Predicting Generic Drug Shortages for Picket Pharmaceutical

Kevin Nguyen

September 27, 2022

1 Executive Summary

Generic drugs typically cost less than brand-name drugs but have identical active ingredients, administration methods, dosage forms, strengths, and indications. Their lower prices are the result of generic drugs being sold after their brand-name equivalents' patents have expired. Drug shortages are a persistent issue in the United States, particularly for off-patent, injectable products with few suppliers. The majority of drug shortages are caused by manufacturing quality issues, which force a pause in production while the problem is addressed. This disruption in production cannot be absorbed by other companies in the event of a product with few competitors, and demand outpaces supply, resulting in a shortage. Picket Pharmaceutical is a generic drug manufacturer interested in leveraging data to better understand the generic drug industry, competition, and drug products at risk of shortage to meet market demand and provide essential medications to patients during the shortage period. All three primary objectives were addressed in this report. To address the first objective, we utilized a simple table to learn PAIN, ANTIBACTERIALS, ANTITHROMBOTICS, ONCOLOGICS, and HOSPITAL SOLUTIONS are the top five drug classes that have at least one shortage. To address the second objective, we implemented binary logistic regression to predict which generic drugs may experience a shortage in the future, which we forecast to be those with the Product IDs 59, 1671, and 3830. To address the third objective, we implemented Poisson regression to predict the number of times in shortage, on average, for each drug class. We were able to forecast the mean number of times each drug class would experience a shortage and determined the generic drugs belonging in the drug classes ANTI ULCERANTS, THYROID ANTI THYROID AND IODINE PREPS, OTHER HORMONES, ANTITHROMBOTICS, OTHER CARDIOVASCULARS, and BLOOD COAGULATION are likely to experience a shortage in the future.

2 Introduction

2.1 General Background

Generic drugs typically cost less than brand-name drugs but have identical active ingredients, administration methods, dosage forms, strengths, and indications. Their lower prices are the result of generic drugs being sold after their brand-name equivalents' patents have expired. The pharmaceutical industry continues to be highly competitive, with 70% of drug manufacturers having three to nine competitors. The number of competitors and rate of price reduction increase with product size. However, a substantial portion of the market has four or fewer competitors in the dermatological, injectable, and ophthalmic branches, which often leads to drug shortages. Drug shortages are a persistent issue in the United States, particularly for off-patent, injectable products with few suppliers. As a result, 90% of emergency room doctors routinely lack access to essential medications, 56% of hospitals postpone treatments due to drug shortages, and 47% of pharmacists report being compelled to use less potent drugs to treat patients. The majority of drug shortages are caused by manufacturing quality issues, which force a pause in production while the problem is addressed. This disruption in production cannot be absorbed by other companies in the event of a product with few competitors, and demand outpaces supply, resulting in a shortage. If a drug is produced exclusively by one company, there are no other options for production, therefore clinicians must either struggle to get a supply of it, compound it if they can, or suggest an alternative therapy if one is available. Picket Pharmaceutical is a generic drug manufacturer interested in leveraging data to better understand the generic drug industry, competition, and drug products at risk of shortage to meet market demand and provide essential medications to patients during the shortage period.

2.2 Objectives

This report will cover three primary objectives for this consultation:

1. Identify the top five generic drug classes that have experienced shortages.
2. Predict which generic drugs may experience a shortage.
3. Predict the number of times in shortage, on average, for each drug class.

3 Project Approach

3.1 Data Description

The client's raw data is composed of 5,231 observations and 615 columns. The 615 columns can be categorized into 22 subgroups:

