

EAS 560 – Spring Internship Report

Name: Sahana Mukherjee

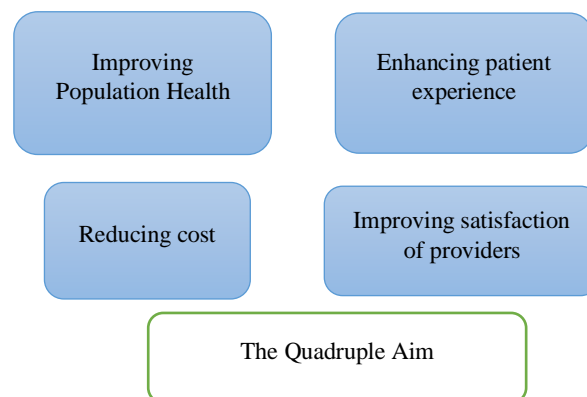
UB Email: mukherj3@buffalo.edu, **UB Person#:** 50373858

1. INTRODUCTION

This report is a comprehensive summary of my internship, pursued during my final semester. It counts towards 3 credits of my degree requirements. This internship was a part time job (20 hours per week) which exposed me to the real-world industry and allowed me to gain skills that would be essential for me to excel in my career going forward. My job aligned with my coursework perfectly, and as a result I got an opportunity to showcase my knowledge of classroom into real solutions. I worked in the capacity of a Data Analyst and Visualization Engineer at Primary Care IPA, Buffalo New York. Primary Care IPA is an independent physician led association that consists of 25 primary care and paediatrics practices of Western New York. The organisation aims to enhance efforts of doctors and primary care providers to improve the overall population health, reduce costs of healthcare and improve the work life balance of physicians.

Reporting to the Population Health Manager of the organisation, I was in the responsibility of leveraging my data-oriented knowledge and skills, to transform how information is being used and introduce a data driven direction to the organisation. Healthcare industry is one such kind, which deals with enormous amount of highly confidential and sensitive data, that includes patient details, electronic health records of patients, medical history, physicians' information, and costs and claims associated with patients. It is quite surprising in this present world of advanced technology, that a major portion of information management in many healthcare companies is still not completely automated and involves a lot of cumbersome manual work. This is mainly due to the gap that exists because of lack of proper communication to the stakeholders of medical world, like doctors and pharmacists. This is where data science and analytics comes into play. Data driven architecture not only has the capacity to retrieve useful information out of massive pools of data, but also communicate in the best way possible for the medical world to understand, with the help of advanced data analytics/visualization and reporting tools.

As a data analyst/visualization engineer, at Primary Care IPA, I was in complete charge of transforming the existing manual infrastructure, introduce data driven strategies, imagine and build a platform that would be capable of using data in a smarter way and allow stakeholders to take business decisions in a more streamlined way. This strategy was aimed at helping doctors know their patients better, understands risk profiles of patients, and provide necessary care before their conditions worsen, thus reducing expensive costs at later stages. This is based on the very foundation of the “Quadruple Aim” on which Primary Care IPA is built: improving population health, reducing cost of care, enhancing patient experience, improving satisfaction of doctors and providers.



2. BACKGROUND

Primary Care IPA being an association of 25 primary care and paediatric practices, deals with a plethora of data. This data essentially comes from the Payers that include Highmark-Western New York (independent licensee of the Blue Cross Blue Shield), Independent Health, Fidelis. This data includes patient level details like demographics, age, gender, the areas of diagnosis, gaps in care, hospitalizations, readmissions, emergency visits and more. It also includes the healthcare costs of patients. Primary Care IPA aims to organise all the data and make the best use of it, so that physicians can take more informed decisions on their patients. Also, the organisation aims to have a clear picture of how the individual practices are performing in treating their patients. This involves creating a system that would ingest all the patient level data across several Payers and create a landscape to visualize the performances of the practices. Out of the 25 practices, some have patients who are mostly above 60 years of age, some have more patients who are in the diabetic stage, some have children with Autism, and some have terminal diseases like Cancer, which augments the healthcare costs of that patient. It is very essential for the doctors to understand the bigger picture, and then be able to drill down into minute details of patients, to enhance their care.

