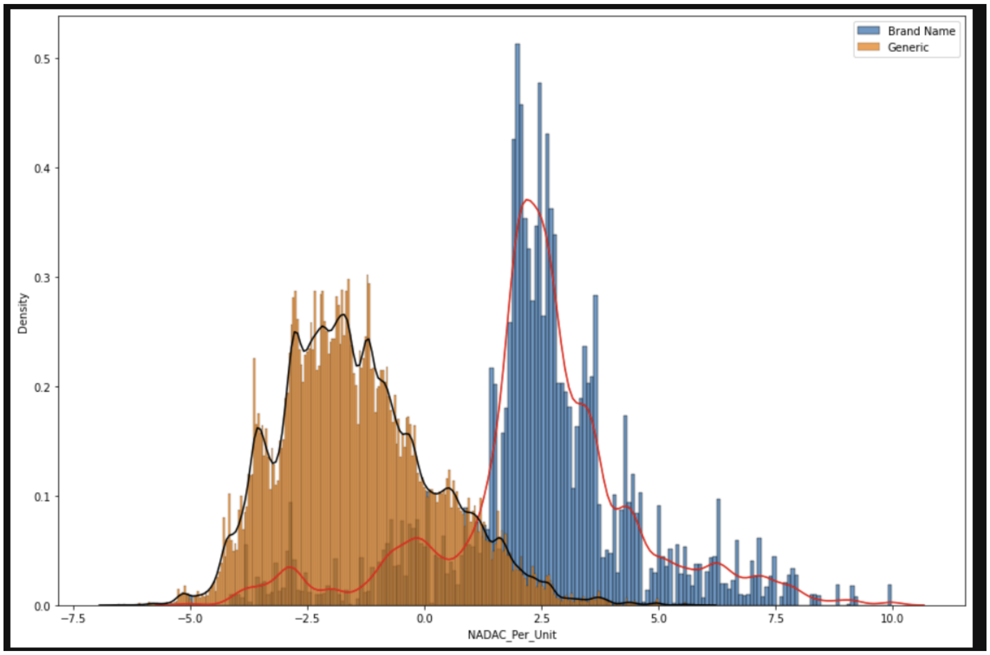


Figure 5: Histograms of log(NADAC Per Unit) of strictly brand name and generic drugs.

5 Brand to Generic

This final section will investigate differences between the NADAC of brand name drugs and their generic counterparts through a more qualitative approach. I first separate the NADAC data into drugs with no generic counterpart, and the drugs where the rate is classified as generic in the NADAC. I then plot a density histogram of the natural logarithm of the NADAC Per Unit for both groups:

Note that the brand name drugs distribution lies farther to the right along the x-axis, which indicates that brand name drugs are priced significantly higher on average compared to generics. The brand name density is also concentrated higher at a significantly higher NADAC per unit for brand name drugs: this tells us that brand name drugs are very consistently priced at a specific high “price range”. Additionally, we can also see that where most of the generic NADAC per unit values are, the brand name drugs are sparse in that same area. The lines drawn over the histograms are kernel density estimations (KDEs) of the respective distributions. Also note the fat tails of the brand name distribution, which indicates significantly high variance compared to the relatively symmetric and approximately normal distribution of the generic NADAC.

Now we consider brand name drugs which have had a corresponding generic introduced at some point. In total, there are 1286 of these drugs after accounting for duplicates.

Table 4: Summary Statistics - Brand Name with Generic

Statistic	N	Mean	St. Dev.	Min	Max
NADAC_Per_Unit	1,286	16.173	41.481	0.005	471.559
Corresponding_Generic_Drug_NADAC_Per_Unit	1,286	4.514	19.310	0.003	330.858
nadac_diff	1,286	−11.659	28.786	−324.811	2.277
nadac_percent_change	1,286	−0.696	0.292	−0.999	1.679

