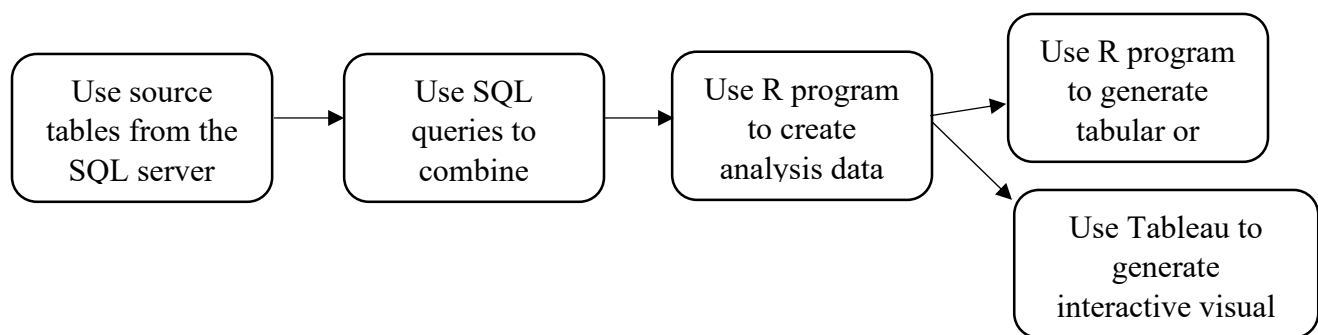# 2 Methods

2.1 Study design

This was a retrospective study of Electronic Health Records (EHR) at TDU. Electronic Health Records of patients from 2011 to 2017 are used. It contained data for more than 51,000 patients, more than 1,50,000 visits, more than 900 variations of disease types, more than 3,000 variations of medical procedures. The study was approved by the authorities of IAIM and TDU (refer Appendix section 6.1). We explored "naturally reported data" for getting insights into demographics, health-seeking behaviors, and other health parameters. Sensitive information related to patients and doctors was not extracted to maintain confidentiality. Data was analyzed through SQL [90] and R programs [91], python [92], Java [93], D3js [94] and tableau [95] software. A high-level pictorial representation of the technical study is displayed below (Figure 2-1).

Figure 2-1: Pictorial representation of analysis

2.2 Data analysis design

The data analysis was represented using many different methods such as: (1) tabular representation using frequency counts [96], (2) descriptive summary statistics [96], (3) data representation on world / country map [97], (4) boxplot representation [98], (5) barplot [99] and (6) dotplot representation [99], (7) radar plot representation [100], (8) individual patient level data listings – line by line data representation, (9) various types of bubble plots [99], [101], (10) circular data representation [102], (11) collapsible tree diagram [103], (12) treemap / mosaic plot [104], (13) butterfly plot [105], (14) area plot [99], (15) calendar plot [99].

After converting the source data into analyzable format, the next logical step is to generate various descriptive Statistics. This assessment is covered by frequencies, measures of central tendency (also called averages), and measures of variability. Frequency statistics means to count the number of times that each variable occurs. E.g., number of males and females, number of diseases reported, number of treatments prescribed, to name a few. This calculation is displayed by both the absolute or actual number and relative or percentage totals. Measures of central tendency provide a number that represents the entire data for a particular variable, such as mean, median. Measures of variability indicate the degree to which scores differ around the average. It is essential for any end user to have a sound understanding of study data. Initial statistics for all studies should include descriptive statistics [96].

For many centuries humans have used maps for various reasons. Maps are used to display geographically linked data providing a clearer and more intuitive visualizations. Maps allow to see the distribution of data in each area going from district, state, country. Technology has enabled creation of interactive maps. These allow zooming in and out, panning around, identifying certain features, querying data by topic(s) or specific indicator(s), producing reports and visualizing information in the map [97].

Boxplots were invented in the 1970s by American statistician John Wilder Tukey. They are also called as a box and whisker plot. Boxplots display five-number summary of data - the minimum, first quartile, median, third quartile, and maximum. Sometimes, the mean is also indicated by a dot or a cross on the box plot [98].

William Playfair is considered to have designed the bar plot. A bar plot is a graph that represents the category of data with rectangular bars. The length and height are proportionate to the values which they represent. The bar plots can be drawn either horizontally or vertically. One of the axes of the plot shows specific categories being compared. The other axis denotes the measured values corresponding to those categories [99].

A dot plot can be used for any graph that is translating data in a dot or small circle. A dot plot is also known as a strip plot. It is a simple form of data visualization consisting of data points plotted as dots on a graph with an x- and y-axis. These types of charts are used to graphically depict certain data trends or groupings [99].

A radar plot is a 2-dimensional representation of multivariate data as a polygon. Each variable included in the plot is represented as an axis. All axes have the same origin. The position and angle of axes are usually not informative, so any variable can be represented in any order. If different axes represent months or seasons, then the order of representation would be important as there is a certain meaning to ordering. The bars on each axis represented by the variable are called radii. A radar plot looks like an irregular polygon arranged on top of each other, all with the same center. These are used in many fields such as medicines, sports, education, and different businesses. Radar plots are also called spider plots, cobweb plots, star plots, polar plots to name a few [100].

A bubble plot is a plot that displays three dimensions of data. Each entity with its triplet variable 1, variable 2, variable 3 of associated data. 2 of the 3 variables are used on the x and y axis and a numeric variable determining bubble size is used to determine the size [99], [101].

An area plot is a variation of line plot. In this plot data points are connected by a continuous line and the area below the lines is filled with colors or textures. This plot compares two or more quantities with an area chart. Area graphs can be effective for showing the fluctuations of various data series over time [99].

A butterfly plot is a comparative bar plot or a histogram that displays the distribution of a variable for two subpopulations. A butterfly plot could be displayed either vertically or horizontally [105].

A calendar plot is a visualization used to show activity over the course of a long span of time, such as months or years. They are used to illustrate how some variable alters depending on the day of the week, or how it changes over time [99].

The treemap functions as a visualization composed of nested rectangles. This was developed by Prof. Ben Shneiderman, from the University of Maryland. The rectangles represent certain categories within a selected dimension and are ordered in a hierarchy, or "tree." Quantities and patterns can be compared and displayed in a limited chart space. Treemaps represent part to whole relationships [95], [104].

Tabular listings of data are generated at individual patient level. These are equivalent to displaying source data. They are the basis for all tables, listings, and figures (TLFs). To make the listings readable different components are data are shown in different listings. Patient profile listing covers all data for a patient in one consolidated report. This is used in data review and clinical review.

These various analyses enable data to be reported in different levels of details. Most of these representations are interactive, end user can perform filtering tasks while using the visualizations. Tableau's drill down facility provides additional ways of analyzing the data. Tooltip functionality allows extra dimension to provide more details [95].

2.3 Converting real life clinical data into analyzable format

2.3.1 Data access

Patient data was stored in the hospital database. "Read only" access was provided to the hospital database, to avoid any accidental updates to the records, thus preventing the risk of source data change or loss. The details for accessing the hospital management system are as follows:

1. Install PostgresSQL locally on the system and then connect to the database as per details below.

2. Install Cygwin terminal locally on the system.

3. Login using the Cygwin terminal (the following command will prompt for password): psql -h xx.yy.zz.ww -p ABCD -d iaim -U iaim_ro

4. Postgress Data Base details are as follows:

   - Hostname: xx.yy.zz.ww

   - port: ABCD

   - user: iaim_ro

   - password: efghijk

The real login and IP details are not presented to keep the confidentiality of secure access.