As a data analyst, my goal was to understand different sources of data, transform them into desirable format, perform statistical analysis of data to come up with metrics that would help the practices to evaluate themselves, leverage my knowledge of probability and statistics to sketch risk levels of patients, and create comprehensive automated reports and dashboards, that would effectively communicate the findings and results to my stakeholders. My knowledge of core subjects that became useful during the tenure of my internship includes EAS 502 Introduction to Probability Theory, EAS595/509 Statistical Learning I/II, EAS 503 Programming Database Fundamentals, CSE4/560 Data Models and Query Languages. Besides this, I involved myself in deep research of the healthcare domain, understand different regulations, MEDICARE/MEDICAID measures, domain specific calculation techniques, and data security issues. I mainly resorted to my organisation's archive for SOPs and online resources for articles on data science/analytics applications to healthcare industry. For data visualization/reporting purpose, I used the platform of Power BI, and created the environment that could capture data from different sources and generate cohesive dashboards for my stakeholders.

3. METHODOLOGY AND TECHNIQUES

3.1. Requirement Analysis

During the initial days of my internship, I engaged myself in understanding my stakeholders and gathering business requirements in detail. Primary Care IPA includes practices for both adult treatment and paediatric treatments. The two follow strictly different approaches towards managing their patients, evaluating practice performances, and calculating total cost of care. To understand these details and become familiarized with the healthcare sector, I held several meetings with my stakeholders. This included meeting with the chief pharmacist, clinical care coordinator, care manager, population health specialist and paediatric specialist. During these meetings, I was able to absorb most of the approaches followed in the industry, and the technical loopholes that existed and could be worked upon.

My stakeholders had varied requirements and goals which included having a clearer picture of the patients and understand their patients better. One of the major requirements for the physicians and care managers included, understanding the diabetic patients more, evaluate the risk level of those patients, provide prior medication and care before diabetes becomes worse, thus preventing increased cost of care at a later stage. Monitoring diabetes, as I learnt, is also essential to drive the many of the diabetic and pre diabetic programs in the city, that my organisation is involved in. I gathered from the paediatric specialist about specific requirements in the children treatment section, that included close monitoring of immunization status of children, screening status of different diseases at defined intervals of age like

Autism. I went through different sources of data, that contained information about patients, diagnosis details, and related costs of care. These data came from the Payers who were in contract with Primary Care IPA. My goal was to understand the data and retrieve useful information that would help my stakeholders answer questions about their patients, in a more structured and confident manner.

3.2. Data Wrangling/ Cleaning

The data came in a variety of formats, including PDF files, Text files, and Excel spreadsheets. A significant portion of my work included combining different sources of data, cleaning the data, and making it suitable for a centralized system. My knowledge of Python libraries and ETL (Extract, Transform and Load), came useful during this exercise. I deployed several techniques that helped me structure the data in a suitable format.

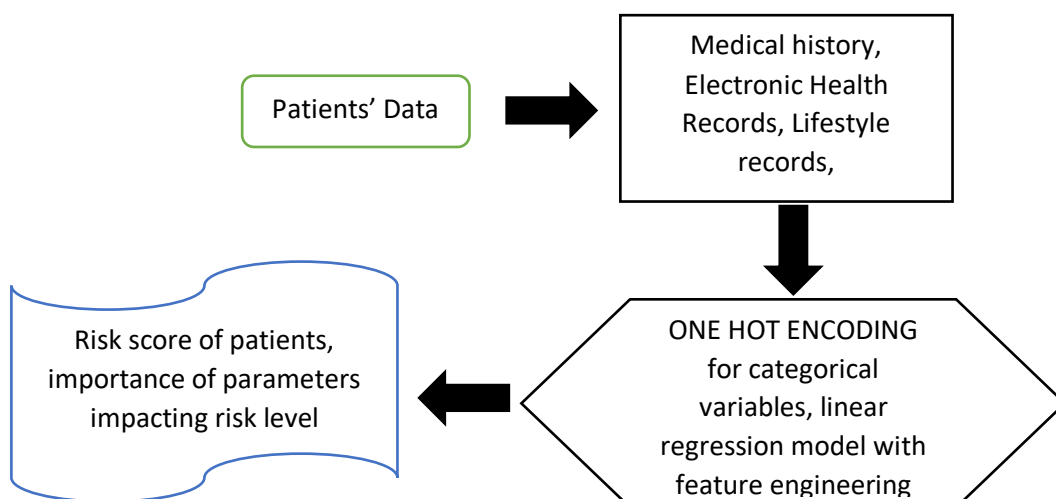
- 1) **Missing data** – One of the major challenges that I faced during this exercise, was dealing with missing values. Patients often do not tend to disclose personal information like date of birth and gender. However, for diagnostic purposes, these form essential factors or features that determine patients' behaviour and treatment courses. I used Python visuals like Heatmaps and histograms to easily identify missing values. I adopted the following techniques in Python (Numpy and Pandas) to deal with missing values:
 - a) Dropping the rows that had missing values – In case the number of rows having missing values were less compared to the entire data set, I dropped them because it did not cause major loss of information.
 - b) Padding the missing values with relevant information – In cases when I encountered greater number of rows with missing values, dropping them would lead to loss of information. So, I researched more into the patients' profiles, conducted meetings with stakeholders and gathered numbers that I could pad into the missing entries.
- 2) **Inconsistent data across Patients' address and names** – The names of patients often had inconsistency, middle names and incorrect capitalizations and special characters, that made mapping of names to unique patient IDs difficult and error prone. The same challenge I encountered while cleaning the address field of the data. Addresses were inconsistent and extraction of zip codes from them became difficult as a result. I used Pandas library of Python to clean these irregularities in data and create consistent clean data frames.
- 3) **Challenge with DOB of patients** – Date of birth of patients came in different formats, as every payer has its own way of formatting date and time in their individual systems. I worked on bringing all the data to the same format for with one common SQL formatting for date/time data so that the data can be easily combined in the database system.
- 4) **Duplicate data** – Payers often make errors while reporting their patients and one of the frequently occurring error was that of duplicate records. Duplicate records resulted in unwanted arithmetic over numeric variables and produced redundant results. I deployed Pandas library methods to identify duplicates and remove them from the data set.
- 5) **Outliers** – Patients' data often came with outliers. The most occurring outliers that I faced were those of patients' age. I came across data where amongst all patients of age 50+, there existed few records with age between 10-15. Those were anomalies, either erroneously reported or special cases. I took note of such outliers and communicated with population health specialist to understand further actions. I used methods like Box Plot and descriptive summary of data in Pandas library to identify outliers and report to my stakeholders.

