Capstone Project

Mei Eisenbach

Introduction

Background

Telesofia is a provider of web-based videos for patients. Their platform allows healthcare providers to automatically generate branded personalized educational videos. The videos are tailored to the specific patient, easy to understand, and available on devices with no additional software installation. They can be used to explain proper use of medication, direct preparations for medical procedures, or provide discharge instructions, among others.[1]

Problem Statement

Telesofia would like to identify potential clients in pharmaceutical companies. They have determined that specialty drugs have the highest potential for revenue and would like to concentrate on companies that manufacture those drugs. Specialty drugs are defined as drugs with high cost, high complexity, and/or high touch. Complexity refers to the difficulty in manufacturing the drug; often they are biologics (drugs derived from living cells). High touch means that the administration of the drug often requires ongoing support from a medical professional.[2]

Telesofia indicated that they were interested in specialty drugs with the following characteristics: [REDACTED]

Data Dictionary

A data dictionary was created using the fields from the sample report that Telesofia provided. This document defined what data would be delivered and was very helpful for ensuring that the team and Telesofia were on the same page when discussing the project. The data dictionary was updated periodically and served as a progress report to the company.

Data dictionary fields

- Field number - The field number

[1] https://www.telesofia.com/about
[2] https://en.wikipedia.org/wiki/Specialty_drugs_in_the_United_States

- % found - The percentage of drugs where there is data for the field
- Name - The name of the field (e.g. Therapeutic areas)
- Description - The description of the field (e.g. The category of the specialty drug)
- Source - Where the data for the field comes from (e.g. Accredo Specialty Drug List)
- Comments - Comments such as “not provided” or “new”

The final data dictionary is attached as Appendix A.

Data sources and processing

Accredo Specialty Drug List

Format: PDF file

Accredo is a pharmacy that focuses on specialty drugs. They publish a list of specialty drugs and indicate which ones they distribute. This file was provided to us by Telesofia.

The fields acquired from this file are: therapeutic area, brand name, and generic name. The list was minimally modified, and a few duplicate drug names were merged. 468 brand names were found, which were used as the unique identifier for our dataset.

FDA

OpenFDA

Source: https://open.fda.gov/
Input format: JSON
Output: CSV
Tools: Python, R

OpenFDA provides APIs and raw download access to several high-value, high priority and scalable structured datasets, including adverse events, drug product labeling, and recall enforcement reports.[3] A python script was written to retrieve the records via the API using the brand name from the specialty drug list.

This data source had major shortcomings: Only 85% of the drugs from the Accredo list were retrieved, and the dosage, indications, and administration fields are unstructured text blocks. Attempts were made to extract information from the dosage field using regular expressions but these were mostly unsuccesful.

[3] https://open.fda.gov/about/

Processing

Using an R script, a word count of the protocol text was calculated (stop words were removed). The assumption is that the longer the text, the more complex the administration protocol is. The count was used to generate the protocol complexity field.

FDA National Drug Code (NDC) Directory

Source: https://www.fda.gov/Drugs/InformationOnDrugs/ucm142438.htm
Input format: Excel file
Output format: CSV?
Tools: SAS

Drug companies are required to provide the Food and Drug Administration (FDA) with a current list of all drugs manufactured, prepared, propagated, compounded, or processed for commercial distribution.[4] The FDA provides this information in via a web search page or a downloadable file. The zipped Excel file that was downloaded contained all drugs.

Processing:

- Filter using the list of specialty drugs
- Merge rows
- Generate new fields and output

Filtering

The NDC file has one record per ProductID (labeler code and product code segments of the National Drug Code number, separated by a hyphen followed by packaging ID). This meant that a drug name could be in multiple rows.

The first task was to filter the rows for only the drugs on the specialty drug list. The full file contains over 115,000 records. The NDC file splits the proprietary name into the name and the name suffix (e.g. Aralast NP → Aralast, NP). Sometimes the parts of the name were in a different order. (e.g. Acthar H.P. Gel → H.P. Acthar). A lookup field was added to translate between the Accredo list and the NDC file. For most drugs, this was systematically generated but there were special cases that needed to be added by hand.