An independent (not interfering in the day-to-day transactions of the hospital), remote access for the specific version of the database was established.
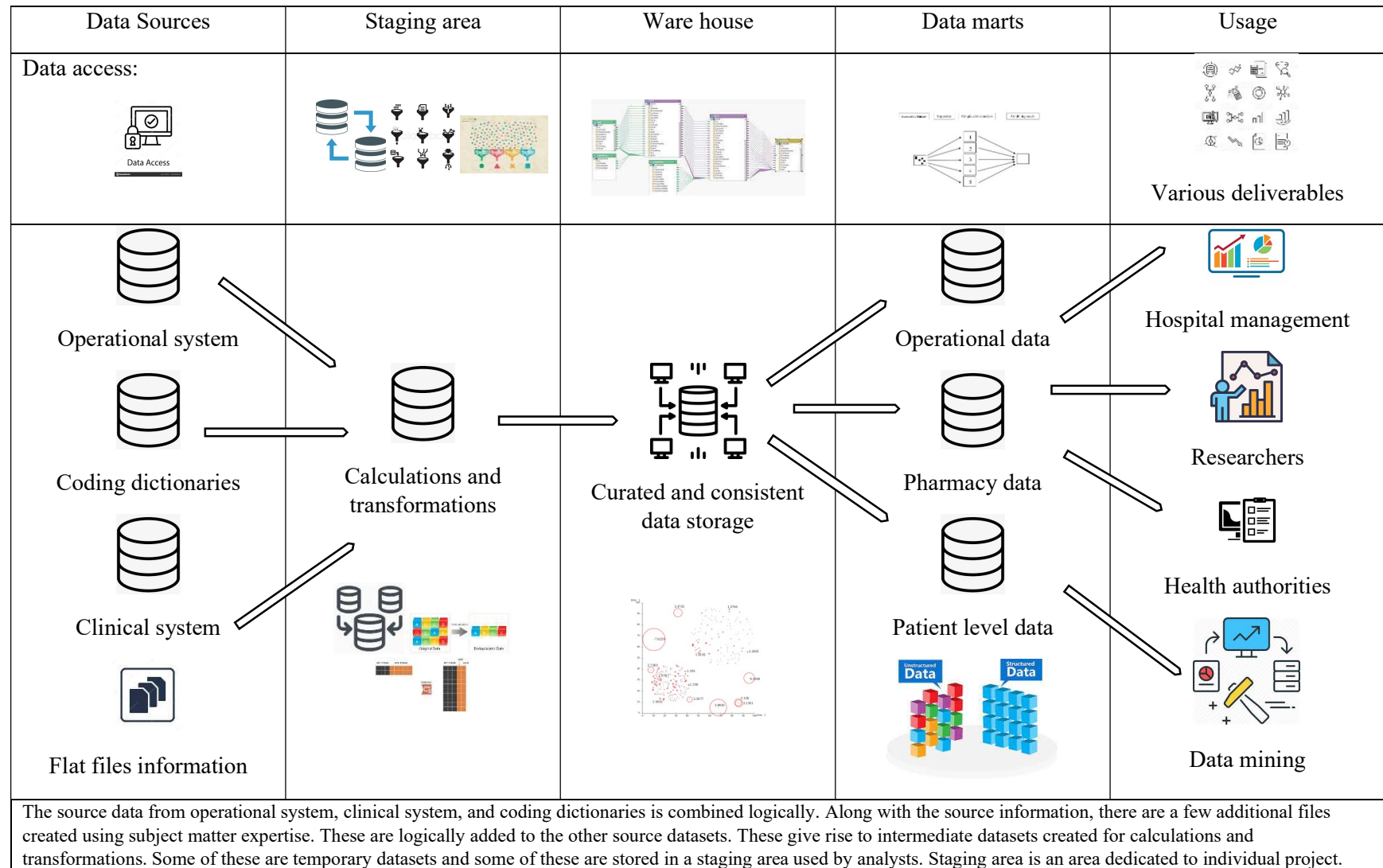
2.3.2 Data preparation

It is very important to think through the data preparation stage about the data holistically. Audience is critical while preparing data. It is important to assess, who will use the data and where and when and for what purpose. The answer to these questions determines how the data should be processed. Data preparation has a lot of different components, from restructuring to reformatting to cleaning, and should not be constrained by a specific order. This data is stored in a central place called as a data warehouse. It is a central repository of data within for an organization. Data flows into a data warehouse from various systems. This data is used to make operational, business, and scientific decisions [106]. This detail influences the data preparation process significantly, determining both the amount of effort and detail [82]. The following steps were followed in data preparation (Figure *2-2*). The relevant details can be found in individual R, SQL, python programs (refer Appendix 6.5):

- Merging, joining: Combine relevant data from different datasets into a new dataset. A "join" is an operation that connects two or more datasets by their matching columns. This establishes a relationship between multiple datasets, which merges data together so a query can be made on the combined data [84].

- Appending: Combine two or more similar datasets into a single dataset [84].

- Filtering: Rule-based reduction of a larger dataset into a smaller dataset. The goal of data filtering is to refine a data source to only what the user needs. Data filtering involves the selection of specific rows, columns, or fields to display from the dataset [82] [84].

- Deduping: Remove duplicates based on a defined criterion. Data deduplication is a data compression process to identify and remove repeated copies of information. Deduplication allows storage of one unique copy of data in the database. This process allows for examining incoming data and compares it to data that is already stored in the system [82] [84].

- Transforming: This involves converting data from one structure (or no structure) to another to integrate it with a data warehouse or with different applications [82] [84] [106].

- Format revision: Format revisions fix problems of different data types. Similar data captured in different formats creates problem for analysis. E.g., one dataset may capture treatment information as a coded numeric variable whereas another dataset may capture the same treatment information as a text. This inconsistency results in misrepresentation as well as sometimes loss of information. Along with the data format it is important to ensure the variables have appropriate lengths so that no data is truncated. Standardizing the data formats and lengths ensures correct data joins and appends. This could involve the conversion of male - female, units from one unit to another, datetime, phone number list, etc. to name a few into a consistent format. Format adjustment could involve dividing a comma-separated list into multiple columns [84].

- Categorical variable creation: This transformation is used to change a numeric series into fixed, categorical ranges, say, from {2,5,8…} to {2-5, 6-9, 10-13…}. E.g., the seasonal fluctuations in diseases using Indian seasons, RMSD and Metabolic disease group creation, treatment groupings into kashaya, asava, arka, etc. [84].

Figure 2-2: Flow diagram from data source to final usage by various usage types

Data Sources	Staging area	Ware house	Data marts	Usage
Data access:				 Various deliverables
 Operational system Coding dictionaries Clinical system Flat files information	 Calculations and transformations 	 Curated and consistent data storage 	 Operational data Pharmacy data Patient level data 	 Hospital management Researchers Health authorities Data mining
The source data from operational system, clinical system, and coding dictionaries is combined logically. Along with the source information, there are a few additional files created using subject matter expertise. These are logically added to the other source datasets. These give rise to intermediate datasets created for calculations and transformations. Some of these are temporary datasets and some of these are stored in a staging area used by analysts. Staging area is an area dedicated to individual project.				

Final output datasets are stored in a central place called as a data warehouse. It is a central repository of data within for an organization. These datasets are used to make operational, business, and scientific decisions by various stakeholders by converting them into interactive analysis, dashboards, formal project reports.

2.3.3 Data derivation

The case report form at each visit captures disease and medication data, along with demographic, background data and a few more characteristics (outlined later in the document). This data creates documented complete picture of each patient from various parts of the database including In-Patient visits, Out-Patient visits, Diseases reported as per Ayurvedic Classification dictionary, Medication prescribed and Ayurvedic services prescribed. These components of data were logically arranged in one dataset by using various data transformation steps. In addition, there were new variables derived to create necessary information for the potential analyses. Some of the challenges experienced to assemble the "reference dataset" from the source data and practical explanation of the "data preparation" steps taken are given below.

1. The database was manually explored using various SQL programming commands to check variables and observations from numerous tables (Figure *2-3*)

2. Patient information and key variables needed to be understood: unique patient ID is MR_NO, and unique ID for individual visit is PATIENT_NO (many tables containing patients' clinical information have this variable as the key variable)

3. Reference files needed to be used to reformat the coded variables

4. First section of the creation:

   - Extract relevant data tables from the source database (Figure *2-4*)

   - Transform the variables, join the tables based on logical link

   - Create "staged data" (Figure *2-5*) which can be also called as snapshot of data

   - Reference files (disease categories, Indian seasons) which were needed for calculations were developed using expert's help (Figure *2-6*)

5. Second section of the program:

   - Cleanse the tables

   - Transform the tables for combining

   - Join the tables using logical link

   - Derive additional variables as necessary

   - Filter the data using reference files created in the earlier section

6. In this process, we used 13 source datasets (5 reference datasets and 8 patient level datasets) and 72 variables to generate the necessary snapshot of the source data. These were re-arranged into 6 datasets and 40 variables. 3 additional reference files were used for further processing. 1 final dataset having 39 variables from source and 33 newly derived variables is built. (Figure *2-7*). All these steps were covered in 50 stages of programming.

