A REPORT

ON

DEVELOPMENT OF AN ALGORITHM TO IDENTIFY INDICATIONS BASED ON TRANSACTION-LEVEL DATA

BY

Sandeep Venkata Kollipara
2013B1A10916G

AT

IQVIA, Inc.

A Practice School – II station of

BIRLA INSTITUTE OF TECHNOLOGY AND SCIENCE, PILANI

10th June, 2019

A REPORT

ON

DEVELOPMENT OF AN ALGORITHM TO IDENTIFY INDICATIONS BASED ON TRANSACTION-LEVEL DATA

BY

Sandeep V. Kollipara	2013B1A10916G	M.Sc.Bio+B.E.Chemical

Prepared in partial fulfillment of the
Practice School – II Course
BITS F412

AT

IQVIA, Inc.

A Practice School – II station of

BIRLA INSTITUTE OF TECHNOLOGY AND SCIENCE, PILANI

10th June, 2019

ACKNOWLEDGEMENT

I wish to express my gratitude to my mentors Mr. Vijay Viswanathan, Senior Consultant and Ms. Mallika Kowshik, Associate Consultant at Real World Analytics Solutions (RWAS) for assigning a project related to my discipline and assisting in getting me acclimatized to the realm of Analytics in Human Data science.

I would like to extend my thanks to Ms. Karanpreet Kaur and Ms. Faizia Arsheen of the HR department for appointing me as an intern at Real World Analytics Solutions (RWAS).

I would also like to thank my Practice School – II Instructor Dr. R. Bharathi for supporting me through this wonderful opportunity of a project. I'd like to end by sending my warmest regards to all the colleagues at Real World Analytics Solutions (RWAS) team of IQVIA, Inc, Bangalore.

BIRLA INSTITUTE OF TECHNOLOGY AND SCIENCE
PILANI (RAJASTHAN)
Practice School Division

Station:	IQVIA, Inc. **Centre:** Bangalore
Duration:	23rd January– 18th June, 2019 **Date of Start:** 23rd January, 2019
Date of Submission:	10th June, 2018
Title of the Project:	Development of an algorithm to identify indications based on transaction-level data
ID No./Name(s)/ Discipline(s)/of the student(s):	2013B1A10916G/ Sandeep V. Kollipara/ M.Sc.Bio+B.E.Chemical
Name(s) and Designation(s) of the expert(s):	Vijay Viswanathan, Senior Consultant Mallika Kowshik, Assoc Consultant
Name(s) of the PS Faculty:	Dr. R. Bharathi, Lecturer, BITS Pilani
Key Words:	Automated, Python, Transaction-level, Data, Analytics
Project Areas:	Real World Analytics

Abstract:

This project utilizes analytics on the lucrative biologic market for autoimmune diseases. The prime objective is to develop an algorithm to identify indications based on transaction-level data with the functionality to update with the evolving biologic market and requiring minimal manual intervention from the analyst. The program is built on Python and utilizes an algorithm based on a previous model in SAS and adapts around including additional features that replace the human involvement in the analysis process. The algorithm is divided into 4 phases based on different levels of analysis,

sub-objectives and complexity. Post development, the datasets generated by the new program are validated against the pre-existing SAS datasets. The automation aspect is realized by allowing the user to update specs and criteria beforehand instead of hardcoding them within the script. The final dataset obtained has six additional columns per prescription with the details of doctor specialization, indication, therapy duration etc. This dataset is then scheduled to be run in IQVIA's cockpit software made ready for client use. Being open source, the program reduces the investment on licensed software by the company while suffering no drawbacks in performance. This program can be used for developing programs for other markets as well. A secondary project in Python for 'Targeting Best Practices' was assigned. It primarily involves ranking the doctors on market sales off their prescriptions while adhering to privacy protection standards by using variables generated indirectly from transaction-level data.

Signature(s) of Student(s) **Signature of PS Faculty**

Date **Date**

BIRLA INSTITUTE OF TECHNOLOGY AND SCIENCE
PILANI (RAJASTHAN)

Practice School Division

Response Option Sheet

Station: IQVIA, Inc. **Centre:** Bangalore

ID. No. & Name(s): 2013B1A10916G & Sandeep Venkata Kollipara

Title of the Project: Development of an algorithm to identify indications based on transaction-level data

Usefulness of the project to the on-campus courses of study in various disciplines. Project should be scrutinized keeping in view the following response options. Write Course No. and Course Name against the option under which the project comes.

Refer Bulletin for Course No. and course Name.

Code No.	Response Options	Course No. (s) & Name
1.	A new course can be designed out of this project	No
2.	The Project can help modification of the course content of Some of the existing Courses	No
3.	The Project can be used directly in some of the existing Compulsory Discipline Courses (CDC)/Discipline Courses Other than Compulsory (DCOC)/ Emerging Area (EA) etc. courses	No
4.	The Project can be used in Preparatory courses like Analysis and Application Oriented Courses (AAOC)/Engineering Science (ES)/Technical Art (TA) and Core Courses	No
5.	This Project cannot come under any of the above mentioned options as it relates to the professional work of the host organization	Yes

_____ _____

Signature of Student Signature of Faculty

Contents:

1. Organization

IMS Health is an American company founded in 1954 by Bill Frohlich and David Dubow. It went private under TPG Capital, CPP Investment Board and Leonard Green & Partners. In 2014, the company went public and started trading in NYSE under the symbol IMS. IMS Health underwent a $17.6 billion merger with Quintiles in 2016 to form QuintilesIMS. In 2017, the company changed its name to IQVIA and changed its symbol to Q in NYSE.

IQVIA is the leading provider in global information and technology services with clients in the Healthcare industry and providing comprehensive solutions to measure and improve their performance. At IQVIA, human data science is applied:

1. by leveraging the analytic rigor and clarity of data science to the ever-expanding scope of human science
2. to enable companies to reimagine and develop new approaches to clinical development and commercialization
3. speed innovation and
4. accelerate improvements in healthcare outcomes.