Merging

[4] https://www.fda.gov/Drugs/InformationOnDrugs/ucm142438.htm

Multiple rows for a drug were combined into one. Unique entries in each field were combined with a semicolon as a separator.

e.g. Labeler: GLAXOSMITHKLINE LLC;NOVARTIS PHARMACEUTICALS CORPORATION

Generated fields

The dosage field needed to be combined with strength and unit fields.

Dosage form name	Strength	Unit
INJECTION, SOLUTION, CONCENTRATE	20	mg/mL
INJECTION, SOLUTION	180	mg/mL

Output: INJECTION, SOLUTION 180 MG/ML;INJECTION, SOLUTION, CONCENTRATE 20 MG/ML

This process was complicated by the fact that occasionally the strength field would have multiple entries.

For each field where there were multiple entries, a new count field was added. For example, for the dosage field, there is also a dosage count field.

Medicare

Source: https://www.cms.gov/Research-Statistics-Data-and-Systems/Statistics-Trends-and-Reports/Information-on-Prescription-Drugs/2015MedicareData.html
Format: Excel file
Tools: Excel

Unsuccessful attempts were made to acquire drug price and revenue information from Drugs.com and SEC filings from drug companies. In lieu of that information, Medicare Part D (Medicare prescription drug benefit) data was used.

Processing

Like the FDA NDC data, the Medicare data also needed a name translation field to match the rows with the Accredo brand names.

There were a few rows with the same name but different data. For the numeric fields, the data from multiple rows were averaged. Count fields were summed.

Output and QA

Delivered Data

Data from all sources were loaded as tables in an Access database. This ensured that the keys (brand name) were unique and allowed the tables to be joined for the output file. The output format was an Excel spreadsheet, as requested by Telesofia. The data acquisition and processing were geared towards producing a flat single file with one drug per row.

Quality Assurance

At various points in the project, quality assurance was performed. Work produced by one team member was checked by a different team member for accuracy and recall rate. Where there were only a small number of rows, all the data was checked. Otherwise, a 10-20% sample was checked. This was instrumental in discovering brand name inconsistencies between data sources.

Data Analysis

[REDACTED]

Rankings Report

A customizable Excel spreadsheet for ranking potential drugs was delivered to Telesofia. This spreadsheet consisted of the fields relevant for ranking, the ranking score for that field, the weights for the composite rank, and the composite rank. When the weights for the composite rank are changed and the composite rank is automatically re-calculated. The user can then re-sort the drugs by the composite rank.

Data processing

Additional data processing was required for the report. Missing or zero values were replaced with appropriate values. The replacement values were chosen based on common sense (e.g. all drugs have at least one pharmaceutical manufacturer) or to not affect the ranking (e.g. using the average).

[REDACTED]

Conditional Formatting

The original field values were conditionally formatted with color.

- For fields where the values started at zero, a white to green color gradient (white = 0).
- For fields which included negative numbers, a red to green gradient was used.
- In the Total Spending field, cells where a missing value had been replaced with the median were colored yellow.
- Start marketing dates before 1/1/2010 were colored red.

Rankings

- Individual rankings for each variable (except the Start marketing date) were calculated using the percentile function in Excel, PERCENTRANK.INC().
- The ranking score for Start Marketing Date was assigned a 0 for dates before 1/1/2010 and a 1 for all other dates.

- The composite rank is weighted average computed using the weights and the individual scores.

Results

[REDACTED]

Conclusion

Telesofia was pleased with our work. We were able to give them most of the data that they requested and analytic insights about the data. The ranking spreadsheet gives them a flexible way to examine the results.

Due to time constraints and Telesofia's requirements for the project, there were some areas where more research could have been done:

- Comparison of specialty drugs with non-specialty drugs to confirm specialty drug characteristics and to find candidates which are not on the Accredo list.
- Segmentation of specialty drugs
- Segmentation of drug companies (size, revenue, number of drugs manufactured)
- Extraction of the protocols from the protocol text using NLP techniques for semi-structured text.

Acknowledgements

Thank you,Telesofia, for working with our team on your business problem. I would also like to thank Professor Sarnikar for his support and guidance. And finally, thanks to the project team for all time and effort they put in.

Telesofia
- Natalie Levy
- Ehud Belder

Faculty Advisor: Surendra Sarnikar

Project team members
- Meghana Hulsure
- Devansh Pandit
- Yiping Wang

Appendix A: Data Dictionary

[REDACTED]

Appendix B: Additional Charts

[REDACTED]