7. The entire workflow is pictorially depicted in (Figure *2-8*).

8. Information about the final dataset is detailed in (Table *6-1*).

Figure 2-3: A glimpse of data tables used to store source data from the database

action_rights	diet_prescribed	hospital_technical	package_componentdetail	patient_registration	section_field_options	store_item_batch_details	test_details
admission	discharge_format_detail	icu_bed_charges	package_item_charges	patient_section_details	service_consumable_usage	store_item_details	test_org_details
anesthesia_type_charges	doctor_charges_backup	ip_bed_details	package_prescribed	patient_section_details_orig	service_documents	store_item_lot_details	test_results_master
area_master	doctor_charges_op_backup	ip_prescription	patient_activities	patient_section_forms	service_master_charges	store_patient_indent_details	test_visit_report_signatures
bed_details	doctor_consultation	item_supplier_prefer_supplier	patient_consultation_field_values	patient_section_image_details	service_master_charges_backup	store_patient_indent_main	test_visit_reports
Bill	doctor_consultation_charge	manf_master	patient_demographics_mod	patient_section_values	service_org_details	store_po	tests_conducted
bill_activity_charge	doctor_medicine_favourites	medicine_dosage_master	patient_deposits	patient_service_prescriptions	services	store_po_main	tests_prescribed
bill_adjustment	doctor_op_consultation_charge	medicine_id_health_authority_unique	patient_deposits_setoff_adjustments	patient_test_prescriptions	services_prescribed	store_reagent_usage_details	theatre_charges
bill_charge	doctor_org_details	message_recipient	patient_details	ppfv_form_detail_id	stk_chkpt	store_reagent_usage_main	diet_charges
bill_receipts	dyna_package_charges	mrd_codes_doctor_master	patient_discharge	preauth_prescription_activities	stock_issue_main	store_retail_customers	user_services_depts.
complaintslog	dyna_package_org_details	mrd_codes_master	patient_documents	prescribed_medicines_master	store_adj_details	store_sales_details	visit_vitals
consultation_charges	equipement_charges	mrd_diagnosis	patient_general_docs	progress_notes	store_adj_main	store_sales_main	vital_reading
consultation_org_details	estimate_bill	mrd_observations	patient_hvf_doc_values	registration_charges	store_checkpoint_details	store_stock_details	section_field_desc
deposit_setoff_total	estimate_charge	operation_charges	patient_medicine_prescriptions	sample_collection	store_estimate_details	store_transaction_lot_details	section_master
diagnostic_charges	favourite_reports	operation_org_details	patient_other_medicine_prescriptions	sch_resource_availability	store_grn_details	store_transfer_details	ha_item_code_type
diagnostic_charges_backup	fixed_asset_master	other_services_prescribed	patient_other_prescriptions	sch_resource_availability_details	store_grn_main	store_transfer_main	package_charges
diagnostic_reagent_usage	follow_up_details	outsource_sample_details	patient_packages	scheduler_appointment_items	store_indent_details	supp_inv_id	patient_prescription
diagnostics	growth_chart_reference_data	pack_org_details	patient_pdf_form_doc_values	scheduler_appointments	store_indent_main	supplier_master	

This table presents a glimpse of inventory of data tables in database. The cells marked in yellow are used for the generation of the analysis ready datasets.

Figure 2-4: Extraction of relevant data from Source database

	Names of source datasets
1	city
2	state_master
3	country_master
4	patient_details
5	patient_registration
6	doctor_consultation
7	mrd_diagnosis
8	patient_prescription
9	patient_medicine_prescriptions
10	ip_prescription
11	services_prescribed
12	services
13	med

Each row in the above figure is one source dataset. Each column represents a variable. The gray-coloured cell denotes the presence of the variable in the dataset. There are 13 datasets, and 72 variables represented in the above table used to derive analysis datasets

Figure 2-5: Staged data converted into 6 datasets

The source datasets have been merged step by step using the variables marked in yellow colour. The above picture shows 16 steps taken to generate 6 datasets, marked in Green for subsequent processing. The variables marked in numbered yellow squares are the logical links between the datasets and are used for data preparation steps.

Figure 2-6: Staged data

This picture lists 6 datasets created by earlier processing + 3 reference files provided by the experts. Using the source variables (gray-coloured columns), additional variables marked in Orange are created. User defined files: Disease group file: this file was created by Dr. Girish Tillu outlining the disease codes for Metabolic and Rheumatic and Musculoskeletal disease (RMSD) areas. Rutus: the calendar months are transformed into Indian rutus, https://www.drikpanchang.com/seasons/season-tropical-timings.html?geoname-id=1277333&year=2010, lookup_medicine file: this file was created by Dr. Prasan Shankar classifying medicines into groups of medicines such as: Ghritam, Kashayam, Asavam, Aristham, Bhasma, Abhyanga, Cream, Rasayanam, Tablet / Gulika / Vati, etc..

Figure 2-7: Final dataset with 39 source variables and 33 new derived variables

The above figure provides a step-by-step flow of creating the final dataset. The final dataset named "all_met_rmsd" is created through the above complex processing, which will form the basis of many analyses explained later in the thesis. This process is followed for every analysis carried out. The variables marked in numbered yellow squares are the logical links between the datasets and are used for data preparation steps.

Figure 2-8: Data flow from source data to interpretable results

Source data (SQL data file)	Staging data (csv files / R data files)	Data ware house (R data files)	Usage
city			Creation of additional analysis datasets
state_master			
country_master	base01_op (all OP data)		
patient_details	base01_ip (all IP data)		Actual analysis
patient_registration	base_01_ser (Services data)		
doctor_consultation	pat_diag_vis (temp30_5)	Longitudinal Patient data with disease, medication and Ayurvedic services information	
mrd_diagnosis	med		Learning from the existing database to be given back as learning
patient_prescription	services	~30 variables from source ~ 30 variables derived	
patient_medicine_prescriptions		~50,000 patients	
ip_prescription	Reference files for derivations and filtering of data	~17,000+ patients: subsetted version for RMSD and Metabolic	
services_prescribed	(Disease group txt file)		Clinical communication
services	(Indian rutus txt file)		
med	(Medicine type txt file)		

SQL data: source data captured in the database, staging data: logically combined intermediate datasets from source data by software development, Reference files: files needed for deriving certain information which is not present in the source database, Longitudinal data: dataset having 1 record per patient, per visit, per disease and per treatment, the dataset is filtered for the Metabolic and Rheumatic and Musculoskeletal disease (RMSD) patients, analysis is carried out using such derived dataset(s).

2.4 Clinical data understanding

2.4.1 Broad checks on the datasets

As a part of clinical understanding, structural and contents checks were performed for completeness, correctness, to identify duplication, to name a few across 90+ datasets and 500+ variables. Some of these were programmatic checks and some were manual checks to make the data available for exploration and "analysis ready". Unique values for each variable were checked to understand the value level detail for consistency and variations. The following data and contents review was done for vital sign dataset, lab measurement dataset, treatment dataset, as well as review of clinically important variables. After reviewing the source datasets for clinical understanding, derived datasets were also reviewed.

2.4.2 Contents checks

500+ variables were captured across many datasets for each visit and each patient (Table *6-3*) were classified and mapped into the following categories, (1) Ayurvedic data, (2) Background, (3) Disease, (4) Doctor's Notes, (5) Food / Exercise, (6) Hospital Visit, (7) Lab report, (8) Measurement, and (9) Treatment / Procedure. If there was any non-missing data present in a particular variable then a pseudo value "Yes" was assigned, if the data was missing then a pseudo "Blank or No" value was assigned for the purpose of analysis. This data was presented as a listing for each patient for each visit (day) by the categorization presented above. If the data was available, it was presented as a color-coded bar. If the data was missing, then it was presented as a white blank space (Figure *3-2*).

