**Progress report – July 2018**

**Title:** Observational analysis of ayurvedic principles, ayurvedic hospital data, and patient outcomes

By Vinay Mahajan, Girish Tillu, Ashwini Mathur

Summary: Progress made on the research work after Sept 2017 till July 2018

**Course work [Online courses on Coursera.org] Completed:**

1. Health Informatics on FHIR, Georgia Institute of Technology, by Mark L. Braunstein, MD
2. eHealth: More than just an electronic record, The University of Sydney, many instructors
3. Data science, Johns Hopkins University, by Jeff Leek, Roger D. Peng, Brian Caffo
   1. The Data Scientist’s Toolbox
   2. R Programming
   3. Getting and Cleaning Data
   4. Exploratory Data Analysis
   5. Reproducible Research
   6. Statistical Inference
   7. Regression Models
   8. Practical Machine Learning
   9. Developing Data Products
   10. Data Science Capstone
4. Online course on National library of Ayurveda medicine (NLAM) website for Overview of Ayurveda course on <http://www.nlam.in/>
   1. Ayurveda Introduction
   2. Ayurveda Avtaran
   3. Ayurveda Literature
   4. SrushtiUtpati [Cosmology]
   5. Purusha – Prakriti
   6. Tridosha
   7. Dhatu
   8. Mala
   9. Prakriti
   10. Dhatu Sara [Prominent tissue]
5. Course work: Trans Disciplinary Knowledge framework course from 6th to 10th March 2017 @TDU campus, Bengaluru.

I was provided access to the hospital database in April 2018. The following section outlines understanding of the database and analysis.

Alternative approach using the SQL database built by the eRX team

Login on the SQL database and access the hospital database. There are approximately 190 tables holding the different components of hospital data. There is a need to understand the data structure, data contents and quantity at disposal.

General flowchart for data processing:

1. Login on the SQL server using the credentials
2. Use the IAIM schema and access the following tables to generate base table with Demographics information, Patient Visit information and prescribed treatments
   1. STATE\_MASTER
   2. COUNTRY\_MASTER
   3. CITY
   4. STATE
   5. PATIENT\_DETAILS
   6. PATIENT\_REGISTRATION
   7. MRD\_DIAGNOSIS
   8. PATIENT\_PRESCRIPTION
   9. PATIENT\_MEDICINE\_PRESCRIPTIONS
   10. IP\_PRESCRIPTION
   11. SERVICES\_PRESCRIBED
   12. SERVICES
   13. MEDICINE\_SALES\_VIEW
3. There are many CRF pages built to collect relevant Ayurvedic data, measurement data, Hospital visit data, food / exercise advice, etc. This data is present in the following tables:
   1. PATIENT\_SECTION\_DETAILS
   2. PATIENT\_SECTION\_VALUES
   3. SECTION\_MASTER
   4. SECTION\_FIELD\_OPTIONS
   5. SECTION\_FIELD\_DESC
   6. PATIENT\_CONSULTATION\_FIELD\_VALUES
4. The datasets created in steps 2 and 3 are further processed using R programming language and the analysis ready datasets are created
5. Patients suffering with metabolic diseases and Rheumatoid / Musculoskeletal diseases are sub-setted for analysis. There are 10 metabolic diseases and 106 Rheumatoid / Musculoskeletal diseases defined by ayurvedic vaidyas, based on ACD classification
6. The following github repository stores the analysis carried out so far, details are as follows: <https://github.com/Coursephd/PostgreSQL>

**Analysis carried out using Hospital Database from April 2018 onwards**

**PostgreSQL**

Hospital data access

Creation of programs using the PostgreSQL database and SQL codes

The following files should be used in the following sequence:

**Dataset creation files:**

| **Program name** | **Description** |
| --- | --- |
| 100\_adsl\_sqlpart.sql | SQL code to download the necessary data from the server |
| 100\_adsl.R | R code to create data for patient level information: Demog + Visit + diseases + treatments + services -- Treatment names are corrected, Type\_med variable is an attempt to create logical treatment grouping |
| 02other\_data.R | SQL code is added at the top of the file and then followed by R code |
| 100\_adsl\_updt\_BasedOn\_otherData.R | Ongoing  Creates various background characteristics variables using other data,  this program uses a few other files,  Used files: lookup.csv, lookup\_backchar004\_003.txt and lookup\_backchar004\_005.txt Allopathic diagnosis variable creation: 060\_allopathic\_diag.R |

| **Excel files** | **Description** |
| --- | --- |
| \_unq\_whichdata\_touse.xlsx | This file contains information related to background characteristics and medical history, - The background history is coded by using ISO standards dictionary, - The diseases are coded using ICD10 codelist |

