**Logo

Description automatically generated**

**Final Project – Pharmacy Claims**

**Data Warehousing and SQL**

Ayeshabi W Tigdikar

Master of Science in Project Management, Northeastern University

Professor Kayal Chandrasekaran

March 30th, 2024

**Table of Contents**

[Introduction 3](#_Toc162561518)

[Normalization 3](#_Toc162561519)

[Primary and Foreign Key 7](#_Toc162561520)

[Entity Relationship Diagram 10](#_Toc162561521)

[Analytics and Reporting 11](#_Toc162561522)

[Conclusion 13](#_Toc162561523)

# Introduction

The pharmaceutical industry generates vast amounts of data related to prescriptions, patients, and medications. Analyzing this data can provide valuable insights into trends, costs, and patient outcomes. In this project, we have explored a dataset containing information about prescriptions, including member details, drug information, fill dates, copay amounts, and insurance payments. Our goal is to gain insights into prescription patterns, identify the amount paid by insurance for the most recent prescription fill date, and analyze prescription data based on member age groups.

# Normalization

In the given dataset, the presence of recurring groupings of data for fill dates, copays, and insurance paid amounts prevents the dataset from being in 1NF (First Normal Form). To accomplish 1NF, these prescriptions—which each member may have—should be organized into distinct rows according to their fill dates, copays, and insurance paid amounts. To meet 1NF, rearrange the table as follows:

A screenshot of a computer

Description automatically generated

The dataset above is in 2NF (Second Normal Form). But the dataset is not in 3NF (Third Normal Form). A table in 3NF must first be in 2NF and guarantee that no transitive dependencies exist. When one non-prime property depends on another non-prime attribute that is not included in the primary key, this is known as a transitive dependency. The following transitive dependencies are visible when examining the dataset:  
member\_first\_name, member\_last\_name, member\_birth\_date, member\_age, and member\_gender are functionally dependent on member\_id.

drug\_name, drug\_form\_code, drug\_form\_desc, drug\_brand\_generic\_code, and drug\_brand\_generic\_desc are functionally dependent on drug\_ndc.  
The attributes are needed to be divided into distinct tables in order to normalize the dataset into 3NF.

There is 1 fact table and 4 dimension tables namely Dim\_Member\_Details\_Table, Dim\_Drugs\_Table, Dim\_DrugForm\_Table, Dim\_DrugGeneric\_Table, Fact\_Prescription\_Table.

**Dim\_Member\_Details\_Table**

A screenshot of a computer

Description automatically generated

**Dim\_Drugs\_Table**

**A screenshot of a computer

Description automatically generated**

**Dim\_DrugForm\_Table**

**A screenshot of a computer

Description automatically generated**

**Dim\_DrugGeneric\_Table**

**A screenshot of a computer

Description automatically generated**

**Fact\_Prescription\_Table**

A screenshot of a computer

Description automatically generated

The two fact variables in the provided dataset's fact table are **copay** and **insurancepaid**. These fact variables are **additive** and may be utilized with any aggregation function because they can be merged across dimensions (member\_id, drug\_ndc, and fill\_date) to generate informative data.

The prescription level represents the grain of the fact table. A member's prescription for a particular medication on a given fill date is represented by each fact row, together with the corresponding copay and insurancepaid amounts.

# Primary and Foreign Key

A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated

**Dim\_Member\_Details\_Table**

Primary Key: member\_id (**Natural Key**)

**Dim\_Drugs\_Table**

Primary Key: drug\_ndc (**Natural Key**)

**Dim\_DrugForm\_Table**

Primary Key: drug\_form\_code (**Natural Key**)

**Dim\_DrugGeneric\_Table**

Primary Key: drug\_brand\_generic\_code (**Natural Key**)

**Fact\_Prescription\_Table**

Primary Key: prescription\_id (**Surrogate Key**)

Foreign Keys: member\_id , drug\_ndc , drug\_form\_code, drug\_brand\_generic\_code

**Reference table and Foreign Keys**

* member\_id is the primary key in **dim\_member\_details** table
* drug\_ndc is the primary key in **dim\_drugs** table
* drug\_form\_code is the primary key in **dim\_drug\_form** table
* drug\_brand\_generic\_code is the primary key in **dim\_drugs\_generic** table

When a referenced row in the parent table is removed or altered in MySQL, the matching foreign key values in the child table are set to NULL. This is why I chose SET NULL for foreign key (FK) actions.

# Entity Relationship Diagram

A screenshot of a computer

Description automatically generated

One fact table, Fact\_Prescription\_Table, and four dimension tables, Dim\_Member\_Details\_Table, Dim\_Drugs\_Table, Dim\_DrugForm\_Table, and Dim\_DrugGeneric\_Table, make up the dataset's Entity-Relationship (ER) diagram. The dimension tables encircle the fact table, Fact\_Prescription\_Table, which is positioned in the center of the ER diagram. This arrangement creates a star schema, a popular data warehouse design in which several dimension tables are connected to a core fact table.

# Analytics and Reporting

1. **Number of Prescriptions grouped by drugname**

A screenshot of a computer

Description automatically generated

A screenshot of a computer

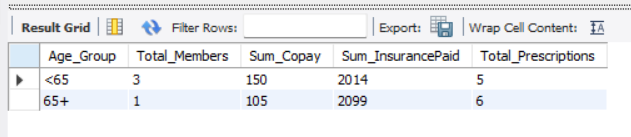
Description automatically generated

There were a total of 5 prescriptions for the drug AMBIEN.

1. **Count of Total Prescriptions, distinct members, sums copay and insurance paid for the Age group 65+ and <65.**

A screenshot of a computer code

Description automatically generated



According to the output, there is 1 unique member who is over the age of 65 years with a total of 6 prescriptions.

1. **Amount paid by the insurance for the most recent prescription fill date.**

A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated

Jane Doe whose member id is 10003, the drug name listed on the most recent fill date 16th May 2018 is AMBIEN. For that medication, the insurance has paid 322.

# Conclusion

Finally, we can say that by normalizing the dataset into 3NF, we were able to remove transitive dependencies and organize the data more effectively. Currently, a fact table and four dimension tables that are connected by main and foreign keys comprise the representation of the dataset.   
The data analysis has yielded insightful information about prescription pricing and trends. For instance, we discovered that the medication AMBIEN was prescribed five times in total and that insurance covered a sizable portion of the cost of prescriptions, especially for members who were 65 years of age or older. All things considered, this research shows how crucial data analysis and normalization are to drawing insightful conclusions from intricate datasets in the pharmaceutical sector.