2.4.3 Visit pattern analysis

Frequency counts of 4 parameters, (1) new Out-Patients added on that day, (2) total number of patients visiting on that day, (3) total number of In-Patient visits on that day, and (4) total number of Out-Patient visits on that day were calculated for each day to understand the patient flow to hospital from year 2011 to 2016. The calculated information was represented on a calendar.

2.4.4 Patient disease and treatment journey view

Patient profile report generation module was also checked to understand the contents. Two longitudinal interactive views were created to display individual patient data. first version of patient profile contains the following information (Figure *3-4*): Patient ID (mr_no), gender, study day, In-Patient visits are displayed in blue colour and Out-Patient visits are displayed in Orange colour. The tooltip of the interactive display holds information about the following data points not displayed on the page: (1) Study day, (2) Total duration of hospital visits, (3) Disease description variable accompanying ACD codes, (4) Medicine provided at that visit, (5) Minday Metabolic: First day on which any metabolic disease has been reported by patient, (6) Minday RMSD: First day on which any RMSD disease has been reported by patient.

Second version of patient profile contains the following information (Figure *3-5*): Patient ID (mr_no), gender, base age, category, Code, description, study day. The Diseases were displayed

in blue coloured bars and treatments prescribed were marked in orange coloured bars. Disease duration and treatment duration bars were created as follows: Duration between minimum and maximum reported date for a disease as well as prescribed treatment was calculated, this duration was displayed on the visualization. The tooltip contains information about the following data points not displayed on the page: (1) Daystt: Start of event in days, (2) Disdur: Duration of event in days, (3) Disstt: Start date of event, (4) Diend: End date of event.

2.5   Studying demographics and patient specific factors

Analysis datasets created in the earlier sections (converting clinical data into analyzable format and Clinical data understanding) are used to generate necessary analysis. If the existing variables were sufficient to produce the results, then these were used as is. In case additional information was need then that was derived as appropriate. Reports using tableau software were created. Multiple types of data visualizations were used so that data was represented appropriately.

Preliminary analysis was carried out by Dr. Girish Tillu and he found that the database contains a lot of patients in Metabolic area and (Rheumatic and Musculoskeletal disease) RMSD area [7]. 10 Metabolic and 97 RMSD disease codes were identified (Table *6-2*). The analysis was split into 2 major sections in this thesis. Reports were created for the complete dataset and additional reports were created on a subset of patients' metabolic and RMSD disease areas.

Following interactive reports were created and were analyzed for the complete set of patients to gain insights into patient demographic and patient specific factors: (1) A tabular summary of total number of patients treated (Figure *3-6*), (2) Patient analysis by country – a Country-wise visualization on the world map (Figure *3-7*), (3) Age distribution by country and gender – 2 boxplot representations (Figure *3-8*), (4) A tabular summary of Blood group distribution by gender (Figure *3-9*), (5) A boxplot representation of analysis of number of visits and types of visit (IP / OP) (Figure *3-10*), (6) Number of diseases reported by gender – a descriptive summary statistics table (Figure *3-11*)

Subsequent reports were created for metabolic and RMSD patients: (7) A bubble plot data tabulation for patients reporting RMSD and Metabolic diseases (Figure *3-12*), (8) Disease distribution by Age and gender – a boxplot representation (Figure *3-13*), (9) A tabular representation of Patient visit duration for Disease categories by Gender (Figure *3-14*), using the following logic: The duration between the first visit and the last visit for each patient has been calculated and categorized as follows: >= 1 day, >= 1 month, >= 2 months, >= 3 months, >= 6 months, >= 1 year, >=2 years, >= 3 years, >= 4 years and >= 5 years. In this analysis patients were counted multiple times as per available data for each time period. A patient visiting for more than 5 years was counted in all categories. If a patient discontinued in the 4th month then that patient was counted in Day 1, >=1 month, >=2 months, >=3 months categories. The colour gradient moves from Red to Green denoting low to high number of patients in each category. (10) Seasonal Variations within Metabolic and RMSD disease areas by Indian rutus (Vasant, Grishma, Varsha, Sharad, Hemant, and Shishir) [107] and gender (Figure *3-15*). Pre and Post Disease Classification Analysis was carried out for Metabolic and RMSD disease areas to

understand the disease trajectories [108] (Figure *3-16*) The underlying data was generated from every day medical practice at the hospital. Hence the diseases were reported almost at random. The following analysis used first occurrence of any disease as day 1 for an individual patient. Using this as a reference day "before period" and "after period" was derived. "Before period" provides significant amount of "baseline data", "after period" provides specific insights into what would happen after the onset of the reference disease. The following algorithm was used to create the underlying data for analysis:

1. Each of the 107 diseases (10 Metabolic and 97 RMSD) was considered as a reference disease.
2. Day 1 was calculated as the reference day 1 for individual patient for each disease.
3. Other diseases for the same patient were arranged either before or after compared to this reference disease.
4. Duration was calculated before and after day 1, which is the reference day. This calculation provided the background view as well as future view.
5. This referencing allowed for more informative background disease as well as background medicine information. The duration was split into the following time points:

Table 2-1: Visit window table for Pre and post analysis

Before	After
Day 1 as reference	
Before 1 month	Within 1 month
Before 2 months	Within 2 months
Before 3 to 6 months	Within 3 to 6 months
Before 7 to 12 months	Within 7 to 12 months
Before 2nd year	Within 2nd year
Before 3rd year	Within 3rd year
Before 4th year	Within 4th year
Before 5 year	Within 5 year

2.6   Studying diagnostics and interventions

Diagnostics and interventions were studied using disease - disease, disease - treatment combinations / co-occurrences by using various methods. (1) Ayurvedic Classification of Disease (ACD) and International Classification of Diseases (ICD) [109] mapping exercise was carried out to understand the underlying disease burden (Figure *3-17*, Figure *3-18*, Figure 3-19, Figure *3-20*), (2) Summary table for disease by Prakriti and gender was created (Figure *3-21*), (3) Co-morbidity analysis was carried out using 3 different approaches (Figure *3-22*, Figure *3-23*, Figure 3-24, Figure *3-25*, and, Figure *3-26*), (4) Treatment and disease analysis at individual patient level was carried out to understand treatment protocol (Figure *3-27*, Figure *3-28*, Figure *3-29*), (5) Area graph representation of diseases was created to show variations related day-to-day, seasons, gender, and diseases (Figure *3-30*). (6) Mosaic plot displays were put together for disease and treatment combinations (Figure *3-31*, Figure *3-33*, Figure *3-34*). (7) Cross tabulation of prescribed treatments and disease group by gender was generated, a couple examples for specific disease conditions or specific treatment were shown to provide the utility of this analysis (Figure *3-35*). (8) Treatment regimen using bhasma is very specific to Ayurveda, an analysis was carried out to understand the duration of treatment pre and post usage of bhasma (Figure *3-36*).

Additional Disease – treatment analysis with pre and post visit window approach was performed as described: (9) Circular view representation was created (Figure 3-37, Figure 3-38), (10) Distance metrics analysis for disease trajectories and medicine trajectories were created (Figure 3-39, Figure 3-40), and (11) Multi-dimensional data representation using radar plot displays were created (Figure 3-41), (12) Dynamic bubble plot visualization was created (Figure 3-42).

International Classification of Diseases (ICD) codes were used in patient paperwork, including hospital records, medical charts, visit summaries, and bills. These codes guarantee that a patient obtains right treatment and were charged appropriately for any medical services. An attempt was made to map Ayurvedic Classification Dictionary (ACD) codes with ICD codes by manually comparing the ACD dictionary. Summary tables and boxplots summarizing the ICD classes with the frequency of patient's visits and classifying by gender along with their duration of visit were created.

Bar graph representation for disease by Prakriti and gender was created.