IQVIA has one of the largest and most comprehensive collections of healthcare information in the world, which includes more than 530 million comprehensive, longitudinal, non-identified patient records spanning sales, prescription and promotional data, medical claims, electronic medical records and social media. Their scaled and growing data set contains approximately 30 petabytes of proprietary data sourced from more than 120,000 data suppliers and covering over 900,000 data feeds globally.

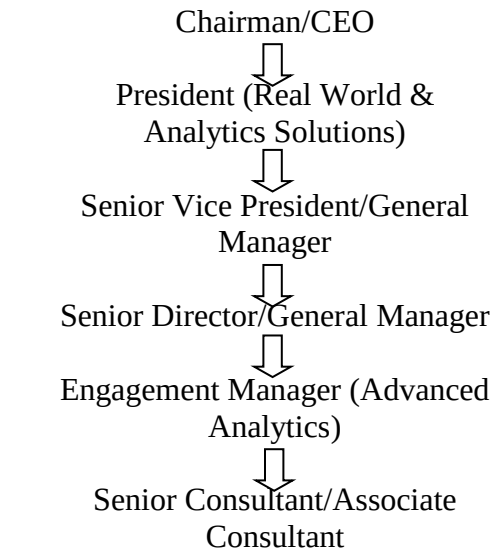
"We are IQVIA™. Our vision is to outpace the inevitable progress of change across the life sciences and accelerate our ability to empower healthcare decision makers to meet the future head on. We bring the future to clients through Human Data Science. We provide solutions that help our clients innovate with confidence, maximize opportunities and, ultimately, drive healthcare forward."

IQVIA, based on the data, delivers information and insights on 85% of the world's pharmaceuticals, as measured by sales revenue in 2016. IQVIA standardizes, organizes, structures and integrates this data by applying their sophisticated analytics and leveraging its global technology infrastructure to help their clients run their organizations more efficiently and make better decisions to improve their operational and financial performance.

With more than 55,000 employees including approximately 19,000 Commercial Services employees, approximately 29,000 Research & Development Solutions employees and approximately 7,000 Integrated Engagement Services employees, IQVIA conducts operations in more than 100 countries.

Organization Structure

Hierarchy:



Business Divisions/Domains:

a. Business Integration
b. Centers of Excellence
c. Chief Information Office
d. Chief Medical and Scientific Office
e. Ethics and Compliance Office
f. Finance
g. Legal Office of General Counsel
h. Global Services
    i. Commercial Analytics
    ii. Commercial Outsourcing Solutions
    iii. Consulting Services
    iv. Primary Intelligence
i. Go-To Market, Sales and Customer Relations
j. Information and Technology Client Solutions
k. Insurance Certificates (Business)
l. Biotech
m. Contract Sales and Medical Solutions
n. Quality Assurance
o. Real World and Analytics Solutions
p. Research and Development Solutions
q. Strategy, Marketing and Communications
r. Business Operations
s. Global Environment Health and Safety

Important Personnel:

- Chairman & CEO: Ari Bousbib
- President (Real World Analytics & Solutions): Jon Resnick
- Senior Vice President & General Manager: Prashant Parab
- Senior Director & General Manager: Jaivardhan Iyer
- Engagement Manager(Advanced Analytics): Sunil Kumar Singh

2. Project Plan:

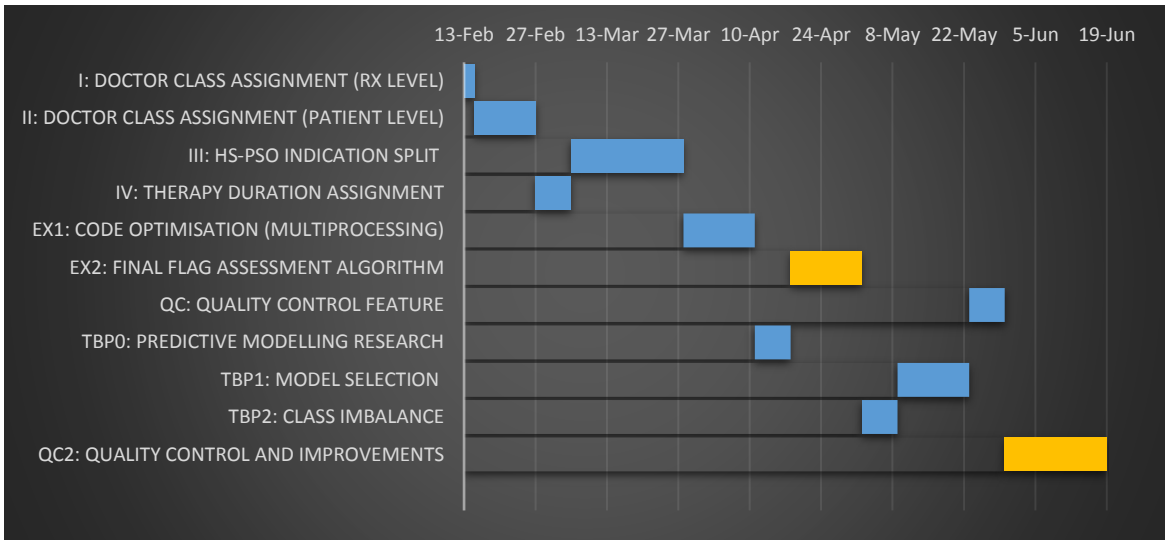


Figure 1: Gannt Chart of the project timeline with blue denoting completed and yellow denoting currently underway

The PS-II internship involved 2 different projects: Autoimmune Biologics Market and Targeting Best Practices both of which independently focus on the patients and the doctors respectively in relation to the drug they are prescribed/prescribing.