3.3. Statistical Analysis of data

Primary Care IPA wanted to know how all the practices (both adults and paediatrics) are performing in treating their patients. Each practice is comprised of assigned physicians and care providers. It is essential to evaluate the practice performances, based on predefined medical metrics, and drill down into how each of the physicians within a particular practice is performing. This exercise is of critical importance to Primary Care as it would allow the organisation to know how practices compare with each other, and better manage the budget and cash flow. This is also critical from patients' point of view and gives a clear and better picture of the patients each of the physicians are treating within a given practice.

Leveraging statistics and probability theory - To achieve this landscape, I deployed my statistical knowledge and came up with calculations that would consider required features and calculate the practice performances. I used methods like hypothesis testing, p-value, and probability, to calculate numbers that would be statistically significant. Methods like z-score, mean/median calculations, and confidence intervals, came useful for me, as I ran the numbers and picked the best performing practices. This decision is critical to the organisation, from a business perspective, and required special attention. I had the responsibility to leverage my knowledge of statistics and probability to evaluate how each of the practices have performed over the entire year of 2021, across every diagnostic measure. Some of the major measures included several treatments of diabetes, cancer, heart diseases and immunization/screening (for children). I had to combine data from all months over the year and rank the practices across every diagnostic measure based on predefined medical metrics. Mapping the numbers with the metrics was critical for me, as it required my statistical knowledge, as well as my domain understanding of the healthcare industry.

Risk Model for Patients using Regression - Besides evaluating the practice and physician performances, I also had the responsibility to create a model that could rank patients based on their risk level. The features that I used to evaluate risk level of patients included among others, hospitalizations, readmissions, emergency visits, gaps in care, number of comorbidities present, claims made, and level of complicity of the conditions present. This exercise was important as it allowed doctors to have a comprehensive picture of their patients, and schedule appointments for those who need immediate need care and identify patients who have gaps in appointments and send out reminders to them. I built the model using Python libraries (Numpy, Pandas, and Scikit Learn) and used regression to predict the risk score of patients based on the input features listed above. I used the claims file, to pull the data of patients that would be instrumental in deciding the risk score. Lifestyle of patients is also critical in determining risk level of patients, and this data came from survey forms and demographic information. I deployed linear regression with feature engineering to predict the risk score of patients. I used feature engineering (and generated interesting visuals depicting the impact of top 10 parameters) to understand which were the most important parameters that impacted the risk score of patients.