- **Product ID**: Five-digit unique product number.
- **Ingredient**: Active pharmaceutical ingredient(s) for specific Product ID.
- **Form**: Dosage form of a specific product.
- **ParentCode**: Parent manufacturer code for specific Product ID.
- **Strength**: Concentration or strength of a specific product.
- **SpecPharm**: Specialty pharma or non-specialty pharma. Specialty pharmacies deliver medications with special handling, storage and distribution requirements.
- **CareClass**: Acute or chronic care classification.
- **DrugClass**: Drug classification.
- **ProductLaunchDate**: Month of product launch. If product was launched before Month0 (first month of sales data), then it will have a "_" before the month number.
- **PackSize**: Number of units per case.
- **PackQuantity**: Number of units in one pack.
- **NDC_Shortage_FL**: NDC (National Drug Code) had at least one shortage during five-year period. There was no shortage if the value is 0, and there was a shortage if the number is 1.
- **StDateX**: Start month of shortage X for a given product.
- **EnDateX**: End month of shortage X for a given product.
- **DurMnthsX**: Duration of shortage X in months (counts). The number "X" indicates the number of shortages that have occurred in the given period.
- **MonthX_RxSales**: Sum of sales for product grouping in month X.
- **MonthX_Units1**: Units (number of packages) summed by product in month X.

- **MonthX_Units2**: Units (bottles/vials) summed by product in month X.
- **MonthX_Units3**: Units (tablets/capsules/mL) summed by product in month X.
- **MonthX_WholesalePrice**: Median wholesale price for product grouping in month X, but does not include rebate offered to distributors.
- **MonthX_ManuPrice**: Median manufacturer price for product grouping in month X.
- **MonthX_MjManuCT**: Count of major manufacturers by Ingredient and Form in month X

3.2 Data Cleaning

We added a new column called ShortageFreq that keeps track of the number of times each generic drug has been in shortage. Furthermore, we created a subset of the data set called DrugShortage in which the values in the NDC_Shortage_FL column are equal to 1. Only generic drugs with at least one shortage during a five-year period will be included in this subset.

3.3 Methodologies

We used the DrugShortage subset to generate a table with the DrugClass column to identify the top five generic drug categories that have experienced shortages.

Then, we implemented logistic regression, a supervised machine learning model, to predict which generic drugs may experience a shortage. Based on a given data set of independent variables, logistic regression calculates the likelihood that an event will occur. A logit transformation is applied on the odds, the probability of success divided by the probability of failure, in logistic regression. The transformed odds are often referred to as log odds. In this model, the beta parameters, or coefficients, are often estimated via maximum likelihood estimation. This algorithm evaluates various beta values over a number of iterations to get the best match for the log odds. To determine the most accurate parameter estimate, logistic regression aims to maximize the log likelihood function, which is produced by all of these iterations. Following the identification of the best coefficients, the conditional probabilities for each observation can be computed, logged, and summed to provide a predicted probability. In the case of binary classification, a probability of less than 0.5 predicts 0 and a probability of more than or equal to 0.5 predicts 1. However, in some circumstances, it may be desired

to modify the threshold of 0.5.

Lastly, we implemented Poisson regression to predict the number of times each drug class has been in shortage. Poisson regression is a special case of the generalized linear model where the random component is defined by the Poisson distribution. Poisson regression typically performs well when the response variable is a count of some occurrence. Poisson regression is better suited in instances where the response variable is a small integer. Similar to logistic regression, the beta parameters are often estimated via maximum likelihood estimation.

4 Results

4.1 Table

We utilized the `table()` function to generate a table of the DrugClass column using the DrugShortage subset. The top five drug classes that have faced shortages are depicted in Figure 1. PAIN accounted for 27.648115% of the observations in the DrugShortage subset, followed by ANTIBACTERIALS (19.120287%), ANTITHROMBOTICS (7.001795%), ONCOLOGICS (6.912029%), and HOSPITAL SOLUTIONS (6.373429%).

PAIN	ANTIBACTERIALS	ANTITHROMBOTICS	ONCOLOGICS	HOSPITAL SOLUTIONS
27.648115	19.120287	7.001795	6.912029	6.373429

Figure 1: Top Five Drug Classes in Shortage

4.2 Binary Logistic Regression

To ensure consistency between the training and test splits, we first set a seed using `set.seed(7995)`. Then, we divided the data into a training set and a test set using an 80/20 split using the `sample.int()` function. The next step was to estimate the following four models, with NDC_Shortage_FL as the response, using the `glm()` function with the argument `family = binomial` and the training set:

