



# HULT HUSTLE

**Final Deliverable**

**Final Deliverable**

**Team 9 Submission**

- Karthik Krishnamoorthi
- Maria Celeste Bustillo
- IK

PharmaCo is pioneering with the concept of decentralized pharmacy bringing the customers, doctors, and the pharmacy under one platform. The company is also investing in drug research and development. By running analytics on the datasets provided, we have come up with strategies which can help in decision making in areas of product development and marketing.

We (Team 9) kickstarted this project with three objectives of helping Pharmaco make ideal decisions in the following

- RND initiative on where to investment their money when it comes to product development.
- Measuring efficacy of the drug at the drug group basis (Type 2 Diabetes – their core product)
- Understanding customer demographics (age, group, medical specialty) to gain insights on the market.

Our entire code is uploaded and available on github.com (<https://github.com/kartskri/pubmed-r>)

## Objective 1

Objective	Product Development – Identify the trending Research Areas in Medicine and Diseases
Dataset	PubMed XML Files Provided (PubMed Case Report Set)
Method	<b>Project Steps</b> <ul style="list-style-type: none"> <li>- Downloaded the PubMed files to a local directory and ran a process to read each file at a time and parse the xml file</li> <li>- Read parsed contents of the XML to a R data frame and saved to an intermediary csv file</li> <li>- Recreated a Data frame from the csv intermediary file and ran LDA algorithm to find trending words and created 3 topics based on Model findings</li> </ul>
Inference	<p>Based on the data analysis, we tried to infer trending journals via most used keywords and topics. Following are our research finding</p> <ul style="list-style-type: none"> <li>- Major areas of research were based on the areas of Cancer, Tumor research or in the areas of pulmonary domain.</li> <li>- Majority of the research were focused on acute disease conditions and based on clinical trials</li> </ul> <p>From the Topics (that we were able to isolate from the PubMed database), we have the following inferences</p> <ul style="list-style-type: none"> <li>- First Topic categorizes the disease (Carcinoma, Cancer, Lung Disease)</li> <li>- Second Topic categorizes the method of the disease and how to discover it (clinical, diagnosis, congenital)</li> <li>- Third Topic talks on the management of the disease (treatment, surgery, valve)</li> </ul> <p>So, we believe this model and topic selection can give great insights to the business on the areas where they could invest in drug research, development, or treatment methodologies.</p>
Code	Data Preparation 1_pubmed_parser.Rmd Model Execution 1_pubmed_text_analyzer.Rmd
Visualizations	<a href="https://raw.githubusercontent.com/kartskri/pubmed-r/main/graphs/1-pubmed-topics.png">https://raw.githubusercontent.com/kartskri/pubmed-r/main/graphs/1-pubmed-topics.png</a> <a href="https://raw.githubusercontent.com/kartskri/pubmed-r/main/graphs/1-pubmed-trending-words.png">https://raw.githubusercontent.com/kartskri/pubmed-r/main/graphs/1-pubmed-trending-words.png</a>