3.4. The Database Design

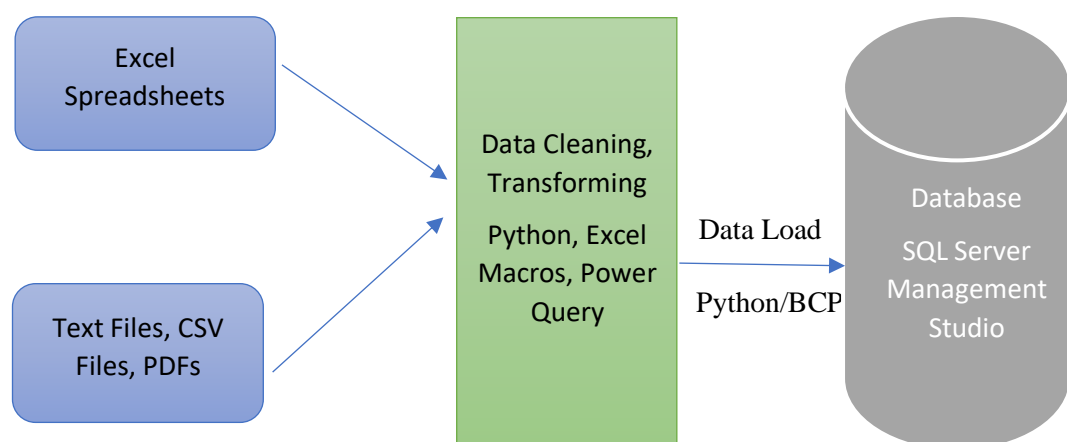
The healthcare industry hasn't completely reached the summit of technical advancements, and many workflows still follow manual and decentralized systems. One of my accomplishments in my internship was to identify a strategy that would enable all the data to be centralized on a database, and operated from there, instead of dealing with multiple sources like Excel, PDFs, and others.

Loading the data into the database - Primary Care already has a database system (SQL Server Management Studio), but I discovered that it is not structured in the most modern way that it should be and many of the data are still not loaded properly into the database. As I was mainly working on practice-performance data, I worked on writing scripts in Python, to convert the data into data frames and create connection between the python script and the database system. This would allow automatic loading of data into the database. I also employed Bulk Insert Commands to bulk load large volume of data directly into database from raw text files. This required me to create master schema in the database, and then bulk load the previously cleaned raw data from multiple sources into the corresponding tables.

Queries to fetch relevant data - I also worked on creating necessary tables, and writing advanced queries using joins, window functions and pivots that would create necessary views, child tables and stored procedures. This approach is going to help the organization maintain data in a more secure and consistent way. From the master table of data from Payers, I worked on queries that would fetch business specific results, like high-risk patients, diabetic patients, patients with more than 2 months of gap in care, patients under a specific physician with a specific diagnostic measure, child patients with screening required in coming dates and so on. Such answers are important to the doctors as well as the organisation, to monitor progress of patients and treatment priorities. Some of the methods that I used during this exercise included joins (left, right and inner joins), pivot function, window functions (rank, dense rank, lead, lag function), group by and order by clause, triggers, and stored procedures.

Combining data from different sources, and creating clean tables was a challenge, as there were number of discrepancies in data types, across the files and lots of null values. Cleaning the data to make it consistent across data types, so that union of tables become possible, and the process can be automated for further loading of data in the future was my primary objective. Methods learnt from my CSE4/560 coursework came useful in this exercise, like schema designing, window functions (rank, dense rank, lead, lag functions), advanced joins and views.

I worked on strategizing and making the system dynamic so that, as every month data comes in from the Payers, the script automatically fetches the latest file from the specific folder and loads the data into the database and subsequently runs the queries in the script that would provide answers and summary to the month's details



3.5. Data visualization: Creating automated reports /Analytics in Power BI

Shift from the paradigm - My work at Primary Care IPA involved creating and maintaining reports for my stakeholders. This was a shift from the paradigm at my organisation. Each of the 25 practices requires a detailed picture of their patients and performances at the end of every month. The report of December is a more comprehensive one, as it contains the trend over the entire year and yearly revenue too. These reports were usually made from Excel files, in PDF format and sent out to the individual practices through email. This made the reports only static and lengthy documents of several pages. One of my critical works was to automate this process and shift the entire approach of creating monthly updates' reports to the platform of Power BI. Power BI is an analytical platform and is a product of Microsoft. It has extensive capabilities of performing advanced analytics, producing a wide range of visuals (bar charts, graphs, trend lines, pie charts, tree maps, geographic maps and more). The property of Power BI, that I mainly pitched when I was leading the meeting with my stakeholders, was the interactive property of this platform. Reports in PDF formats, lack interactivity, highlighting options, and do not give the user the freedom to drill down into details, or filter any specific portion that he might be interested in, rather than looking at the entire report.