Autoimmune Biologics Market program analyses patients 'on drug' to determine what indication they are being prescribed the biologic for. The base dataset for the analysis in this program consists of all biologic transactions made up until the latest quarter. The transaction-level data consists of patient ID, drug name, molecule name, pharmacy ID, prescribed doctor's ID (encrypted), (drug) pack name, doctor's specialty, transaction date and month, pack size, strength per unit (mg/unit) and number of units. The dataset is analyzed for patterns of transactions and dosage to cross-evaluate against known therapies for indications.

Initially, the project spanned 6 weeks from 4th Feb to 20th March with the first week spent training on Python and learning the data structures but was later extended to

optimize the performance of the code and to develop an algorithm and incorporate its additional functionality.

The aim of the project is to develop an algorithm to identify indication based on transaction-level data with robustness to update. The project consists of different phases, namely:

I. Prescription-level Assignment of Doctor Specialization
II. Patient-level Assignment of Doctor Specialization and Classification
III. Indication Split: Dermatology patients into Hidradenitis Suppurative (HS) and Psoriasis (PSO).
IV. Therapy Duration Calculation per dosage unit and Assignment.

EX1. Code Optimization.
EX2. Final-Flag Assessment Algorithm.
QC. Feature to conduct Quality Control.

The initial 2 phases require fundamental understanding of the dataset and importing them into python environment to access, search and modify for analysis. The SAS model must be deciphered alongside development of the code especially in the 3rd Phase where SQL server access is to be implemented. The final phase involves calculating the therapy duration spec based on multiple criteria from different levels of the dataset.

The automation aspect of the program is to be implemented in the 2nd and 4th phases where specific criteria are imported (input in an excel file beforehand) to classify data which change overtime with the evolution of biologics market.

The Targeting Best Practices program is basically a model that predicts the sales of a drug by a doctor and ranks them on the basis of attributes derived from the doctor's bio data, sales history, specialization and other transaction-level data.

The aim of the project is to develop a model that accurately predicts and ranks doctors at 80% accuracy and quality. The following are the different phases of the project:

TBP0. Study Predictive Modelling and its implementation in Python.
TBP1. Model Selection for the program.
TBP2. Class Imbalance problem and its various solutions.
QC2. Quality Control Check for fine-tuning of the model.

This program requires implementation knowledge of special Python packages along with trials on model selection for best accuracy. The QC stage is still underway at the time of writing of this report and changes are due.

3. Literature survey

3.1 Biologic Market and Autoimmune Diseases:

The market in focus is of Biologic Therapies for Autoimmune diseases. A Biologic is a product extracted or semi-synthesized from living organism most of them manufactured utilizing Recombinant DNA Technology. Their structure is not characterized and are significantly larger in size and complicated compared to drugs. Delivery systems are primarily in parenteral route, they are injected or infused directly in the target region or released into the vein.

"The process is the product" aptly emphasizes the quality control restraints and effectiveness of the molecule as manufacturing is concerned. Biologics market requires heavy nascent investments. They also impact economically, for e.g., Rheumatoid

Arthritis, Drug therapy like methotrexate of cost is less than $100 and its counterpart biologics certolizumab and abatacept costs $300-6000.

The pros brought by biologics are:

- High selectivity in action,
- potent therapeutic efficiency and
- limited side effects.

The risks/drawbacks entailing biologics are:

- High cost,
- long term use increasing risk of cancer and
- less experience in clinical field

Biologics being a lucrative market and Autoimmune diseases being lifestyle-affecting to fatal, the prescriptions and injections are taken strictly on time thereby making this market analysis most accurate.

The biologics for autoimmune diseases are broadly classified into 4 categories which are explained as follows:

TNF Inhibitors:

Examples of these drugs certolizumab (*Cimzia*), etanercept, golimumab (*Simponi*), adalimumab (*Humira*) and infliximab (*Remicade*). The reduce inflammation and can be used in combinations of 2-3 doses. All of them have been clinically approved for children and certolizumab also for pregnant women.

IL Inhibitors:

Examples include anakinra (*Kineret*)[IL-1 inhibitor], tocilizumab (*Actemra*)[IL-6 inhibitor], canakinumab (*Ilaris*), secukinumab (*Cosentyx*) and ustekinumab (*Stelara*).

They are used after TNF inhibitors known to be effective and safe. Rarely bowel perforations are seen in some cases.

B-cell Inhibitors

Examples include belimumab (*Benlysta*) and rituximab (*Rituxan*). They interfere with production of abnormal antibodies (produced from B-cells. About 2 infusions are conducted in a year with relative long-term safety. Risks include blood pressure changes, chest pain, difficulty breathing, rash, dizziness and/or flu-like symptoms which need additional control medications.

T-cell Inhibitors

Some of them are also known as Selective Co-stimulation Modulators with examples like Abatacept (*Orencia*). Effects are not seen until 4-6 weeks after treatment and is more effective when used in combination with other common drugs. Risks involve being susceptible to infections such as pneumonia, tuberculosis and influenza.

3.2 Common Autoimmune Diseases:

Rheumatoid Arthritis: Inflammation of Joint Linings. All TNF inhibitors and adalimumab-atto (*Amjevita*) a biosimilar to Humira can be prescribed.

Multiple Sclerosis: Immune system attacks myelin sheath and deteriorates nerves with potential permanent damage. It affects limbs and the root cause is unknown. Biologic natalizumab (*Tysabri*) by Novartis in the market with Avonex (*Interferon B*) & Capaxone (*glatiramer acetate*) drugs (non-biologics).

Lupus: Immune system attacks tissues and organs and its severity can cause permanent tissue damage. Biologic belimumab (*Benlysta*) is the first approved biologic for Lupus

treatment. It can also be used alongside other immunosuppressive drugs (non-biologics.

Psoriasis: Immune reaction on skin resulting in scale formation and discoloration. Most TNF inhibitors like ustekinumab (*Stelara*) [IL-12/23 inhibitor], secukinumab (*Cosentyx*) and ixekizumab (*Taltz*) [IL-17 inhibitors], guselkumab (*Tremfya*) and tildrakizumab-asmn (*Ilumya*) [T-cell inhibitor].