The nadac_diff is calculated by taking the difference between the corresponding generic NADAC per unit and the regular NADAC per unit. nadac_percent_change is calculated by taking nadac_diff and dividing it by the corresponding NADAC per unit. One thing we can immediately notice is that the NADAC per unit for most drugs significantly decreases when a generic is introduced. However, we can also see that there are some drugs which happen to experience an increase in price when the generic is introduced: in particular, one drug increases by 67%! After filtering, there are 16 drugs which increase in price when the generic is introduced, with no obvious reason why. We can look at the summary statistics of these drugs:

Table 5: Summary Statistics - Generic is More Expensive

Statistic	N	Mean	St. Dev.	Min	Max
NADAC_Per_Unit	16	2.018	2.150	0.081	7.245
Corresponding_Generic_Drug_NADAC_Per_Unit	16	2.493	2.257	0.116	7.686
nadac_diff	16	0.474	0.656	−0.027	2.277
nadac_percent_change	16	0.457	0.452	0.057	1.679

For the drugs where the generic ends up more expensive, it is on average 45% more expensive. However, since there are only 16 drugs, an increase in price in the generic is the exception rather than the rule.

In any case, we can take a look at the distribution of the percent change:

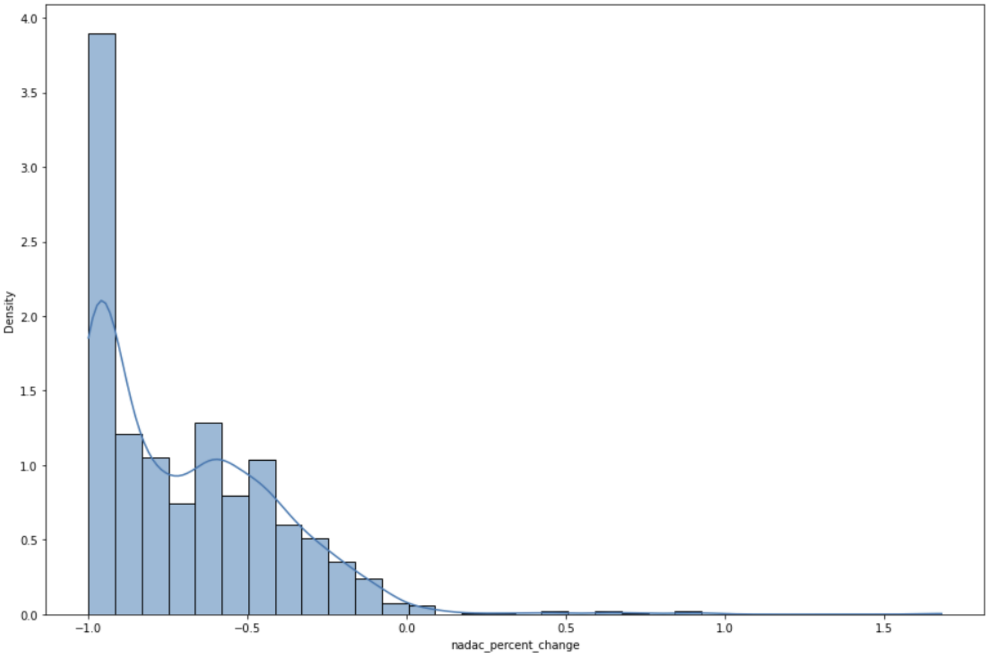


Figure 6: Density histogram of the percent change in NADAC

We can see that a significant proportion of drugs experience an approximately 90% decrease in NADAC per unit, comparing the brand name NADAC to the generic counterpart. This is a drastic decrease, and indicates that there is a high degree of “unnecessary” pricing when it comes to certain brand name drugs, in that the brand name drug pricing is significantly higher than what is needed to make non-zero profit. These are telltale signs of monopoly pricing, as discussed in the theoretical background section.