1. $\hat{y} = \beta_0 + \hat{\beta}_1 Form + \hat{\beta}_2 SpecPharm + \hat{\beta}_3 CareClass + \hat{\beta}_4 DrugClass$
2. $\hat{y} = \beta_0 + \hat{\beta}_1 Form + \hat{\beta}_2 SpecPharm + \hat{\beta}_3 CareClass + \hat{\beta}_4 DrugClass + \hat{\beta}_5 PackSize$
3. $\hat{y} = \beta_0 + \hat{\beta}_1 Form + \hat{\beta}_2 SpecPharm + \hat{\beta}_3 CareClass + \hat{\beta}_4 DrugClass + \hat{\beta}_5 PackQuantity$
4. $\hat{y} = \beta_0 + \hat{\beta}_1 Form + \hat{\beta}_2 SpecPharm + \hat{\beta}_3 CareClass + \hat{\beta}_4 DrugClass + \hat{\beta}_5 PackSize + \hat{\beta}_6 PackQuantity$

Afterwards, we used the `predict()` function with the argument `type = response` to calculate the predicted probabilities of the NDC_Shortage_FL values in the test data set. Then, we created a confusion matrix using the `table()` function. A confusion matrix is a table that is used to summarize and visualize a classification algorithm's performance. As previously noted, in some cases, modifying the 0.5 threshold may be desirable. The threshold should be chosen based on the type of error we are most concerned with. In other words, we are more concerned about misclassifying generic drugs that may be in short supply. Therefore, we should lower our threshold to a value that both minimizes the misclassification of generic medications that might experience a shortage and does not significantly decrease accuracy. We made the decision to lower the threshold from 0.5 to 0.3. The confusion matrix is made up of four main properties (numbers) that are utilized to define the classifier's measuring metrics. These four numbers are:

1. True Positive (TP): The observation is predicted to be positive and is in fact positive.
2. False Positive (FP): The observation is predicted to be positive but is actually negative.
3. True Negative (TN): The observation is predicted to be negative and is in fact negative.
4. False Negative (FN): The observation is predicted to be negative but is actually positive.

Note: In our case, positive indicates being classified as 1, and negative indicates being classified as 0.

A confusion matrix is shown in Figure 2 as an example. In Figure 2, TP = 12, FP = 2, TN = 27, and FN = 9.

	FALSE	TRUE
0	27	2
1	9	12

Figure 2: Example of a Confusion Matrix

These four numbers can be used to calculate the accuracy of our binary logistic regression models.

$$Accuracy = \frac{TN + TP}{TN + FP + FN + TP} * 100 = \frac{27 + 12}{27 + 2 + 9 + 12} * 100 = 78\%$$

The computed accuracy for each model is shown in Figure 3. The first and second models has the best accuracy, and the third and fourth models has the worst accuracy.

	Model1	Model2	Model3	Model4
Accuracy	73.16141	73.16141	72.97039	72.97039

Figure 3: Accuracy of Each Model

A binary logistic regression model with an accuracy of 73.16141% is not ideal. As one of the assumptions of logistic regression, we opted to test the assumption that there is minimal to no multicollinearity between the independent variables using the `vif()` function on all the models. Figure 4 displays the output of the vif() function for each model. The GVIF indicates the increase in the variance of a regression coefficient due to its collinearity with the other variables. As we can see, for every model, the GVIF value for DrugClass is significantly greater than the GVIF values for the other variables. As a result, we decided to remove DrugClass from all of our models because of the assumption violation and to see if we might get an improvement in accuracy.