Type-I Diabetes: Immune system destroys pancreatic beta cells completely leaving the victim unable to secrete insulin for metabolism. Common daily insulin injections are taken by the patients. Insulin by itself is a biological compound albeit regulated as a common drug.

Inflammatory Bowel Disease: Crohn's disease and ulcerative colitis come under this. Most TNF inhibitors can be prescribed and biologics ustekinumab (only for Crohn's disease) and vedolizumab (for both diseases) are new.

3.3 The Best Practices and Privacy:

IQVIA, bound by the right to privacy, protects the doctors' identities by encrypting their IDs and grouping them in batches while deriving the attributes for the Targeting Best Practices model. The data protection setup utilized in IQVIA prevents the onshore team from accessing the transaction-level data (patient data) while preventing the offshore team from accessing the means of identification of doctors. To that end, the program must be designed to predict close to 80% accuracy (correctly predicted doctors out of actual doctors of the class) and 80% quality (correctly predicted doctors out of total predicted doctors of the class) for maintaining a level of inconstancy in deliveries. The 20% imprecision is attributed to the privacy protection due to derived variables.

4. Technology and Tools used:

Software:

- Spyder 3.2, a Python 3.6 Integrated Development Environment (IDE)
- Microsoft Excel 2016, data viewer in spreadsheets
- RapidMiner, a data miner tool with process flow-based interface.
- SAS, a data management tool as a reference (control) for QC of Python model.

Technology:

- Pandas Framework, to use Data frames and to read, access and modify Excel, SAS datasets.
- Multiprocessing package, to improve the runtime of the Python code.
- Scikit-Learn package, to import predictive models.

Database:

- Anonymized IQVIA datasets of longitudinal transactions, stored on Microsoft SQL server.
- Anonymized IQVIA derived datasets of doctors' data.

5. Research Methodology:

With the project being a culmination of healthcare and analytics, there are a plethora of concepts that come into play, the healthcare part requires knowledge of the biologics, indications and the therapies involved to formulate an algorithm whereas the analytics part requires technical knowledge in programming to implement the idea into a program. Hence, the concepts are vastly different for the algorithm and its implementation.

5.1 Autoimmune Biologics Market:

5.1.1 Algorithm:

The 4 phases of the program require selection, classification, spec calculation and spec assignment at prescription (referred to as 'rx' in the algorithm), patient, indication and doctor specialty levels.

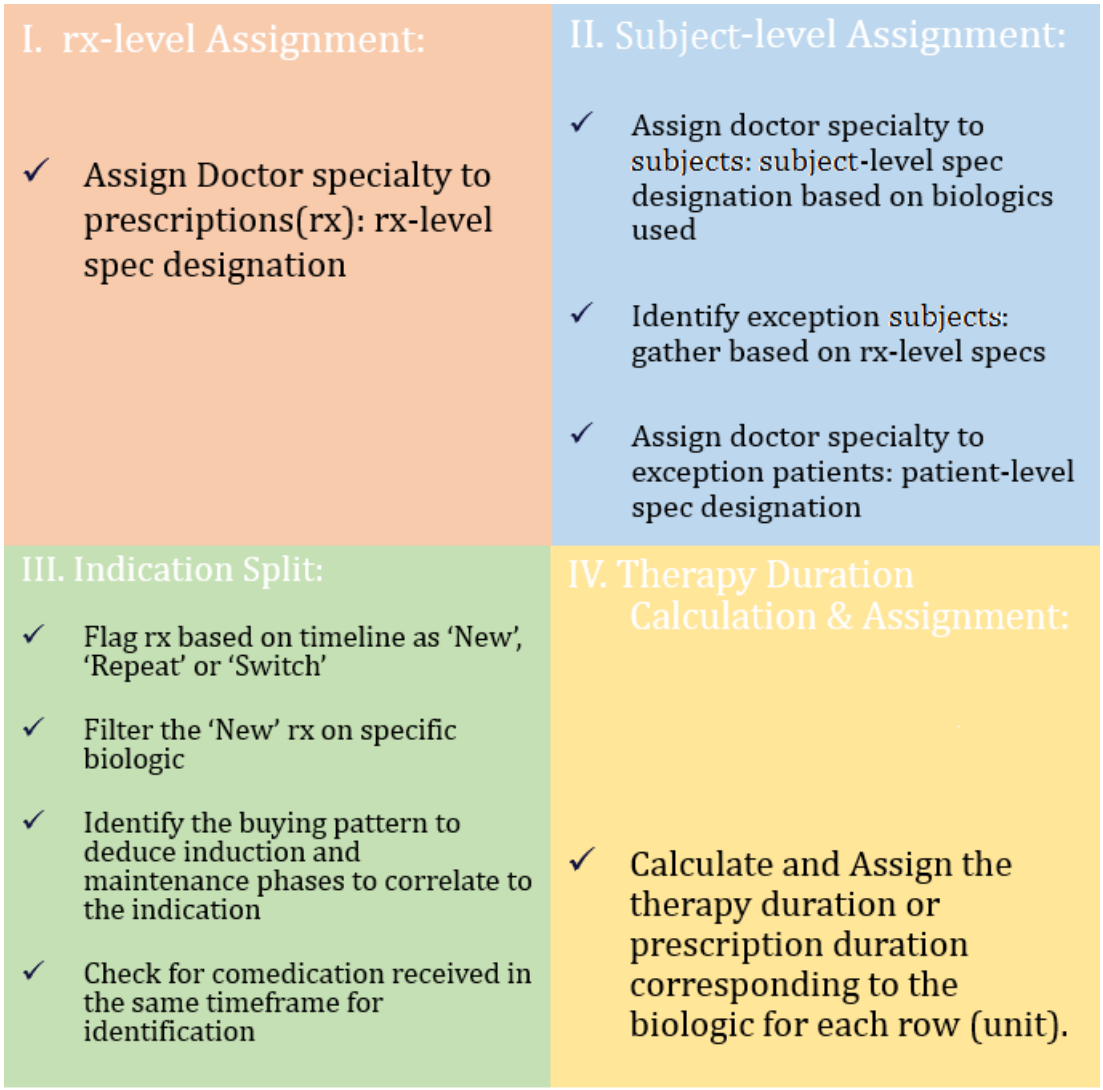


Figure 2: The algorithm of the 4 phases of project depicted above.