In summary, the NADAC per unit of generic drugs tend to be significantly lower than those which are strictly brand name drugs. Additionally, the NADAC per unit for strictly brand name drugs has very high variance, while the NADAC per unit for generic drugs has much more normal variance. For the drugs which have a generic counterpart introduced at some point, a significant proportion of the drugs experience at least an 80% decrease in the NADAC per unit and the vast majority of the drugs experience a decrease in the NADAC in general. A select few end up experiencing an increase in the NADAC per unit when the generic is introduced, but this is a very small proportion and may be a result of confounding factors which have to do with the specific circumstances of those drugs. Most importantly, these results line up with many of the claims in the literature, that the introduction of generic drugs results in a significant decrease in the cost that consumers face. Unfortunately, this analysis is only for the year of 2019, and so a much more robust way of confirming this pattern of decreasing prices is to analyze this same type of data over time. Due to poor data formatting and lack of computational power, it is extraordinarily difficult to merge all of the data provided of the NADAC and analyze this over time, and so it must be taken with a grain of salt. However, there is no immediately obvious reason to believe that this relationship would not generally hold over time.

6 Conclusion

From the analysis of the Cost Plus Drugs data, it suggests that manufacturing costs are a factor when it comes to understanding the retail (brand-name) prices. However, they explain little about the markup from the manufacturing price of the drug. Additionally, the regressions indicate that the retail price and the markup are inelastic to changes in the manufacturing cost i.e. the retail price and markup do not change much when the manufacturing cost changes. Unfortunately, there may be better indicators of what determines the retail price, but this data is not available. Also, the lack of controls for the regressions and the ambiguity of the calculation of the retail price do not make for robust results, but the outcomes of the regressions are still great preliminary analyses of the relationships with no obvious omitted variable bias or other bias that hasn't bee mentioned.

The regressions from the Medicaid data do confirm Rajkumar and Kesselheim et al (2016). The number of manufacturers for a particular drug seem to have a causal relationship with the cost of the drug on average, as measured through three cost metrics in the data. These relationships stand even when controlling for fixed effects and demand for that drug, which indicates that the relationship is quite robust, and suggests a causal relationship. This helps to further confirm intuitions founded on microeconomic theory. The R^2 values for the regressions are not very high, which may suggest that Rajkumar and Kesselheim et al (2016) overstate the relationship between monopoly power and high drug prices. Although the generalizability outside of Medicaid is questionable, I argue that the relationship is strong and very relevant for Medicaid.

Based on a qualitative analysis of the NADAC data, we can see that brand name drugs tend to cost significantly more on average compared to strictly generic drugs. When looking at brand name drugs which have had a generic counterpart released at some point, the cost significantly decreases for the vast majority of these drugs. However, more work is needed to verify that these results hold over time.

7 Data Overview Tables

Features	*Description*
Brnd_Name	The brand name of the drug, if there is a brand name (trademarked name). Otherwise, the generic name is listed.
Gnrc_Name	The name of the chemical compound which makes up the drug.
Tot_Mftr	The total number of manufacturers for each drug.
Mftr_Name	Name of the manufacturer of the drug.
Year	Year of the particular observation.
Tot_Spndng	Total Medicaid spending for the corresponding drug and year.
Tot_Dsg_Unts	Total dosage units of the drug which are dispensed in the corresponding year. The dosage unit refers to the lowest dispensable amount for the corresponding drug.
Tot_Clms	The total number of prescription fills for the drug, consisting of original prescriptions and refills.
Avg_Spnd_Per_Dsg_Unt_Wghtd	This is the Medicaid spending divided by the total number of dosage units. This is weighted by the proportion of total claims with respect to different forms of the drug, such as different dosage strengths, forms, etc.
Outlier_Flag	An indicator variable which equals 1 if the drug's statistics are affected by some form of outlier circumstances. Further details can be found below.

Table 6: Description of the features in the Medicaid data.