Use Cases - Use cases included, drilling down into only diabetic patients from the cohort of all patients and look at their profiles, compare different doctors and practices over the months and over years using trend lines and be able to see quarters' data and specific months' data, filter patients over 65 years of age and look at their diagnostic measures and identify risk levels. Use case also included scheduling of appointments by running the report, as this report would give the physician the freedom to see who all his/her patients are, what are the conditions they have, what are their risk levels and who is nearing an appointment or who needs a prescription refill and so on. Such detailed, yet compact landscape is not possible through PDF reports.

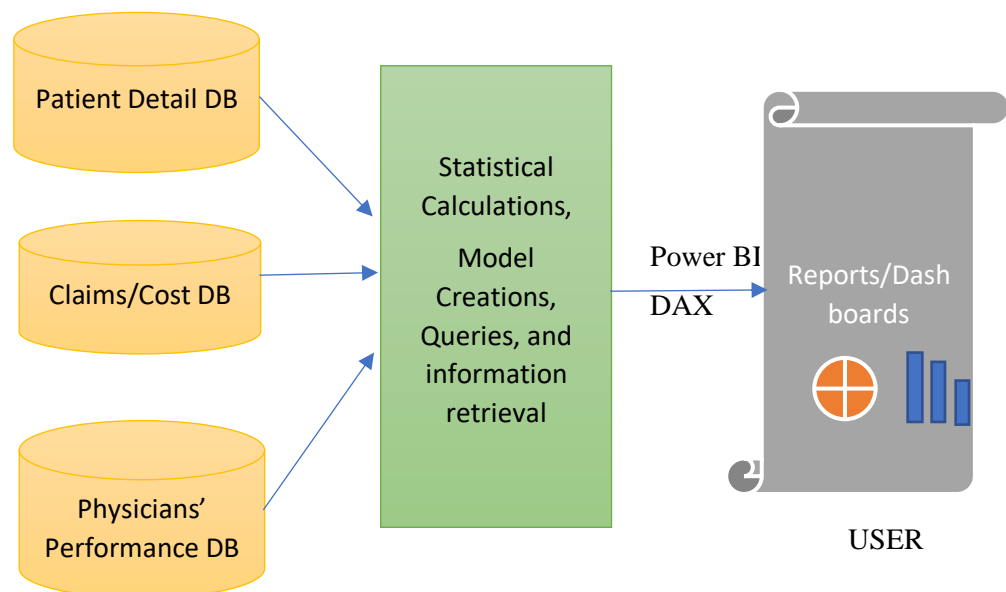
Generating dashboards and Scripting with DAX Queries - My work included connecting Power BI to the data source, importing the data, and performing analytics on the data inside the BI platform. I generated dashboards in Power BI, that included detailed picture of monthly updates, including revenue and patient information. I used bar charts, trend lines, tree maps, pie charts, gauge, and dual charts to represent the data that would answer the use cases of my stakeholders. The analytics involved creating custom columns using BI's DAX language and showcasing relevant numbers on the visuals. I had to identify and if necessary, create relations between tables, to connect related data. I leveraged my knowledge of relational database, and Entity Relationship Diagram, that allowed me to create connections among large number of data tables based on common identifiers across them. Such connections were essential to introduce filters and slicers in the visuals that would enable all relevant visuals to change based on the selected filter or slicer. I introduced maximum interactivity in these reports, by including relevant filters, drag-drop features, custom buttons, page navigation options and hovering capabilities. Such features made the reports neat and compact and capable of answering variety of business questions.

Automating the generation of reports - My work in Power BI Analytics also involved making the report generation as easy and automatic as possible. As each of the 25 practices require reports at the end of each month, it is a cumbersome process to generate reports every time. I researched and came up with the idea scripting a DAX query, that would parametrize the file path, instead of hard coding the file in the data source. I leveraged the feature of Power BI Template, which allowed me to make one report and save it as a template. This template has the data source set as dynamic, i.e., every time the report opens, it prompts for a file location from the user. The user then passes the file location in the prompt box, and reports are generated using that specific data file, in the same template with all the source formatting and interactive capabilities intact. This makes the reports dynamic in nature and reduces the need for creating dashboards every month. As long as the schema of the data remains same, Power BI template will fetch the new data passed as the parameter and generate similar reports every time.

3.6 Sharing reports with Stakeholders: Role based access control mechanism

PDF reports were sent out to stakeholders by email every month. I worked on strategizing a more convenient and secure way of sharing Power BI reports with the stakeholders.