Co-morbidity analysis was carried out by using 3 different approaches: First approach produced a bubble plot, boxplots, summary statistics tables and frequency tables for number of other diseases reported along with the primary disease. Second approach recreated the same analysis by visit window of each month. Third approach created disease trajectories are visually displayed in a form of collapsible tree (Figure *3-22*, Figure *3-23*, Figure 3-24, Figure *3-25*, and, Figure *3-26*).

Algorithm for first approach: (1) A unique combination of Patient ID, gender and reported disease at any given time point was created. (2) Subsequently, a dataset having combination of diseases for an individual patient was created. E.g., if a patient had reported 5 unique diseases,

then all the combinations of these 5 diseases were created i.e. $^5C_2$ combinations were created i.e. 10 combinations. (3) The resulting data had the following structure: Patient ID, Disease1, Disease2, and Gender. (4) Frequency count of distinct patients was calculated for each Disease1, Disease2 combination and gender. (5) Using this data following analysis was carried out: Summary statistics of age group for each disease by gender; Boxplot of age group for each disease by gender; Bubble plot for each disease where the bubble size was determined by the count of unique patient IDs. For each disease number of other unique diseases by gender were reported. Tooltip on the bubble plot provides information about count of distinct number of patients, and summary statistics for age group. The dashboard is controlled by a "Primary Code" or a reference disease and relevant data is displayed on the page. Other bubbles in the bubble plot, display the diseases reported by this subset of patients at any point in time (these could be clinically related or unrelated or could have occurred before or after the occurrence of reference disease). The tooltip shows minimum, median, and maximum age and distinct counts of patients. A table on the left side shows number of other diseases experienced by the patient (Figure *3-22,* Figure *3-23,* Figure 3-24).

Algorithm for second approach: Same calculations for first algorithm were followed to create co-morbidities, but now in addition the time factor of month was added to get insights into seasonal variation and bubble plots by gender and month using a reference disease were created (Figure *3-25*).

Algorithm for third approach: (1) Diseases experienced by each patient were sorted by date and only the first instance of a disease was retained. This chronological list of diseases is labelled as "disease trajectory". Similar analysis when carried out for medicines, the trajectory created is labelled as "medicine trajectory" (2) For each disease trajectory the frequency counts were created and were displayed as a collapsible tree. (3) The tree has filled blue dots which open additional branches, white filled blue dots are the end of the branch, (N=xx) at each of the branches display number of patients reporting that disease trajectory. Disease trajectories were created using R programming. Final output was stored in Json file. Json file was used as the input to the D3js Java programming. Index.html file was hosted on the Github page to create the interactive page https://coursephd.github.io (Figure *3-26*).

Treatment and disease analysis at individual patient level was carried out. (1) When a disease was reported for the first time then that was counted as "first time disease reported", any subsequent repetition was counted as "Repeat". (2) When a treatment was prescribed for the very first time then that was counted "first time treatment prescribed", any subsequent repetition was counted as "Repeat". (3) These two calculations were repeated throughout the complete duration for each patient.

Area graph representation of diseases was created to show variations related to day-to-day changes, seasonal changes, by gender, and diseases. This analysis provides frequency count of patients for each disease by month and by gender. (1) Unique combinations of patients, diseases by date and gender were created, (2) Frequency counts of females were displayed in blue color and counts for males were displayed in orange color. (3) this visual opens for each day, by

clicking on "+" sign on the x-axis, providing monthly to weekly to daily view without having to go through multiple visualizations.

A mosaic view of disease and intervention was created to explore the following: (1) Total number of interventions prescribed during one disease. (2) To check if there were possible relationships between different diseases and interventions considering multiple diseases reported and multiple interventions prescribed, layers of visualizations were created in the following manner: (1) TreeMapDisMed-Parameter sheet is used to filter a particular disease, which is displayed as a green colored box, (2) smaller boxes inside each disease display one intervention each, (3) Medicine-count sheet provides information on total number of different interventions prescribed for the selected disease, (4) Medicine-list sheet shows a detailed list of interventions with the total patients prescribed with them as well as total patients suffering from the disease (Figure *3-31*).

TreeMapDisMed-Parameter sheet can be used to filter a particular intervention and the whole analysis could be performed from an intervention's perspective. (1) TreeMapDisMed-Parameter sheet is used to filter a particular intervention, which was displayed as a single or multiple green colored boxes across multiple disease boxes, (2) Disease-count sheet provides information on total number of different diseases for which this medicine was prescribed, (4) Medicine-list sheet shows a detailed list of diseases with the total number of patients prescribed with the interventionas well as total patients suffering from different diseases (Figure *3-34*).

Cross tabulation of prescribed treatments and disease group by gender was also generated. The interactive visualization was used to create a few examples for specific disease conditions or specific treatment. First example is created using Balaristham and second example is created using bhasma. An attempt was made to understand the impact of usage of Bhasma on patient visit duration. Dataset for individual patients with the following variables was created: Visit duration: cdur – "Total duration in days": cdur = End visit date – start visit date +1, "Pre bhasma duration": prebhasmadur = bhasmamin (start date of bhasma intake) – 1, "Post bhasma duration": postbhasmadur = cdur – bhasmamin (start date of bhasma intake) + 1.

A simple t-test analysis was performed on the created data. The pre bhasma duration and post bhasma duration was analyzed by t-test. This analysis can be carried out by an appropriate non parametric test as well.

Analysis for disease – treatment with pre and post visit window approaches:

The circular visualization allows a single page view of relation between disease – disease and / or disease – treatment across multiple time points. This view shows the following information: (1) A table on the middle row: On day 1 of a disease how many distinct diseases have been reported and how many distinct medicines prescribed, this same information is shown as the green bars inside a circle, (2) Pre and post time windows are displayed and for each of the time window a similar table is represented in the upper section of the visualization. (3) In the lower section of the visualization, 1st row represents the co-occurrence of disease – disease and / or

disease – treatment before day 1 of the reference disease. (4) Last row represents the same co-occurrence data after day 1 of the reference disease (Figure *3-37,* Figure *3-38*).

An attempt was made to understand the disease trajectories for patients by using mathematical distances. There are numerous distance measures available in mathematics and statistics which allows understanding of similarity and dis-similarity between objects, in our case disease trajectories [110].

Following assumptions were used to derive the disease trajectory: (1) Diseases experienced by each patient were sorted by date and only first instance of a disease was retained. (2) This enabled in creation of a disease trajectory for each patient for each reference disease, before and after the occurrence of the reference disease. (3) Cartesian product of patients was created for each reference disease, so that distances could be calculated. A cartesian product is a set of all possible pairs, in this case all possible pairs of reference disease and other diseases for individual patient. (4) The similarity measure was calculated for each disease trajectory, e.g., Jaccard distance was used as a distance measure for this display [111]. The Jaccard distance highlights similarity between finite sample sets. It is defined as the size of the intersection divided by the size of the union of the sample sets [112]. For our example, this will be calculated as the common diseases reported divided all the diseases reported in each case. (5) Jaccard distance closer to 0 shows dissimilarities and closer to 1 show similarities. (6) The distances were divided into 4 categories 0 to 0.25, 0.25 to 0.5, 0.5 to 0.75 and 0.75 to 1 for data visualization perspective. (7) These calculated distances are displayed as a butterfly plot for easy comparison of underlying values [105]. This plot is a comparative bar plot to display comparison of a continuous variable across groups. Similar Analysis to understand the medicinal trajectory was performed.

Radar plot representation: a multidimensional, comparative view of the different diseases was created to understand at various aspects of the diseases. The radar plot chart presents multidimensional metrics. Radar plots can convey a large amount of information [100]. They provide a standardized view of different indicators on one scale. The following information for each disease was visualized as a percentile and is represented as a dimension on a heptagon (as there are 7 parameters considered in this example): (1) Distinct number of patients for each disease, (2) Number of times a disease is reported, (3) Number for a specific disease (chronological number of disease reported by a patient) e.g. a disease is reported as the very first disease or third disease or fifth disease, etc., (4) Number of diseases before the specific disease, (5) Number of diseases after the specific disease, (6) Number of treatments before the specific disease, (7) Number of treatments after the specific disease. Trellis plot display allows multiple representations of same kind next to each other [99].