**Analysis creation files:**

| **Program name** | **Description** |
| --- | --- |
| Ongoing: |  |
| 100\_adsl\_analysis.rmd | R Knitr file to produce a word document / HTML file |
| 01adsl\_primary\_related\_diseases.R | Primary and related disease/medication caretsian product creation program, use in Tableau |
| 01adsl\_primary\_related\_diseases\_subgroup.R | Primary and related disease/medication by gender, used in network graph type of analysis |
| 01adsl\_primary\_related\_diseases\_seasonGender\_subgroup\_Network.R | Primary and related disease/medication by gender and season, used in network graph type of analysis |
| 20\_cooccur.R | Co-occurence analysis for the diseases, may help in printing the network graph |
| 101\_age\_sparkline.Rmd | Sparkline analysis to show many graphs at one go. |
| 01\_traxminer\_disease.R | Event sequence analysis using R |
| 05\_slopegraph.R | Edward Tufte style slopegraph to understand the trends in the data |
| 060\_allopathic\_diag.R | ICD10 coding for background diseases used in **Tableau** displays |
| 070\_Diff\_Visit\_Studyday\_analysis | Create a file for differences in 2 consecutive visits Overall and by Disease, these 2 files are used in **Tableau** displays |
| 070\_Diff\_Visit\_Studyday\_Medicine\_analysis | Create a file for differences in 2 consecutive visits by treatment, this file is used in **Tableau** displays |
| 080\_medicine\_repeat\_prop.R | Create a file to understand the first occurrence of the disease, repeat occurrences of diseases as well as for the medicines |
| 080\_medicine\_repeat\_prop\_addnl\_cal.R | Create a file to understand the first occurrence of the disease, repeat occurrences of diseases as well as for the medicines, this inlcudes number of times a disease and treatment is co-occurring |
| 085\_dis\_1st\_time\_refCal\_NodesEdges.R | This program creates **all\_met\_rmsd02** dataset, this allows the "before" and "after" view for each disease for each patient, this provides a comprehensive view of background and foreground for each disease |
| 085\_dis\_counts\_bruce\_java.R | Bruce McPherson, a programmer, has provided a Java utility to create a **collapsible view**. This is implemented using R, Java and Github, this program creates the necessary Json file. This view allows combination of diseases and treatments assigned |
| 085\_dis\_counts\_bruce\_java\_byperiod.R | Explanation similar to the above program, data by period "before" and "after" view is possible |
| 085\_dis\_counts\_edges.R | This program creates a csv file used in a Circular Tableau display, the inner circle represents the diseases, the outer circle represents the medicines given, the data is visualised for each reference disease and period |
| 085\_dis\_counts\_edges\_3rdbyPeriod\_circular17.R | Explanation similar to the above program |
| 086\_dis\_patterns\_HumanBody.R | This program creates a dataset useful for display of disease trajectories on a human body tableau display |
| 086\_dis\_patterns\_combinations.R | This program creates a dataset of disease trajectories |
| 086\_dis\_patterns\_combinations\_gender\_Macro.R | This programs creates multiple files, similarity measurements for each patient by reference disease, similarity measurements for medicines, these files are used in Tableau display |
| 086\_fisher.R | This program creates singificant combinations of disease trajectories |

**Documents created in RPubs based on the following programs [**[**http://rpubs.com/mahajvi**](http://rpubs.com/mahajvi)**]:**

| **Program name** | **Description** |
| --- | --- |
| 100\_adsl\_analysis.rmd | Metabolic and RMSD disease analysis |
| 100\_adsl\_analysis\_survival\_rmsd.Rmd | RMSD Survival analysis |
| 100\_adsl\_analysis\_survival\_met.Rmd | Metabolic Survival analysis |
| 100\_adsl\_medicine.Rmd | Metabolic and RMSD Medicine analysis |

**D3 network program and the corresponding files are: [**[**https://coursephd.github.io**](https://coursephd.github.io/)**]**

-- This tree shows progression of diseases as experienced by patients in the database.  
-- This tree shows approximately 12,500 lines of data in very short space.  
-- Some diseases are experienced more by males or by females.  
-- Some diseases have many more branches than a few others.

**Displays created in Tableau [**[**https://public.tableau.com/profile/frlht#!/**](https://public.tableau.com/profile/frlht#!/)**]**

| **Viz name** | **Description** |
| --- | --- |
| **Based on old data till July 2016** |  |
| 04\_patient\_analysis\_tablaeu | Demographic information of patients: - 01NoOfPatients - 02 Country - 03 AgeBoxplotCountry - 04 AgeBoxplotGroup - 05 NoOfVisitsBox - 05a NoOfDis\_age - 05b NoOfDis\_agebox - 06 NoOfDiseases - 07 BloodGroup |
| IndividualPatientCalendar | This visual has 3 sheets (ignore sheet 3): - IndividualPatientCal: Individual patient visit and recorded disease - MonthDiseases: This provides a frequency count of diseases by month and gender, the month and year view can be opened up and details for each day can be found |
| Primary\_disease\_and\_all\_other\_diseases | - This provides a view of diseases and other diseases reported by patients. The dashboard contains comprehensive information about summary statistics of age by gender, boxplots and bubble plot - The bubble plot is shown for each disease and additional diseases experienced |
| PrimDis\_otherDis\_ByMonth | - This display is similar to the **Primary\_disease\_and\_all\_other\_diseases** display, the disease distribution over different months is displayed. - This should provide a good idea about seasonal variations of diseases |
| 01RMSD\_MET | Detailed RMSD and Metabolic disease analysis - 01TotalPatRMSD\_Metabolic - 02AgeGroupByDisease - 03AgeDistByDisease - 08CumDisplayByDuration - 08UniqueVisWindow - 08UnqBoxAgeGender - 08CumDisplayAgeDist - 08CumDispByDurDisease - 09CumDisByDurVitals - 09CumStatsByDurVitals - 09PatProfile |