The base dataset containing the transaction-level data corresponding to their prescriptions is of only biologics. Each transaction is designated a 'Doctor Specialty' spec based on what doctor prescribed it. The spec designated based on the doctor's specialization. The ends Phase I.

The dataset resulting from last phase is grouped by patient's ID and the transactions of each patient are checked. If the biologic they are purchasing are exclusively prescribed by the doctor of the respective specialization, they're designated 'Doctor Specialty' spec on a patient-level. The remaining patients without any exclusive purchases are checked for the number of transactions per unit belonging to the respective 'Doctor Specialty' spec assigned in Phase I. Furthermore, a patient-level spec is assigned on the basis of the higher number of transaction-level spec to those without such exclusive prescriptions. If there is an equality, a precedence order is followed depending on the indication and biologic of interest. This ends Phase II.

The dataset resulting from last phase is considered again and flagged transaction-wise 'New' based on the business definition and market which is 18 months in case of biologics i.e., if it is the first ever biologic prescribed or initiated after a gap of more than 18 months. The new patients are filtered on the biologic of interest and checked for comedications that come with the indications and split based on them. By the end, the dataset is split across Hidradenitis Suppurativa or Psoriasis and flagged accordingly as 'HS' or 'PSO' under Indication column. This concludes Phase III.

The Indication flagged dataset from Phase III is currently in transaction per row format. It is converted to unit (of product) per row format and then numbered pack-wise from their newest transaction to the end or another 'new' transaction after a 18 month gap. Then they are assigned therapy duration (days) per unit calculated on the basis of the induction and maintenance phases of the therapy followed by them. The Induction phase of a therapy is the dosage taken over a period by a patient who has been prescribed the product for the first time. The Maintenance phase of the therapy is the dosage taken over a period when prescribed for a continuation of the product. Both phases vary depending on the product and indication. This concludes Phase IV.

5.1.2 Data Analysis:

The primary tool used by an analyst in IQVIA is SAS, a licensed software suite with point-and-click graphical interface for non-technical users along with advanced SAS code capabilities. Python is free, open source software which is lightweight and rich in resources in the form of libraries extending its reaches to Analytics making it the best open source alternative to SAS for Data Scientists. This project involves developing a Python program as an alternative to the existing SAS model in Autoimmune Biologics market.

5.1.2.1 Pandas Framework:

The ‘pandas.DataFrame’ module is used as the primary data structure to encapsulate the datasets of longitudinal transactions in SAS files and SQL server. The ‘pandas.iloc’ function was used to access the dataset index-wise and also modify it. The ‘pandas.concat’ function was used to append data frames post classification by ‘pandas.groupby’. ‘pandas.iterrows’ was used for row-wise access.

5.1.2.2 PYODBC Package:

The ‘pyodbc’ python package is an Open Database Connectivity (ODBC) tool allows connection to multiple types of SQL based servers such as Oracle, MySQL, Microsoft SQL Server, PostgreSQL, SAP HANA, Sybase ASE and DB2. This feature is not implemented due to restrictions in user previleges.

5.1.2.3 User-defined Modules:

A collection of functions encapsulated in a user-defined module were developed for conversion of datatypes and objects from SAS to Python environment, mainly the date conversion and data structure handling. SAS utilizes a single numerical representation of date in integer datatype format whereas python possessing its own datetime datatype.

For e.g.: 2019-03-20 is Python's representation whereas 21628 is SAS's representation. Appropriate functions were created to retrieve data from data frames for creation of summary file. For e.g., a function to retrieve the number of distinct patients for a data frame.

5.1.3 Multiprocessing module:

Python's 'multiprocessing' package enables the program to run parts of the code parallelly across multiple CPU cores effectively multiplying the processing speed of the sub-routine/program directly by the number of cores available. When applied to parts of code involving complex iterative tasks on huge amount of data requiring copious amounts of time, multiprocessing enables the program to run multiple instances of the task simultaneously effectively cutting the timespan by the number of instances (normally the CPU core count). For analysis of either big data or while using advanced analytics, the runtime of the program plays a decisive role in determining the project's timeline. The major factors affecting the runtime are dependent on the efficiency of the code (which in turn depends on complexity of the analysis) and size of the dataset to be analyzed, the former being the preferable option to be optimized. The following code is used to invoke multiprocessing:

```
from multiprocessing import cpu_count, Process
def Iterative_function(dataframe):
      #Process to be iterated
def PartitionDataFramebyCores(dataframe)
      #Partitions dataframe by patient's transactions
if __name__ == '__main__':
      if cpu_count() < 2:
          partitions=1;
      elif cpu_count() < 4:
          partitions=2;
      elif cpu_count() < 8:
          partitions=4;
      else:
          partitions=8;
      processedpartitionframelist=[];
      partitionframelist=PartitionDataFramebyCores(dataframe, partitions);
      pool=Pool();
      processedpartitionframelist.append(pool.map(Iterative_function,
      partitionframelist));
      pool.close();
      pool.join();
```

The program's process generates subprocesses when multiprocessing is enabled which are optimal when equal to the number of CPU cores. The above code calls for 1, 2, 4 and 8 subprocesses for single-core, dual-core, quad-core and octa-core processors respectively (Incase of hexacore processors, only 4 subprocesses are created). Before calling the subprocesses, the dataset (data frame) to be processed is partitioned by a function 'PartitionDataFramebyCores' which slices the main data frame to subsets whose number corresponds to the number of subprocesses to be called. In addition to that, functionality is enabled to slice the data frame to not split a patient's transactions across subsets. This is to not compromise the patient-level function 'Iterative_function' which requires all the patient's transactions to analyzed.