- 1) **Role based Access Control for BI Dashboards** - I deployed the method of creating separate workspaces for the individual practices on Power BI Service environment. I collated the reports generated in Power BI Desktop, published them to practice specific workspaces in Power BI Service (which is the online service of Power BI). I developed role-based access control for these reports. Only people in the organisation will have access to the reports. Engineers at Primary Care, who developed reports were given ADMIN access and had the authority to read and edit the reports. Physicians were given READ ONLY access, so that they could not edit any data or visuals.
- 2) **Preliminary demonstration sessions with Practices/Doctors** - The approach that I strategized made the sharing of reports within our organisation simple and secure. Whenever a report is published to a particular workspace by any engineer on the Primary Care end, the corresponding practice/physician gets an update/email and simply clicks on the link to visit the Power BI Workspaces page to view the latest report shared with them. I also virtually met few of the practices where I had the responsibility to introduce the platform of Power BI to the physicians and show them how to access and navigate through reports. It was challenging and exciting at the same time, to explain the technicalities to the medical domain experts, who were not so familiar with the data architecture and back-end development processes. I pitched about the platform, its capabilities, and showcased it in a simpler way, by identifying similarities with Microsoft Excel. I detailed them on a step-by-step approach to access and navigate the dashboards and make the best use of the interactive capabilities that Power BI offers.



4. RESULTS / DISCUSSIONS

During the tenure of my internship, I was involved in a variety of tasks, that included driving the end-to-end strategy for transforming the existing legacy system, into a data-driven architecture. I was responsible for managing and cleaning data, performing calculations, loading data, and writing advanced queries to generate answers for business. I also took lead in creating visuals for my stakeholders using Power BI. Besides this, I also brainstormed on improving the website design of the organisation. I proposed that the dashboards should be hosted on the website, which would make the

website more interesting and informative. These dashboards include generic visuals and do not contain any confidential patient information. I worked with the back-end development team to host some of the generic Power BI dashboards on the website. These were mostly overview of all practice- performances, bird's eye view of all diagnostic measures across different regions, and interactive comparisons of different physicians and care providers.

From a manual and cumbersome process of managing large number of Excel spreadsheets, and manually sending out monthly reports to stakeholders in static PDF format, I, under the supervision of population health manager and in collaboration with back-end development team, have been in the responsibility of shifting this system into a structured and more automated process. Data is now being handled in the database system, and reports are henceforth generated in Power BI platform, which makes them a lot more comprehensive and interactive. Monthly reports need not to be generated every time, as long as the underlying schema of the data remains same. Similar reports can be generated for every month, using the template reports that have been created, simply by passing the new file path to the user prompt.

Doctors/care providers who have license to Power BI, and are within the organization's users list, have the authority to view and navigate through these reports easily.

A few of my accomplishments during my internship tenure, to summarize will be:

- 1) Extraction and Transformation of raw data, into proper clean format that is suitable for loading and reporting purposes.
- 2) Performing statistical calculations on data, to answer business specific questions.
- 3) Regression model creation for predicting risk level of patients.
- 4) Scripting for loading data into database system.
- 5) Writing advanced queries in database to create necessary views, tables and retrieve relevant information.
- 6) Strategizing the transformation in reporting process from PDFs to interactive dashboards in Microsoft Power BI.
- 7) Connecting multiple data sources in Power BI, scripting DAX queries to generate cohesive reports/dashboards for stakeholders.

5. FUTURE OUTLOOK

With the introduction of centralized data management system, and automated report generation in Power BI platform, I was a part of several discussions at my organisation regarding future goals and improvements. We are imagining the company's own platform that would serve as a one-stop portal for the doctors/practices and care managers. This platform would allow physicians to look up any of their patients and get all information as required. I brainstormed on this conducted several meetings with the leadership and executive team members. Me and my team are thinking of developing an entire web application that would have all the capabilities and features required for continuous monitoring of patients and healthcare costs.

Leading the idea of a web application for Primary Care doctors/practices/care managers:

This application is essentially going to be all-in-one shop for the practices/Primary Care engineers and pharmacists who are in contract. I researched on the idea and came up with few interesting user experiences (UX) that the application must cater to –

- a) **Practices/Doctors Block** – This part of the application should be able to contain all relevant information about the doctors in the practices. Such information might include their specialities, achievements, publications, and office hours. Such a comprehensive platform would allow young doctors to showcase their priorities.