	GVIF	Df	GVIF^(1/(2*Df))
Form	1.745871	4	1.072140
SpecPharm	7.945472	1	2.818771
CareClass	4.285033	1	2.070032
DrugClass	54.841813	45	1.045499

	GVIF	Df	GVIF^(1/(2*Df))
Form	1.795517	4	1.075904
SpecPharm	7.953873	1	2.820261
CareClass	4.302196	1	2.074174
DrugClass	56.612546	45	1.045868
PackSize	1.066495	1	1.032712

	GVIF	Df	GVIF^(1/(2*Df))
Form	3.459747	4	1.167832
SpecPharm	7.975958	1	2.824174
CareClass	4.369926	1	2.090437
DrugClass	60.297870	45	1.046601
PackQuantity	2.141499	1	1.463386

	GVIF	Df	GVIF^(1/(2*Df))
Form	3.559706	4	1.171998
SpecPharm	7.985272	1	2.825822
CareClass	4.388907	1	2.094972
DrugClass	62.326174	45	1.046986
PackQuantity	2.144231	1	1.464319
PackSize	1.067018	1	1.032966

Figure 4: VIF Output for Each Logistic Regression Model

The computed accuracy for each new model is shown in Figure 5. After DrugClass was

dropped, all models had a noticeable improvement in accuracy. The first model has the best accuracy, and the third model has the worst accuracy. Furthermore, the accuracy of the first model is approximately 80%, which is ideal, so we are pleased with this model.

	Model1	Model2	Model3	Model4
Accuracy	79.84718	79.75167	79.17861	79.36963

Figure 5: Accuracy of Each Logistic Regression Model without DrugClass

We used the `which()` function to find the Product IDs of observations in the test set that were predicted to experience a shortage and actually experienced at least one during a five-year period. We discovered a future shortage is expected for the generic drugs with the Product IDs 159, 1671, and 3830.

4.3 Poisson Regression

To ensure consistency between the training and test splits, we first set a seed using `set.seed(7995)`. Then, we divided the data into a training set and a test set using an 80/20 split using the `sample.int()` function. The next step was to estimate the following four models, with ShortageFreq as the response, using the `glm()` function with the argument `family = poisson` and the training set:

1. $\hat{y} = \beta_0 + \hat{\beta}_1 Form + \hat{\beta}_2 SpecPharm + \hat{\beta}_3 CareClass + \hat{\beta}_4 DrugClass$
2. $\hat{y} = \beta_0 + \hat{\beta}_1 Form + \hat{\beta}_2 SpecPharm + \hat{\beta}_3 CareClass + \hat{\beta}_4 DrugClass + \hat{\beta}_5 PackSize$
3. $\hat{y} = \beta_0 + \hat{\beta}_1 Form + \hat{\beta}_2 SpecPharm + \hat{\beta}_3 CareClass + \hat{\beta}_4 DrugClass + \hat{\beta}_5 PackQuantity$
4. $\hat{y} = \beta_0 + \hat{\beta}_1 Form + \hat{\beta}_2 SpecPharm + \hat{\beta}_3 CareClass + \hat{\beta}_4 DrugClass + \hat{\beta}_5 PackSize + \hat{\beta}_6 PackQuantity$

Considering 3313 out of 4184 observations had a value of 0 in the ShortageFreq column, we only included observations in the training set where ShortageFreq > 0. If these 3313 observations are included in the models, there is a low possibility of producing a reasonably accurate model for predicting ShortageFreq for each drug class.

After estimating the models, we checked for multicollinearity using the `vif()` function. As we can see in Figure 6, for every model, the GVIF value for DrugClass and SpecPharm are significantly greater than the GVIF values for the other variables. As a result, we decided to remove DrugClass and SpecPharm from all of our models to avoid multicollinearity.

```
                GVIF Df GVIF^(1/(2*Df))
Form        1.839597  4        1.079171
SpecPharm  51.880714  1        7.202827
CareClass   5.663917  1        2.379899
DrugClass 504.404577 32        1.102125
                GVIF Df GVIF^(1/(2*Df))
Form        1.881074  4        1.082183
SpecPharm  51.967287  1        7.208834
CareClass   6.097921  1        2.469397
PackSize    1.437654  1        1.199022
DrugClass 600.534698 32        1.105133
                 GVIF Df GVIF^(1/(2*Df))
Form         3.518534  4        1.170295
SpecPharm   51.881072  1        7.202852
CareClass    5.760683  1        2.400142
PackQuantity 2.285796  1        1.511885
DrugClass  578.295597 32        1.104482
                 GVIF Df GVIF^(1/(2*Df))
Form         3.636571  4        1.175132
SpecPharm   51.969484  1        7.208986
CareClass    6.232132  1        2.496424
PackQuantity 2.317048  1        1.522185
PackSize     1.451298  1        1.204698
DrugClass  696.577134 32        1.107698
```