## Objective 2

Objective	Drug Group Analysis
Why?	Drug Group is the drug formulation behind any brand name drug. In our case, analyzing the drug group is critical as it can help Pharmaco Corp (Decentralized Pharmacy) to measure the efficacy of them in real world scenario and also in invest in generic drugs to lower costs
Method	<p>The dataset contains <b>21 medications</b> (Insulin, Metformin, Glimepiride, ...) however they can be categorized under the following 8 drug groups (Biguanides, Meglitinides, Sulfonylureas, Thiazolidinediones, Alpha-glucosidase, gliptins).</p> <p>We adopted the following methodology for our analysis</p> <ul style="list-style-type: none"><li>i) Ran data analytics over the drug groups, tried to find the distribution of each drug group</li><li>ii) Shortlist 2 top used drugs groups</li><li>iii) Run Efficacy test on top 2 drug groups based on “readmission” field</li></ul>
Inference	<ul style="list-style-type: none"><li>1. Based on our analysis “Biguanides and Sulfonylureas” were by far the most used drug groups in treating type 2 diabetes</li><li>2. Efficacy test for Drug Group based on “Readmission Rate” Data field<ul style="list-style-type: none"><li>- Biguanides fared better at 57%</li><li>- Sulfonylureas value was at 53%</li></ul></li></ul>
Code	2_drug_analyzer.Rmd
Visualizations	<p>Distribution Graphs</p> <p><a href="https://raw.githubusercontent.com/kartskri/pubmed-r/main/graphs/2-sulfonylureas-dist.png">https://raw.githubusercontent.com/kartskri/pubmed-r/main/graphs/2-sulfonylureas-dist.png</a></p> <p><a href="https://raw.githubusercontent.com/kartskri/pubmed-r/main/graphs/2-thiazolidinediones-dist.png">https://raw.githubusercontent.com/kartskri/pubmed-r/main/graphs/2-thiazolidinediones-dist.png</a></p> <p><a href="https://raw.githubusercontent.com/kartskri/pubmed-r/main/graphs/2-alphaglucoasidase-dist.png">https://raw.githubusercontent.com/kartskri/pubmed-r/main/graphs/2-alphaglucoasidase-dist.png</a></p> <p><a href="https://raw.githubusercontent.com/kartskri/pubmed-r/main/graphs/2-biguanides-dist.png">https://raw.githubusercontent.com/kartskri/pubmed-r/main/graphs/2-biguanides-dist.png</a></p> <p><a href="https://raw.githubusercontent.com/kartskri/pubmed-r/main/graphs/2-meglitinides-dist.png">https://raw.githubusercontent.com/kartskri/pubmed-r/main/graphs/2-meglitinides-dist.png</a></p> <p>Efficacy Analysis Graphs</p> <p><a href="https://raw.githubusercontent.com/kartskri/pubmed-r/main/graphs/2-biguanides-readmission.png">https://raw.githubusercontent.com/kartskri/pubmed-r/main/graphs/2-biguanides-readmission.png</a></p> <p><a href="https://raw.githubusercontent.com/kartskri/pubmed-r/main/graphs/2-sulfonylureas-readmission.png">https://raw.githubusercontent.com/kartskri/pubmed-r/main/graphs/2-sulfonylureas-readmission.png</a></p>

## Target Customers

Objective	Market Analysis – Age, Gender, Admission History, Medical Specialty Analysis
Source	Diabetes Dataset – Excel file
Feature Set	Demographic Data, Hospitalization Record, Medicine Usage and Efficacy Measure
Method	“Data Analytics” Over Diabetes Dataset “Decentralized pharmacy” is the core business of Pharmaco, and it is imperative for them to reach out and target the right kind of audience. They are market leaders in Type 2 Diabetes drugs. To acquire more customers, we present few insights based on our dataset on customer demographics based on the available dataset. Admission Source and Type analysis were done combining the data dictionary with the original dataset.
Inference	<ul style="list-style-type: none"><li>- Gender is almost evenly distributed over Age till 60-70 band after which female patients are more in number compared with male counterparts.</li><li>- Emergency room was top Admission Source followed by Physician Referral</li><li>- Emergency also topped admission type followed by elective and urgent care</li><li>- Newborn Admission Type number was significant number suggesting wider prevalence of “Gestational Diabetes”</li><li>- Internal Medicine, Family / General Practice, Emergency/Trauma, Cardiology are the four top medical specialty groups when it comes diabetic referrals.</li></ul>
Code	3_diabetes_demographics_visualizer.Rmd
Visualization	<a href="https://raw.githubusercontent.com/kartskri/pubmed-r/main/graphs/3-admission-source.png">https://raw.githubusercontent.com/kartskri/pubmed-r/main/graphs/3-admission-source.png</a> <a href="https://raw.githubusercontent.com/kartskri/pubmed-r/main/graphs/3-age-gender-distribution.png">https://raw.githubusercontent.com/kartskri/pubmed-r/main/graphs/3-age-gender-distribution.png</a> <a href="https://raw.githubusercontent.com/kartskri/pubmed-r/main/graphs/3-demographics-admissiontype.png">https://raw.githubusercontent.com/kartskri/pubmed-r/main/graphs/3-demographics-admissiontype.png</a> <a href="https://raw.githubusercontent.com/kartskri/pubmed-r/main/graphs/3-medical-speciality-distribution.png">https://raw.githubusercontent.com/kartskri/pubmed-r/main/graphs/3-medical-speciality-distribution.png</a>

## Conclusion

We believe we were able to cover the company’s requirements at the macro level. However, we believe additional research on PubMed data can deliver huge insights on many levels. However, additional research on PubMed would require a combination of in-depth analytics and web scraping of the website pubmed.com based on clearly defined requirements. However, we believe the information we have provided is great place to start as we have identified the following information

- Objective 1 – PubMed Analysis – Identify areas of research and paper authors could be great to partner for research and development.
- Objective 2 – Drug Group Analysis – Great for investing in “Generics” and lower cost of establish drugs
- Objective 3 – Demographical Analysis - Identify medical professionals to target when spreading the word on drugs. Insights on customer demographics, Admission sources and types can also be very insightful from a marketing standpoint.