| **Viz name** | **Description** |
| --- | --- |
| **Based on data from SQL database till Oct 2017** |  |
| 01SQL\_Dis\_Med\_Ser | Detailed RMSD and Metabolic disease analysis - RMSD\_Met\_patients (Frequency table) - Visit\_Duration (Boxplot) - Patient\_Visit\_View (Patient Profile 1 row per patient) - 1stDay\_Met\_Disease (Summary statistics of metabolic disease reporting) - DisType\_Diseases (Disease frequency by gender and type) - MedType\_DisType (Medicine by gender and type) - Medicine\_DisCode (Medicine by disease, gender and type) - DiseaseByStudyDay (Frequency of diseases by each day and gender) - DiseaseByStudyDaySeason - Grishma Ritu (Frequency of diseases by each day, gender and **Rutus**) - MedByStudyDay (Medicine prescribed by each study day) - SeasonDisease (Frequency of diseases by gender and **Rutu**) - SeasonMedicine (Frequency of medicines by gender and **Rutu**) - PatProf\_season - Aamavaata (Patient Profile 1 row per patient by **Rutus**) - Box\_AgeMed - Box plot for Age and medicine - Slopegraph\_disPatients -- Number of patients per year for each disease - Slopegraph\_disVisit -- Number of visits per year for each disease - MedicineByDay -- Medicines prescribed by day |
| Diff\_Visit\_Studyday | Boxplots for difference between 2 consecutive visits - The number of days between 2 consecutive visits is quite large in initial visits, the difference reduces as the visits increase |
| Diff\_Visit\_Studyday\_ByDisease | Boxplots for difference between 2 consecutive visits for individual disease - This helps understand diseases reported more than once for individual patients. |
| Allopathic\_diag | These displays provide background disease information for 8000+ patients. |
| 080\_medicine\_dis\_repeat\_prop | These displays may be inaccurate (over representing number of treatments and diseases) |
| 080\_medicine\_dis\_repeat\_prop\_cumulative | These displays may be inaccurate (over representing number of treatments and diseases) |
| 080\_medicine\_dis\_all\_met\_rmsd\_prop | These displays provide patient level information by visit, number of medicines and diseases, 1st time reporting and repeat reporting |
| 085\_dis\_1st\_time\_refCal\_NodesEdges | These displays provide display for each reference disease by period "before" and "after" , split into month and year, frequency of distinct patients by each disease and treatment |
| 085\_dis\_count\_edges\_3rd\_byPeriod02try | This display provides a comprehensive view of disease and treatment combinations across "before" and "after" periods. The inner circle represents the diseases, the outer circle represents the medicines given, the data is visualised for each reference disease and period |
| 086\_dis\_patterns\_combinations\_distance | Diease trajectory similarity distance calculations are displayed as a matrix form, before and after each reference disease, Distance (0 dissimilar, 1 similar) is color coded. A lot of dissimilarity is visible before day 1 of reference disease, it becomes more homogeneous after day 1 |
| DistanceMeasures | Butterfly views of the similarity distance matrix, diease distance before and after, maximum distance (similarity) before and after, Maximum distance for each patient for each reference disease, distance for each disease by gender before and after day 1 |
| DistanceMeasures-Medicines | Butterfly views for medicines similar to the disease views |
| HumanBody02 | This display show disease trajectory for each reference disease on human body |

**Madhumeha**

**Madhumeha specific analysis: This is an attempt to understand patients with primary diagnosis of madhumeha "M2.0", there are approximately 1400 patients in this subset.**

| **Program name** | **Description** |
| --- | --- |
| 01\_Primary\_madhumeha.R | - This program creates a dataset for Madhumeha patients M2.0 listed as Primary diagnosis. This dataset is used in Tableau display. - The diseases and corresponding treatments are displayed on calendar days as well as Study day view in Tableau |
| 02\_Primary\_madhumeha\_comb\_medicine | - This program creates a cumulative view of diseases and corresponding treatments prescribed. There are 1026 combinations of diseases observed in the data |

**Madhumeha specific analysis in Tableau:**

| **Viz name** | **Description** |
| --- | --- |
| **Based on data from SQL database till Oct 2017** |  |
| 01\_Primary\_madhumeha | The diseases and corresponding treatments are displayed on calendar days as well as Study day view |
| 02\_Primary\_madhumeha\_comb\_medicine | A cumulative view of diseases and corresponding treatments prescribed. - There are 1026 combinations of diseases observed in the data |

**Gridhrasee**

**Gridhrasee specific analysis: This is an attempt to understand patients with primary diagnosis of Gridhrasee "V2.23", there are approximately 2000 patients in this subset. Additional analysis will be carried out subsequently.**