Figure 6: VIF Output for Each Poisson Regression Model

The mean squared errors (MSE) for each model are shown in Figure 7. Although the MSEs of the four models were very comparable, the fourth model had the lowest MSE. So, going forward, we employed the fourth model.

```
       Model1    Model2    Model3    Model4
MSE 0.365483 0.3630767 0.3654168 0.3630679
```

Figure 7: MSE for Each Poisson Regression Model without DrugClass and SpecPharm

We estimated the number of times each observation in the test set was in a shortage using the `predict()` function. After that, for each observation in the test set, we produced a data frame comprising of the predicted values and the appropriate drug class. The mean predicted number of times each drug class has experienced a shortage was then determined using the `tapply()` function. Figure 8 displays the mean number of shortages for each drug class. On average, ANTI ULCERANTS, THYROID ANTI THYROID AND IODINE PREPS, OTHER HORMONES, ANTITHROMBOTICS, OTHER CARDIOVASCULARS, and BLOOD COAGULATION appear to be in shortage the most.

ANTI ULCERANTS	OTHER CNS	ALLERGY SYSTEMIC NASAL
2.527590	2.514817	2.504702
THYROID ANTI THYROID AND IODINE PREPS	RESPIRATORY AGENTS	ANTIHYPERTENSIVES PLAIN COMBO
2.497689	2.453669	2.440316
OTHER HORMONES	ANTIVIRALS HERPES	MENTAL HEALTH
2.387504	2.370993	2.315896
ANTITHROMBOTICS	NERVOUS SYSTEM DISORDERS	HOSPITAL SOLUTIONS
2.287212	2.282182	2.267933
OTHER CARDIOVASCULARS	VITAMINS MINERALS	CORTICOSTEROIDS PLAIN COMBO
2.263012	2.219346	2.213079
BLOOD COAGULATION	ANTI PARASITICS ANTIMALARIALS INSECTICIDES	LABOUR INDUCERS
2.206453	2.202985	2.202985
ANTI ANAEMICS IRON AND ALL COMBINATIONS	PAIN	ANTIBACTERIALS
2.177390	2.166562	2.133950
MULTIPLE SCLEROSIS	OTHER ALIMENTARY TRACT AND METABOLISM PRODUCTS	CANCER DETOX AG ANTI NAUSEANTS
2.131903	2.125096	2.114416
OTHER THERAPEUTICS	IMMUNOSUPPRESSANTS	BISPHOSPHONATES TUMOR RELATED BONY METASTASES
2.114176	2.110092	2.102207
ONCOLOGICS	HYPOTHALAMIC HORMONES	ADHD
2.092648	2.079692	2.076011
POLYVAL IMMUNOGLOBLULINS IV IM	GASTRO PRODUCTS	IMAGING
2.076011	2.067490	2.065125
ANTIDIABETICS	OTHER HAEMATOLOGICALS	OSTEOPOROSIS
2.050149	2.050149	1.939524
SYSTEIMC ANTIFUNGALS	ANTICOAGULANTS	
1.896102	1.873566	

Figure 8: Mean Number of Shortages for Each Drug Class

5 Conclusions

In this report, we are able to adequately address the three primary objective. In response to the first objective, PAIN, ANTIBACTERIALS, ANTITHROMBOTICS, ONCOLOGICS, and HOSPITAL SOLUTIONS are the top five drug classes that have at least one shortage. In response to the second objective, the generic drugs with the Product IDs 59, 1671, and 3830 are predicted to experience a shortage in the future. In response to the third objective, we were able to forecast the mean number of times each drug class would experience a shortage. We found there would likely be a future shortage of generic drugs in the classes ANTI ULCERANTS, THYROID ANTI THYROID AND IODINE PREPS, OTHER HORMONES, ANTITHROMBOTICS, OTHER CARDIOVASCULARS, and BLOOD COAGULATION.