In the Autoimmune biologics market algorithm, multiprocessing is used thrice during phase II for classification of patients by their doctor's specialty, phase III for flagging of 'New' on biologic patients and phase IV for pack numbering of transactions by units.

5.2 Targeting Best Practices:

5.2.1 Predictive Modelling:

Predictive modeling is a process that uses data mining and probability to forecast outcomes. Each model is made up of a number of predictors, which are variables that are likely to influence future results. Once data has been collected for relevant predictors, a statistical model is formulated. The model may employ a simple linear equation, or it may be a complex neural network, mapped out by sophisticated software. As additional data becomes available, the statistical analysis model is validated or revised.

Predictive modelling is used in business forecast of statistics (sales, web traffic, weather, etc.) proposed by the stakeholders or clients and predicted using an array of parameters which may or may not be dependent on the predicted variable. Different types of datasets fit variedly with different models and selection of the right model for the working data structure is paramount in model building and predictive analysis.

5.2.2 Data Treatment – Class Imbalance:

Data needs to be pre-processed before loading onto a model. Missing values, null representations ('_NULL' in SAS), spaces, etc. must be treated appropriately by replacing with a zero or an average suitably. Some models require normalization of the data while most do not accept non-numeric categorical data and hence must be encoded. In Python, StandardScaler() in there are 2 options in Scikit-Learn package namely: LabelEncoder() and OneHotEncoder() in 'preprocessing' module.

The StandardScaler() rescales the data between -1 and 1 with the average of the data congruent at 0. Normalization restructures data in a way that their initial ranges and magnitudes do not bias the model. Not all models require normalization but doing so prevents the model form default feature selection and weighing features unequally.

The LabelEncoder() converts a column of non-numeric data to numeric data by creating a dictionary containing keys for non-numeric data and replacing them respectively. By using a LabelEncoder(), data is represented ordinally in integer format. This type of encoding allows categorical data to be accepted into most models but may cause inaccuracy in predictions owing to the ordinality of the classes especially with models based on equations (for e.g., a test case with attributer pertaining to class 1 and class 3 will be predicted class 2 as 2 is the average of 1 and 3).

The OneHotEncoder() converts a column of numeric or non-numeric data of ‘n’ classes to ‘n’ columns of binomial data where each column represents a class from the original column. This type of encoding is required for models like Logistic Regression and data with independent classes (non-ordinal).

It is observed that fitting the model with more data doesn’t necessarily increase its accuracy but using the right distribution of data (for classification problem) does. That brings the problem of Class Imbalance. Most dataset containing categorical data doesn’t have equal distributions for training the model. When a model is trained on data skewed towards one class’s majority, the model accurately predicts most test examples only for that class. This condition may not necessarily lower the accuracy of the predictions as the total accuracy itself is biased towards the majority class, but the quality of the predictions drops.

```
import random
Import pandas
def ClassImbalanceTrainTestSplit70Percent(inputframe, labelname):
    testframe=inputframe[0:0];
    trainframe=inputframe[0:0];
    if labelname in inputframe.columns:
        print('Label found.');
        ifgroup=inputframe.groupby(labelname);
        num_of_labels=0;
        min_data=len(inputframe);
        for a_label, grouped_data in ifgroup:
            if len(grouped_data) < min_data:
                min_data=len(grouped_data);
            num_of_labels+=1;
        if num_of_labels*min_data > 0.5*len(inputframe):
            data_per_label=int(0.5*len(inputframe)/num_of_labels);
        else:
            data_per_label=min_data;
        for a_label, grouped_data in ifgroup:
            locality=list(range(len(grouped_data)));
            random.shuffle(locality);

            trainframe=trainframe.append(grouped_data.iloc[locality[0:int(dat
            a_per_label*0.7)]]);

            testframe=testframe.append(grouped_data.iloc[locality[int(data_pe
            r_label*0.7):]]);
    splitframelist=[trainframe, testframe];
    return splitframelist;
```

The above code creates a training set with equal number of class examples selected randomly. Despite solving class imbalance problem, there is still the problem of the majority classes being non-characteristically sampled.

5.2.3 Model Selection:

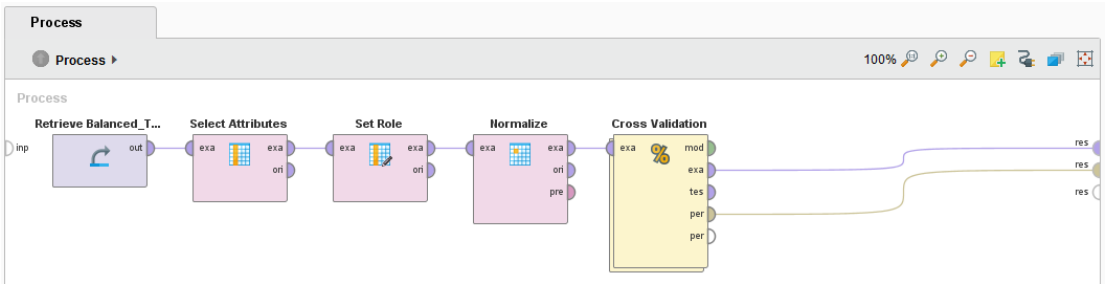


Figure 3(a): The Process Overview in RapidMiner

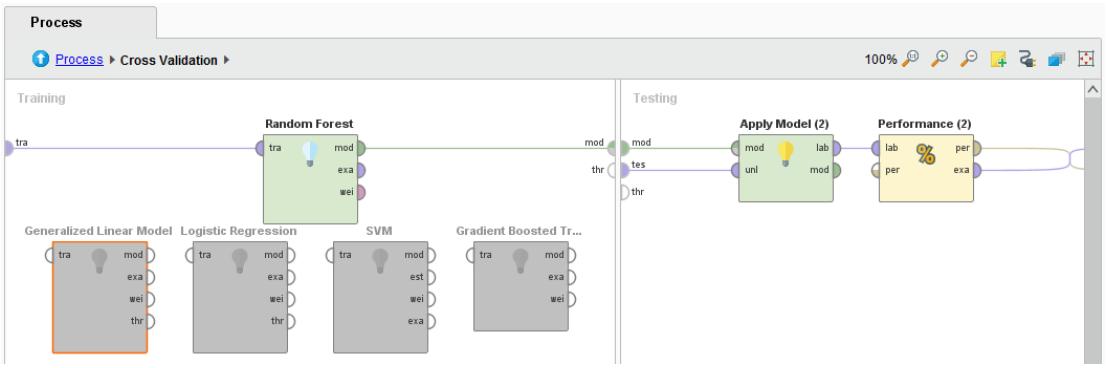


Figure 3(b): The Cross-Validation operator with different options of models used for training.