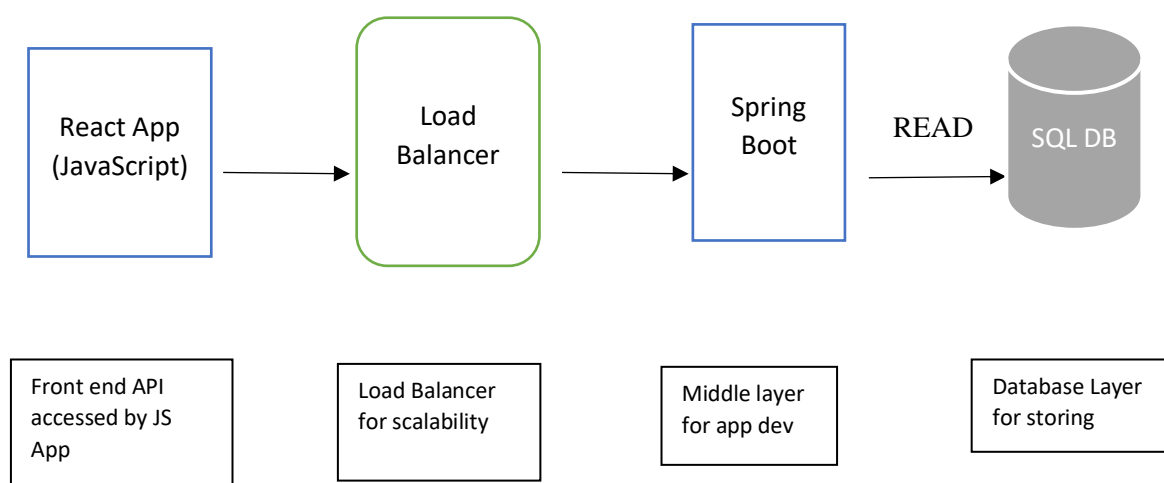
- b) **Patients' Block** – This one-stop application is intended to help Primary Care IPA know its patients better and enhance care for them. This block would contain all information about patients under respective doctors, their medical history, gaps in care, demographics, hospitalizations, diagnostic measures, and others. Doctors/ care managers should be able to see all the patients they serve, pull up any patient and retrieve all their information, and identify high risk patients, schedule appointments, and send reminders to patients.
- c) **Claims and costs** – This are essential for determining how Primary Care IPA is performing and what the monthly and yearly earnings are. It is essential for the business to know how the patients are spending on healthcare, which are the diseases or treatments that are costing the most, how the respective Payers are covering insurance for their patients, how coverages can be improved and more.

Envisioning a bird's eye view of the design of the web application:

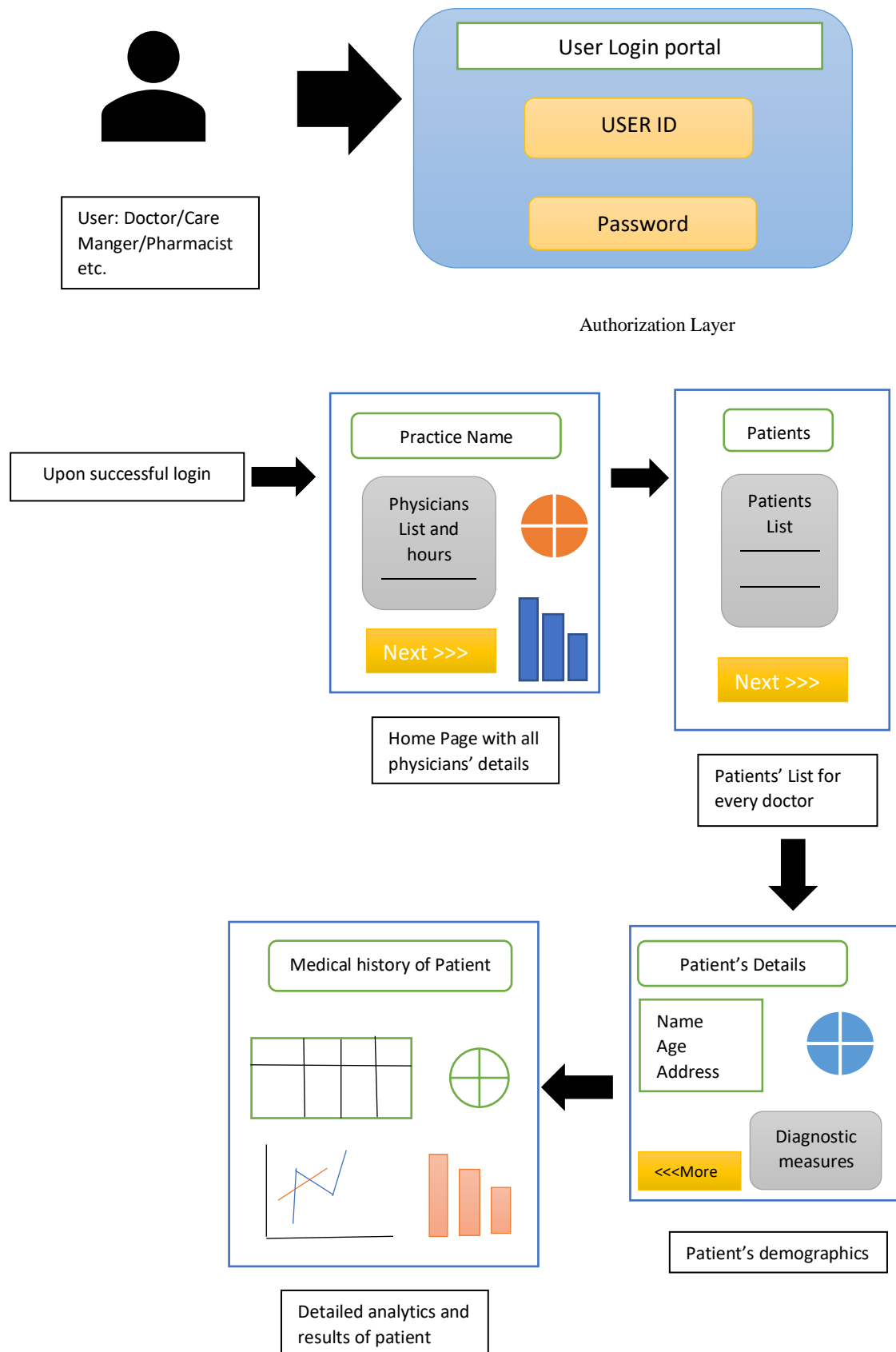
After understanding the requirements and use cases of this new application, I spent the last few days of my tenure brainstorming over designing the white board layout of the idea. I researched on web applications and existing healthcare applications (both web and mobile) and was able to put together a schema for the idea.