Dynamic bubble plot visualization: explanation of an algorithm using Amavaata (ACD code A6.0) as an example: (1) Identified unique patients who have had Amavaata reported at least once, (2) All the other diseases and prescribed medicines for this subset of patients, (3) Created an input Json file to be passed into a D3js java program. The underlying utility generated a

dynamic bubble plot. The size of bubble is proportional to the total number of patients. The links display relationships between diseases and treatments. If a bubble is "double clicked" then all the "unrelated data" to that bubble vanishes and only relevant data is retained on the screen. Once double clicked again the complete data is displayed again (Figure *3-42*).

Same type of analysis was carried out for the Pre and post period, example for Amavaata (A6.0) by period [https://coursephd.github.io/nodediagram/A2_0byperiod/] Similar, views can be created for any number of diseases and treatments, links below provide similar examples for the disease Prameha (P5.0) and its treatments and comorbidities

Prameha [https://coursephd.github.io/nodediagram/P5_0_Prameha/]
Prameha by period [https://coursephd.github.io/nodediagram/P5_0_Pramehabyperiod/]

2.7 Summary of methods section
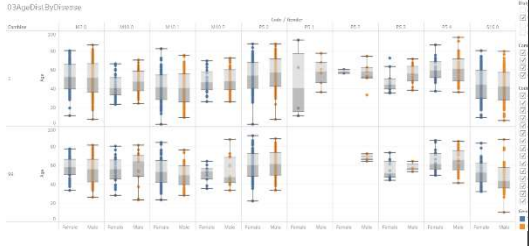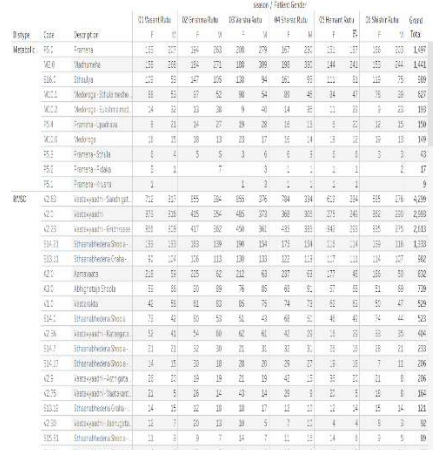
Various methods getting employed have been explained in the earlier sections of chapter 2. These data explorations can be used by different stake holders like Hospital management (HM) from administration point of view, treating doctors (Medics), and by basic researchers (Scientists). Table below (Table *2-2*) systematically presents a consolidated view of various analysis, context about different analysis and their possible relationships with each other. The screenshot attached in the table is not of the best resolution but is only representative of the actual analysis. There is a hyperlink provided below each of the figure which takes the user to the actual analysis stored on Tableau website. The user would have freedom to zoom in and out the visualizations as well as additional interactive abilities offered by the system.
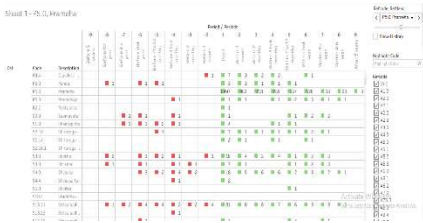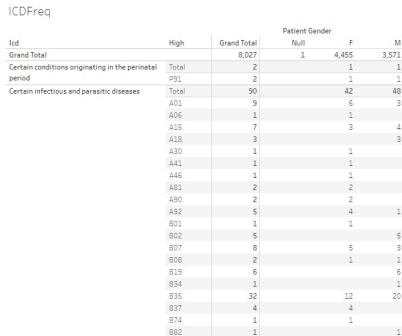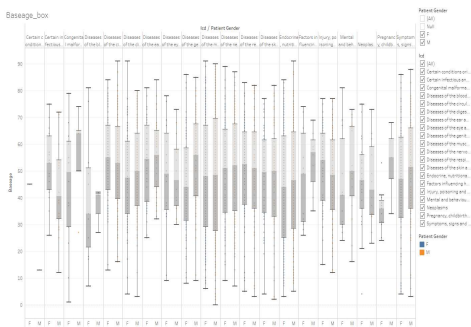
Table 2-2: Proposed methods and analysis use cases
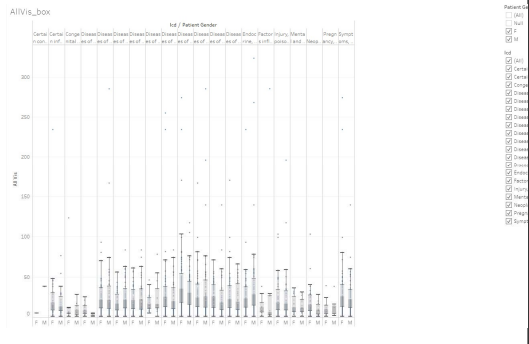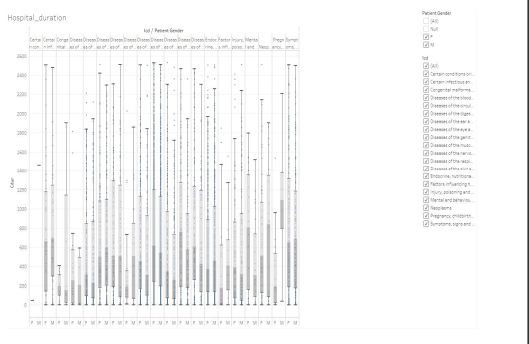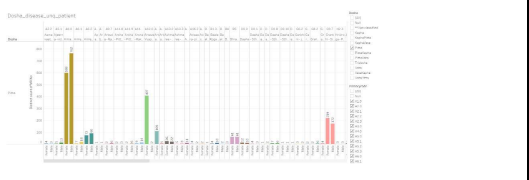
Stake holders	Classification	Type of analysis	Figure number and analysis name (linked to the actual figure)	Snapshot of the analysis (the picture is attached here to provide a quick look into the type of analysis)	Context of the proposed analysis
HM, Medics, Scientists	Clinical data understanding	Tabular frequency table	Figure 3-1: A snippet of disease table by gender		(1) Data quality check based on examples of a couple of diseases. (2) Generation of ideas about how "end users" are using the system
HM, Medics, Scientists	Clinical data understanding	Individual patient level data listing	Figure 3-2: Variable classification by categories		(1) Understanding data generation process at each visit for each patient to gain present status of data and provide feedback into improvements of system architecture. (2) Improve operational efficiencies of the Hospital Management Information system
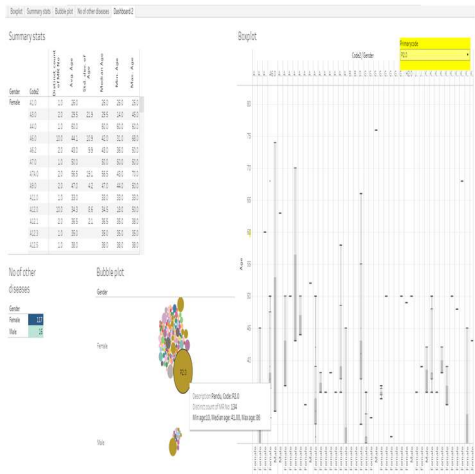
Stake holders	Classification	Type of analysis	Figure number and analysis name (linked to the actual figure)	Snapshot of the analysis (the picture is attached here to provide a quick look into the type of analysis)	Context of the proposed analysis
HM	Clinical data understanding, operational efficiencies	Calendar plot	Figure 3-3: Visit pattern analysis		(1) Learn more about patient inflow and study how to increase outreach to the society. (2) Improve operational efficiencies across various departments based on patient visit patterns
Medics	Patient disease and treatment journey	Individual patient level data listing	Figure 3-4: Patient visit profile – Horizontal view		(1) An analysis of how a patient journey is documented, how can this study be used prospectively and retrospectively to understand disease progression and treatment protocols (2) Convert treatment ideas into documented material (3) Improvements to the data standards and system architecture
Medics		Individual patient level data listing	Figure 3-5: Patient visit profile – Vertical view		
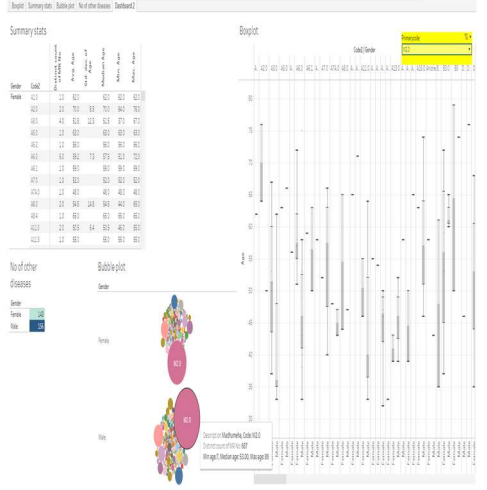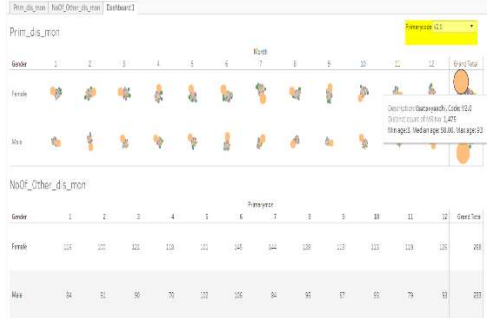
Stake holders	Classification	Type of analysis	Figure number and analysis name (linked to the actual figure)	Snapshot of the analysis (the picture is attached here to provide a quick look into the type of analysis)	Context of the proposed analysis
HM, Medics, Scientists	Study of demographics and patient specific factors	Tabular frequency table	Figure 3-6: Total Number of Patients		(1) Demographic profiling analysis to understand health seeking behaviours of patients
HM, Medics, Scientists		Data on world map	Figure 3-7: Country-wise Visualization		(2) Disease profiling based on preliminary disease analysis, visit patterns and types of visits (In-patient and out-patient visits) (3) Use such analysis to compare against demographic profiling of any other main-stream hospital to understand health seeking behaviours (4) Use disease profiling study to understand epidemiology (5) Use this analysis for defining public health policies to local and central authorities
HM, Medics, Scientists		Boxplot	Figure 3-8: Age distribution by country, age distribution by gender		(6) Secondary use of this analysis is to find data quality
HM, Medics, Scientists		Tabular frequency table	Figure 3-9: Blood-group Distribution by gender		