The above process flow diagram in Fig.3 was designed and run in RapidMiner post manual pre-processing of (sample) dataset with 5 different models: Generalized Linear Model, Logistic Regression, Support Vector Machine, Random Forest and Gradient Boosting Trees. The process was run with and without the class imbalance treatment. It was observed that without class imbalance, all the models exhibited high accuracy, but none had quality in predictions. With class imbalance treatment, the Random Forest and Gradient Boosting showed high accuracy and acceptable quality with the former displaying relatively higher stats. Hence, Random Forest was chosen as the model for the Targeting Best Practices program with the inclusion of class imbalance treatment. Theoretically, stacking multiple models can augment accuracy of the model and the application of stack generalization with Random Forest and SVM is carried out. By

stacking the results of SVM to Random Forest, accuracy can be improved (needs to be tested at the time of writing of report).

6. Results and Discussion:

6.1 Autoimmune Biologics Market:

	combor	DDMS_F	nsactio	FCC	Units	ShortCo	P_ID	saction	atc	prod	pack	str_un
REDACTED	REDACTED	101371	20180529	336819	1	00ACPYH	REDACTED	201805	L04B0	REDACTED	REDACTED	40
REDACTED	REDACTED	101371	20180529	336819	1	00ACPYH	REDACTED	201805	L04B0	REDACTED	REDACTED	40
REDACTED	REDACTED	101371	20180529	336819	1	00ACPYH	REDACTED	201805	L04B0	REDACTED	REDACTED	40
REDACTED	REDACTED	101371	20180717	336819	1	00ACPYH	REDACTED	201807	L04B0	REDACTED	REDACTED	40
REDACTED	REDACTED	101371	20180717	336819	1	00ACPYH	REDACTED	201807	L04B0	REDACTED	REDACTED	40
REDACTED	REDACTED	101371	20181002	363672	1	00ACPYH	REDACTED	201810	L04B0	REDACTED	REDACTED	40
REDACTED	REDACTED	101371	20181002	336819	1	00ACPYH	REDACTED	201810	L04B0	REDACTED	REDACTED	40
REDACTED	REDACTED	101371	20181002	336819	1	00ACPYH	REDACTED	201810	L04B0	REDACTED	REDACTED	40

str_me	gene	ctorSpe	CU	PSIZE	numda	DoctorC	ientLev	dTrans	Indicati	ackNum	RXDurat
MG		31	0.8	2	21333	Dermato	Dermato	New	PSO	1	7
MG		31	0.8	2	21333	Dermato	Dermato	Repeat	PSO	2	28
MG		31	0.8	2	21333	Dermato	Dermato	Repeat	PSO	3	28
MG		31	0.8	2	21382	Dermato	Dermato	Repeat	PSO	4	28
MG		31	0.8	2	21382	Dermato	Dermato	Repeat	PSO	5	28
MG		31	2.4	6	21459	Dermato	Dermato	Repeat	PSO	6	84
MG		31	0.8	2	21459	Dermato	Dermato	Repeat	PSO	7	28
MG		31	0.8	2	21459	Dermato	Dermato	Repeat	PSO	8	28

Figure 4 (a) & (b): Data of a single patient post-analysis by the program.

The final dataset exported to an excel spreadsheet (Fig.4) is filtered to a Psoriasis patient's transactions and shows the additional 6 columns added to it during the analysis. The last 6 columns are:

- DoctorClass: the prescription-level doctor specialty assignment
- PatientLevelClass: the patient-level doctor specialty assignment
- ProdTransaction: the flag describing if prescription is new, repetition or switched depending on patient's history
- Indication: The flag assigned in 3rd Phase identifying him/her with the indication.

- ❖ PackNumber: The count of the units of product used by the patient in their history
- ❖ RXDuration: The length of duration of the particular unit of prescription depending on induction or maintenance phase.

This dataset is then inserted to the cockpit for further analysis. Along with this, a summary file is created for inspection by experts in the field.

6.2 Targeting Best Practices:

The Results of TBP are the Key Performance Indicator(KPI) and the final prediction dataset retrieved by reverse-encoding the predictions back to classes and exported to an excel file. The performance of predicted classes is compared against the actual classes in a pivot table represented as a 2D-matrix namely KPI.

KPI of a Random Forest model						
	1.VH	**2.H**	**3.M**	**4.L**	**5.NON**	**`Total**
1.VH	109	10	0	0	0	119
2.H	18	153	16	0	0	187
3.M	0	7	197	11	0	215
4.L	0	0	35	191	74	300
5.NON	0	0	0	1	17	18
Total	127	170	248	203	91	839
Accuracy	91.6	81.82	91.63	63.67	94.44	79
	Total Accuracy is 79%					

Figure 5. KPI/Confusion matrix of data for Random Forest.

Above is a KPI for a Random Forest Model's prediction on doctor's sales in a de-identified market. In an ideal case, the above model's predictions clear both the accuracy and quality criteria of ~80% in total. The KPI/Performance matrix/confusion matrix is a metric by which the actual classes and predicted classes are compared with respect their distributions. It is also necessary for each individual

class to possess an accuracy and quality close to ~80% as well. Hence, the Direct Ratio (DR) and Loss Quality (LQ) metrics are calculated along with KPI to gauge the in-class accuracy and quality respectively.