Table 7: Summary Statistics (Medicaid, Unfiltered)

Statistic	N	Mean	St. Dev.	Min	Max
Tot_Mftr	80,730	1.462	2.289	1	44
Tot_Spndng	80,730	9,105,021	55,096,357	0	2,730,141,376
Tot_Dsg_Unts	80,730	7,056,597	283,508,476	0	56,150,332,658
Tot_Clms	80,730	83,800.920	478,348.500	0	15,084,149
Avg_Spnd_Per_Dsg_Unt_Wghtd	80,730	279.376	14,558.810	0	1,898,689
Avg_Spnd_Per_Clm	80,730	1,152.983	14,688.430	0	1,717,862

Table 8: Summary Statistics (Medicaid, Filtered)

Statistic	N	Mean	St. Dev.	Min	Max
Tot_Mftr	12,740	3.338	4.789	1	44
Tot_Spndng	12,740	26,330,881	94,373,353	216.710	2,730,141,376
Tot_Dsg_Unts	12,740	22,030,741	529,482,180	57.273	56,150,332,658
Tot_Clms	12,740	251,271.300	1,039,414	36	15,084,149
Avg_Spnd_Per_Dsg_Unt_Wghtd	12,740	283.404	1,699.071	0.0003	39,359.910
Avg_Spnd_Per_Clm	12,740	2,097.988	7,255.865	0.895	187,153.900

Features	*Description*
medication	The generic name of the corresponding drug. Also notes the brand name counterpart.
brand_name	The brand (trademarked) name of the corresponding drug.
form	The dosage form of the drug when purchased from the website.
retail_price	The brand name price as reported by Cost Plus Drugs. According to the website, this is calculated as ”an average of retail prices across other pharmacies that is sourced from third-party data. This information is published at the time we add the medication so may not reflect current market prices.”
CP_price	The sale price of the medication when purchasing from Cost Plus.
manuf_cost	Estimate of the manufacturing costs of the drug. Method of calculation is discussed in the Manufacturing Costs and Pricing section.
markup	Percentage markup from the CP_price to the retail_price. Method of calculation is discussed in the Manufacturing Costs and Pricing section.

Table 9: Description of the features in the Cost Plus Drugs data.

Statistic	N	Mean	St. Dev.	Min	Max
retail_price	442	229.526	838.190	4.500	13,067.140
CP_price	442	22.543	98.288	3.600	1,900.500
manuf_cost	442	16.994	85.468	0.522	1,650.000
markup	442	2,426.230	3,730.934	26.099	31,876.390

Table 10: Summary Statistics (Cost Plus)

Features	*Description*
NDC Description	Includes the name, strength, and dosage form of the particular drug.
NDC	11 digit code which uniquely identifies the drug. Specifically, it is the NDC package code.
NADAC Per Unit	National Average Drug Acquistion Cost per unit.
Effective Date	The effective date of the NADAC entry for the drug.
Pricing Unit	The unit which the price is calculated with. “EA” stands for discrete units, such as pills. “ML” stands for milliliters, and “GM” stands for grams.
Pharmacy Type Indicator	The source of pharmacy survey data used for the NADAC. The only entry is ”C/I”, which stands for chain and independent pharmacies. This column does not give new information, since all entries are the same.
OTC	Indicates whether the drug is sold over-the-counter or not. ”Y” means yes, ”N” means no.
Explanation Code	Number from 1-10 which indicates how the NADAC was calculated. None are particularly problematic for this investigation, so I will omit the descriptions for brevity.
Classification for Rate Setting	Indicates whether the drug is considered brand (‘B’) or generic (‘G’) for the NADAC rate calculation process. Two other indicators specify further circumstances for brand name drugs: they are not problematic for this investigation.
Corresponding Generic Drug NADAC Per Unit	NADAC for the corresponding generic drug. The entry is NaN if there is no generic.
Corresponding Generic Drug Effective Date	The effective date of when the Corresponding Generic Drug NADAC Per Unit is assigned to a drug. This date may not correspond to the NADAC effective date for the generic drug.

Table 11: Description of the features in the NADAC data.

Table 12: Summary Statistics (NADAC)

Statistic	N	Mean	St. Dev.	Min	Max
NADAC_Per_Unit	1,298,197	10.529	240.682	0.001	21,513.920
Corresponding_Generic_Drug_NADAC_Per_Unit	50,251	4.502	20.375	0.003	372.074