Stake holders	Classification	Type of analysis	Figure number and analysis name (linked to the actual figure)	Snapshot of the analysis (the picture is attached here to provide a quick look into the type of analysis)	Context of the proposed analysis
HM		Boxplot	Figure 3-10: Number of Visits, and Visit Types		
HM, Medics, Scientists		Descriptive summary statistics	Figure 3-11: Descriptive summary statistics by number of Diseases by Age and Gender		
HM, Medics	Study of demographics and patient specific factors	Bubble plot	Figure 3-12: Data tabulation for patients reporting RMSD and Metabolic diseases		Analysis similar to described above on a subset of Metabolic and Rheumatic and Musculoskeletal disease (RMSD) patients

Stake holders	Classification	Type of analysis	Figure number and analysis name (linked to the actual figure)	Snapshot of the analysis (the picture is attached here to provide a quick look into the type of analysis)	Context of the proposed analysis
HM, Medics		Boxplot	Figure 3-13: Disease distribution by age and gender		
HM, Medics		Tabular frequency table	Figure 3-14: Patient visit duration for Disease categories by Gender		
HM, Medics	Disease pattern analysis	Tabular frequency table	Figure 3-15: Disease distribution by Seasonal Variations and gender		(1) Diseases are differently experienced by females and males as well as natural variations affect the prevalence – can we detect if the natural variations are reported in our database (2) A byproduct of this analysis is understanding of data entry and disease classification process at each visit
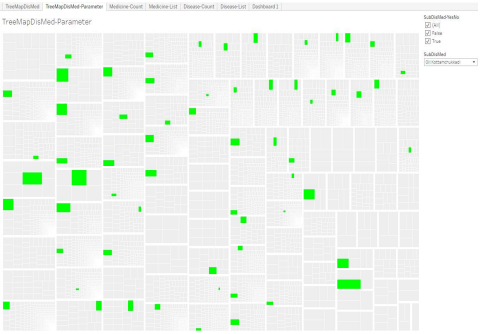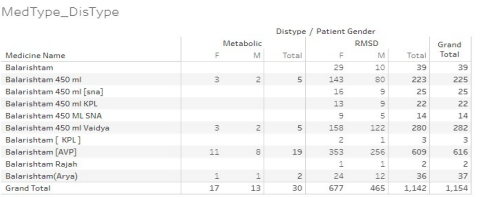
Stake holders	Classification	Type of analysis	Figure number and analysis name (linked to the actual figure)	Snapshot of the analysis (the picture is attached here to provide a quick look into the type of analysis)	Context of the proposed analysis
HM, Medics	Co-morbidity analysis	Tabular frequency table	Figure 3-16: Pre and Post Disease Classification Analysis		(1) Discover disease relationships by creating chronological view (2) Differentiate pre and post diseases so that diagnosis and prognosis can be formed (3) Pre and post differentiation is carried out for the prescribed medicines so that use of medicines could be understood at a summary level
HM, Medics, Scientists	Health seeking behaviour and public policy	Tabular frequency table	Figure 3-17: ICD classification by Gender		(1) ICD classification of disease is used globally to understand the disease burden. What patterns emerge via this analysis, what health seeking behaviors can be understood?
HM, Medics, Scientists		Boxplot	Figure 3-18: Age distribution by ICD classification and Gender		
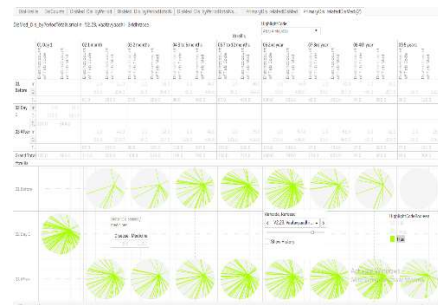
Stake holders	Classification	Type of analysis	Figure number and analysis name (linked to the actual figure)	Snapshot of the analysis (the picture is attached here to provide a quick look into the type of analysis)	Context of the proposed analysis
HM, Medics, Scientists		Boxplot	Figure 3-19: Visit distribution by ICD classification and Gender		
HM, Medics, Scientists		Boxplot	Figure 3-20: Duration distribution by ICD classification and Gender		
Medics, Scientists	Disease and Prakriti analysis	Barplot	Figure 3-21: Disease classification by Prakriti and Gender		What does Prakriti and disease data
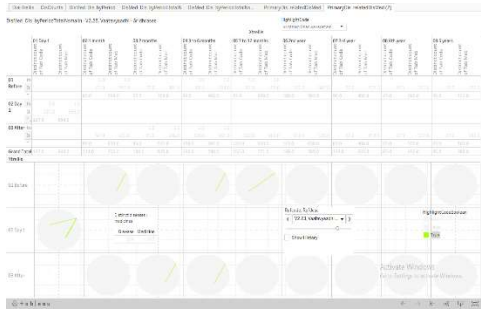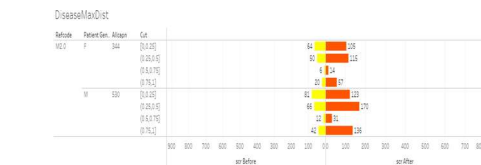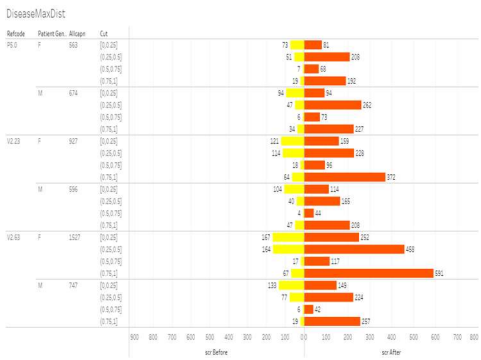
Stake holders	Classification	Type of analysis	Figure number and analysis name (linked to the actual figure)	Snapshot of the analysis (the picture is attached here to provide a quick look into the type of analysis)	Context of the proposed analysis
Scientists	Co-morbidity analysis	Bubble plot + Boxplot + Descriptive summary statistics	Figure 3-22: Co-morbidity analysis approach 1 example 1: Vaatavyadhi		Disease co-morbidities generate insights
Scientists		Bubble plot + Boxplot + Descriptive summary statistics	Figure 3-23: Co-morbidity analysis approach 1 example 2: Pandu		
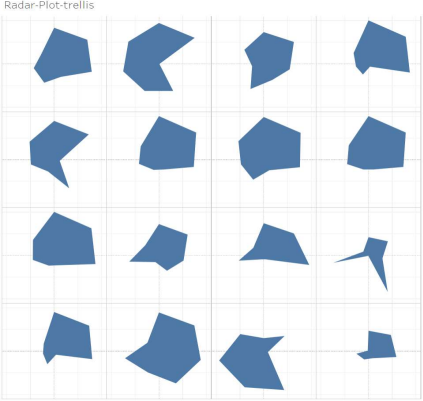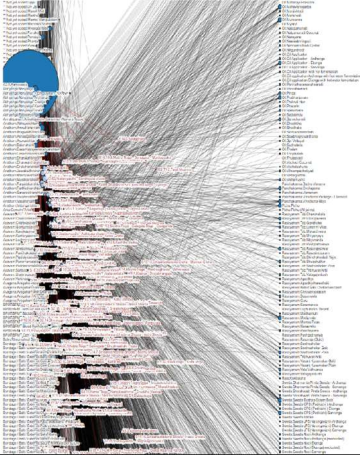
Stake holders	Classification	Type of analysis	Figure number and analysis name (linked to the actual figure)	Snapshot of the analysis (the picture is attached here to provide a quick look into the type of analysis)	Context of the proposed analysis
Scientists		Bubble plot + Boxplot + Descriptive summary statistics	Figure 3-24: Co-morbidity analysis approach 1 example 3: Madhumeha		
Scientists		Bubble plot + Tabular frequency table	Figure 3-25: Co-morbidity analysis approach 2		