- a) **Database layer** – This is going to be the foundation layer. It is going to be built with either existing SQL Server or any other more powerful database engine. All the CRUD (Create, Read, Update and Delete) operations are performed in this layer.
- b) **Persistence layer** – Contains logics for storage. It is responsible for converting business objects into rows of the database.
- c) **Business layer** – This is the layer responsible for business logics, authorization, certificates, secrets, and validation.
- d) **Presentation layer** – This is the front-end interface that will consist of all buttons and navigations to dashboards and reports. Power BI will be connected to the database system and used to generate dashboards. This interface will be easy for doctors to understand and provide them a complete picture of their patients upon simple clicks. This layer is responsible for handling all the HTTP requests from users, authenticating, and transferring requests to business layer.

Underlying System Architecture:



The User Experience:



I conducted meetings with the leadership team to put together requirements for this development. I presented my design of the architecture and the User Experience to the leadership team. I held several discussions on recruitment strategy for more back-end engineers skilled in JavaScript, HTML and CSS, and the budget the organisation must put in into this development. The discussions revolved around what benefits this web application would bring to Primary Care and how it would benefit doctors and patients too. I also held meetings to discuss in brief the extension of the web application into a mobile application later, that would bring enable physicians to easily navigate and continuously monitor patients.

6. CONCLUSION

My internship at Primary Care IPA was full of learning, knowledge transfers, researching, and applying my classroom skills at an industrial level to develop real time solutions. Although, it was a part time internship, I am pleased to share that my organization treated me no less than a full-time employee and trusted me with critical assignments. This internship will be instrumental in shaping my career ahead. Besides applying my data science and analytics skills to a wide range of industrial problems, I got the opportunity to interact with a broad variety of stakeholders, ranging from doctors, pharmacists, care givers, care managers and population health specialists.

I was able to go deep into the healthcare industry domain and understand the vast potential that data science and analytics holds in this sector. Unlike other sectors like banking, finance and travel and hospitality, the healthcare industry still has wide areas unexplored. Through my experience I learnt that this is mainly due to the highly confidential and sensitive data about patients, and the fact that a minor error can risk someone's life. Data management, application of analytics and data science techniques under exceptional supervision, in this domain can transform healthcare and treatment procedures and make a better tomorrow for mankind.

This internship gave me an opportunity to contribute my analytical and data science skills for the improvement of patient care, enhance population health and reduce healthcare costs. I look forward to keeping in touch with my colleagues who guided me throughout my tenure. I am hopeful that the organization takes a giant step from here and goes above and beyond to transform patient care for the good.