DR	1.VH	2.H	3.M	4.L	5.NON
1.VH	91.60				
2.H		81.82			
3.M			91.63		
4.L				63.67	
5.NON					94.44

LQ	1.VH	2.H	3.M	4.L	5.NON
1.VH	85.83				
2.H		90.00			
3.M			79.44		
4.L				94.09	
5.NON					18.68

Figure 6. The Direct Ratio (DR) and Loss Quality (LQ) matrices

The DR and LQ values for classes can be calculated together for adjacent classes (such as VH and H) to account for standard deviation in problems with multiple classes.

7. Conclusion

At the time of writing this report, the Final flag assessment algorithm is yet to be finalized but the developed program is fully functional on the working algorithm. Functionality for updating was implemented which allows user to set criterion for analysis by retrieving data from Excel worksheets in 2nd and 4th phases. The verification of the data is done up to Phase III code and the QC for the Summary file is done with the data from the latest quarter March 2019. The project objectives are effectively complete save for implementation of SQL access, GUI and finalised Final flag Assessment Algorithm.

The program once completed will require only the base dataset (biologics), SQL dataset (comedication) and Patient profile dataset (patients' profile and history) after which the analysis is automated completely. With the evolving biologics markets, the program allows for updating specs with minimal manual intervention of the user. The fully automated program with the means of updating to the current market developments with the rise of new biologics and biosimilar means the analyst not requiring coding or have any technical knowledge to run the analysis on a dataset. Biologics being a lucrative market and Autoimmune diseases being lifestyle-affecting to fatal, the prescriptions and injections are taken strictly on time thereby making this market analysis most accurate.

8. Future Work

Being open source, the program reduces the investment on licensed software by the company while suffering no drawbacks in performance. This Biologics model can be used for developing programs for other markets and indications as well.

Targeting Best Practices program needs to be updated with better class imbalance treatment features like up-sampling, down-sampling and auto-sampling using 'imblearn' package containg 'SMOTE' module.

8.1 Challenges faced:

1. SAS Programming language
    - To understand the pre-existing model by studying the code for deducing the algorithm.
2. Learning the structure of transaction-level data
    - Selection of python data structure for optimal analysis among csv (comma separated values), lists and Dataframes.
3. Conversion of datasets from SAS to Python environment

- With each data structure having its own features and limitations with respect to semantics and warnings.
- 'SettingWithCopyWarning' in Pandas while modifying dataframes.

4. Update Capability
    - To allow input of user defined criteria for analysis by creating a robust program.
5. SQL Access
    - Learning about different SQL servers like MSSQL, PostgreSQL, MySQL etc
    - Accessing the database using python (to be done at a later stage).
6. Application of Stack Generalization
    - At present, there are no modules for stacking and the algorithm needs to be manually coded and data pre-processed properly for stacked models individually.
7. Implementing features of Class Imbalance
    - The implementation of Class Imbalance for dynamically managing the quality of the predictors is still underway.

9. Bibliography

Arthritis Foundation. (n.d.). *Biologics (Biologic Response Modifier) Overview*. Retrieved from Arthritis Foundation: https://www.arthritis.org/living-with-arthritis/treatments/medication/drug-types/biologics/drug-guide-biologics.php

Chron's & Colitis UK. (2017). *Treatments*. Retrieved from Chron's & Colitis UK: https://www.crohnsandcolitis.org.uk/about-inflammatory-bowel-disease/treatments

Denise Mann. (n.d.). *New Drugs for Rheumatoid Arthritis: Is a Biologic Pill on the Way?* Retrieved from WebMD: https://www.webmd.com/rheumatoid-arthritis/features/new-drugs-for-ra#1

FDA. (n.d.). *U.S. Food & Drug Administration*. Retrieved from Fractionated Plasma Products > PANZYGA: https://www.fda.gov/BiologicsBloodVaccines/BloodBloodProducts/ApprovedProducts/LicensedProductsBLAs/FractionatedPlasmaProducts/ucm615698.htm

FDA. (n.d.). *Vaccines, Blood & Biologics > CUTAQUIG*. Retrieved from U.S. Food & Drug Administration: https://www.fda.gov/BiologicsBloodVaccines/ucm628258.htm

Lindsey, H. (2018, 3 7). *Common Treatments for Lupus*. Retrieved from Everyday Health: https://www.everydayhealth.com/lupus/guide/treatment/

Morriss, E. (2019, 2 4). *Rise of the biosimilars*. Retrieved from Pharmafield: https://pharmafield.co.uk/in_depth/rise-of-biosimilars/

N., S.-B. (2014). Biologics: the role of delivery systems in improved therapy. *Biologics*, 8:107-14.

Rath, L. (2017, June 1). *FDA OKs a New Biologic for RA*. Retrieved from Arthritis Foundation: http://blog.arthritis.org/news/fda-approves-new-rheumatoid-arthritis-biologic-sarilumab/

Rossi, K. (2018, OCTOBER 31). *Biosimilar, Hyrimoz, Approved by FDA for Host of Chronic Conditions*. Retrieved from MD Magazine: https://www.mdmag.com/medical-news/biosimilar-hyrimoz-approved-fda-chronic-conditions

Rouse, M. (2018, March). *What is predictive modelling? - Definition from WhatIS.com*. Retrieved from TechTarget.com: https://searchenterpriseai.techtarget.com/definition/predictive-modeling

Shanley, M. (2017, August 29). *FDA Approves Biosimilar for Chronic Inflammatory Diseases*. Retrieved from Rare Disease Report: https://www.raredr.com/news/fda-approves-inflammatory-diseases-biosimilar

Weiss, M. (2017, 12 14). *FSA Approves Biosimilar for Autoimmune Diseases*. Retrieved from The Dermatologist: https://www.the-dermatologist.com/news/fda-approves-biosimilar-autoimmune-diseases