Stake holders	Classification	Type of analysis	Figure number and analysis name (linked to the actual figure)	Snapshot of the analysis (the picture is attached here to provide a quick look into the type of analysis)	Context of the proposed analysis
Medics, Scientists		Collapsible tree diagram	Figure 3-26: Co-morbidity analysis approach 3: collapsible tree view		
Medics	Patient disease and treatment journey	Individual patient level data listing + Tabular frequency table	Figure 3-27: Patient Disease and Treatment administration by Study Day		(1) An analysis of how a patient journey is documented, how can this study be used prospectively and retrospectively to understand disease progression and treatment protocols (2) This could help study the severity of the disease, co-morbidities and the number of medications prescribed to treat the condition (3) This can also provide an overview of the practicing physician's style of treatment and may be help draw parallels in treating medical conditions (4) Improvements to the data standards and system architecture is a byproduct of this analysis
Medics		Individual patient level data listing + Tabular frequency table	Figure 3-28: Patient Disease by Study Day and Treatment administration by Study Day		

Stake holders	Classification	Type of analysis	Figure number and analysis name (linked to the actual figure)	Snapshot of the analysis (the picture is attached here to provide a quick look into the type of analysis)	Context of the proposed analysis
Medics		Individual patient level data listing + Tabular frequency table	Figure 3-29: Patient Cumulative Disease and Treatment administration by Visit		
Medics, Scientists	Disease pattern analysis	Area plot	Figure 3-30: Area graph representation of diseases		(1) Due to this analysis representation, information on many diseases can be viewed in very short space (diseases plotted side by side) (2) Diseases vary by seasons, by gender as well as some diseases are more prevalent than others which can be seen, and clinical interpretations can be drawn (3) Operational and clinical insights can be generated with help of this visualization
Medics, Scientists	Co-morbidities and concomitant medications	Mosaic plot	Figure 3-31: Mosaic plot: Disease and treatment representation example 1: Prameha		(1) In ayurveda, a treatment is used for multiple diseases and multiple treatments are used for the same disease based on the context of disease. This relationship gives rise to many to many associations between disease and treatment (2) Many to many relationships which are hard to visualize are generated through this analysis

Stake holders	Classification	Type of analysis	Figure number and analysis name (linked to the actual figure)	Snapshot of the analysis (the picture is attached here to provide a quick look into the type of analysis)	Context of the proposed analysis
Medics, Scientists		Tabular frequency table	Figure 3-32: Disease and treatment example 2: P5.0: Prameha and Oil: Kottamchukkadi		(3) This analysis produces a mosaic display for diseases showing what kinds of treatments are prescribed (4) Another mosaic display is created for treatments showing what kinds of diseases are getting treated by the underlying treatment (5) Clinically meaningful as well as not so meaningful relationships are visualized, rare disease – disease combinations or disease – medicine combinations also can be studied
Medics, Scientists		Tabular frequency table	Figure 3-33: Disease and treatment example 3: P5.0: Prameha and Vati: Diabecon DS		
Medics, Scientists		Mosaic plot	Figure 3-34: Mosaic plot Disease and treatment representation example 4: Treatment: Oil: Kottamchukkadi		
Medics, Scientists		Tabular frequency table	Figure 3-35: Cross tabulation of prescribed treatments and disease group by gender Example 1		

Stake holders	Classification	Type of analysis	Figure number and analysis name (linked to the actual figure)	Snapshot of the analysis (the picture is attached here to provide a quick look into the type of analysis)	Context of the proposed analysis
Medics, Scientists		Tabular frequency table	Figure 3-36: Cross tabulation of prescribed treatments and disease group by gender Example 2		
Medics, Scientists	Co-morbidities and concomitant medications	Circular data representation	Figure 3-37: Circular view: Co-occurrences of disease – disease Example 1		(1) This analysis provides a view for a disease – disease combination or disease – medicine combination longitudinally (day 1 of disease, diseases reported at different time points before and after day 1) (2) Disease and treatment information is represented on a circular display with green spokes representing co-occurrences of disease combination or disease – medicine

Stake holders	Classification	Type of analysis	Figure number and analysis name (linked to the actual figure)	Snapshot of the analysis (the picture is attached here to provide a quick look into the type of analysis)	Context of the proposed analysis
Medics, Scientists		Circular data representation	Figure 3-38: Circular view: Co-occurrences of disease – treatment Example 2		(3) Clinically meaningful as well as not so meaningful relationships are visualized, rare disease – disease combinations or disease – medicine combinations also can be studied
Scientists	Similarity analysis in disease and medicine trajectories	Butterfly plot	Figure 3-39: Pre and Post distance analysis for disease: M2.0: Madhumeha		(1) Disease trajectories (chronological list of diseases reported) and medicine trajectories (chronological list of medicines prescribed) are generated to study if similar diseases are experienced after an onset of a particular disease
Scientists		Butterfly plot	Figure 3-40: Pre and Post distance analysis for medicines given for diseases: P5.0, V2.23, V2.63		(2) This analysis helps in understanding biological changes, represented by subsequent reported diseases triggered by an underlying disease (3) Similarly, which medicines are prescribed provide insights into treatment protocols

Stake holders	Classification	Type of analysis	Figure number and analysis name (linked to the actual figure)	Snapshot of the analysis (the picture is attached here to provide a quick look into the type of analysis)	Context of the proposed analysis
Scientists	Multi-dimensional data analysis	Radar plot	Figure 3-41: Radar plot		(1) This analysis developed showed 7 parameters on 7 vertices. Different diseases were displayed next to each other as trellis radar display. (2) If there are different shapes for different diseases then this suggests that there are underlying differences in the data representing differences in diseases.
Scientists	Co-morbidities and concomitant medications	Dynamic bubble plot	Figure 3-42: Dynamic bubble plot: Example 1: Disease: A6.0: Amavaata		(1) Intricate relationship between disease and medicines in a very short space, at present this type of data representation approach may not have a direct application, but consultation with "end users" may provide appropriate use case.