Analysis of hospital based Ayurvedic clinical practice to gain Real World data knowledge

Vinay Mahajan

PhD thesis

Guides: Dr. Ashwini Godbole,

Dr. Girish Tillu,

Dr. Ashwini Mathur

Table of Contents

[Introduction 4](#_Toc106879343)

[Methods 5](#_Toc106879344)

[Converting real life data into analyzable format 5](#_Toc106879345)

[Study design 5](#_Toc106879346)

[Data access 5](#_Toc106879347)

[Data derivation 5](#_Toc106879348)

[Clinical data understanding 7](#_Toc106879349)

[Studying demographics and patient specific factors 7](#_Toc106879350)

[Studying diagnostics and interventions 8](#_Toc106879351)

[Results 9](#_Toc106879352)

[Converting real life data into analyzable format 9](#_Toc106879353)

[Details of the database 9](#_Toc106879354)

[Data Extracted from Hospital Database 9](#_Toc106879355)

[Clinical data understanding 10](#_Toc106879356)

[Studying demographics and patient specific factors 12](#_Toc106879357)

[Studying diagnostics and interventions 13](#_Toc106879358)

[Discussions 14](#_Toc106879359)

[Converting real life data into analyzable format 14](#_Toc106879360)

[Clinical data understanding 14](#_Toc106879361)

[Studying demographics and patient specific factors 19](#_Toc106879362)

[Studying diagnostics and interventions 20](#_Toc106879363)

[Conclusion 20](#_Toc106879364)

[Appendix 23](#_Toc106879365)

[Approval from the hospital management to carry out the retrospective study 23](#_Toc106879366)

[11.1 ANNEXURE 1 –NOC FROM IAIM MEDICAL DIRECTOR. 23](#_Toc106879367)

[1.1.1. Total Number of Patients 24](#_Toc106879368)

[1.1.2. Country-wise Visualization 24](#_Toc106879369)

[1.1.3. Age and Gender Distribution 25](#_Toc106879370)

[1.1.4. Blood-group Distribution 26](#_Toc106879371)

[1.1.5. Number of Visits, and Visit Types 27](#_Toc106879372)

[1.1.6. Number of Diseases by Age and Gender 27](#_Toc106879373)

[1.1.7. Data tabulation for patients reporting RMSD and Metabolic diseases 28](#_Toc106879374)

[1.1.8. Disease distribution by Age and gender 28](#_Toc106879375)

[1.1.9. Patient visit duration for Disease categories by Gender 29](#_Toc106879376)

[1.1.10. Seasonal Variations 30](#_Toc106879377)

[1.1.11. Pre and Post Disease Classification Analysis 30](#_Toc106879378)

[1.1.11.1. Staged data converted into 6 datasets 35](#_Toc106879379)

[4.1.1.1. Staged data (6 datasets, ~40 variables) + 3 reference files created based on inputs from experts 36](#_Toc106879380)

[4.1.1.1. 1 Final dataset with ~30 source variables and ~30 new derived variables 37](#_Toc106879381)

[4.1.1.1. Data flow 38](#_Toc106879382)

[4.1.1.1. Details of the Reference Dataset “01adsl\_met\_rmsd” 39](#_Toc106879383)

[1.1.11.1.1. Metabolic and RMSD disease code and de-code 45](#_Toc106879384)

[References 52](#_Toc106879385)

# Introduction

# Methods

## Converting real life data into analyzable format

### Study design

This is a retrospective study of Electronic Health Records (EHR) at TDU. Electronic Health Records of patients from 2011 to 2017 are used. It contains data for >51,000 patients, >1,50,000 visits, > 900 variations of disease types, >3,000 variations of medical procedures. In this study, we will explore “naturally reported data” from this database for getting insights into demographics, health-seeking behaviors and other health parameters. Sensitive information related to patients and doctors is not extracted to maintain confidentiality.

### Data access

Patient data is stored in the hospital database. “Read only” access was provided to the hospital database, in order 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 54.244.12.255 -p 5432 -d iaim -U iaim\_ro
4. Postgress Data Base details are as follows:
   * Hostname: 54.244.12.255
   * port: 5432
   * user: iaim\_ro
   * password: a1b2c3

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

### 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, Ayurvedic services prescribed. These components of data are logically arranged in one dataset by using various data transformation steps. In addition, there are new variables derived to create necessary information for the potential analyses. Let us go through the challenges experienced to assemble the “reference dataset” from the source data and practical explanation of the “data preparation” steps.

1. The database was manually explored using various SQL programming commands, the variables and observations were checked from numerous tables
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 (Refer picture [2.4.1.1](#_Picture_option_3a:))
* Transform the variables, join the tables based on logical link
* Create “staged data” or “snapshot of source data” (Refer picture [2.4.1.2](#_Staged_data_converted))
* Reference files (disease categories, Indian seasons) which are needed for calculations are developed using expert’s help (Refer picture [2.4.1.](#_Staged_data_(6)3)

1. 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

1. In this process, we have used 13 source datasets (5 reference datasets and 8 patient level datasets) and ~65 variables to generate the necessary snapshot of the source data. These have been re-arranged into 6 datasets and ~40 variables. 3 additional reference files are used for further processing. 1 final dataset having ~30 variables from source and ~30 newly derived variables is built. (Refer picture [2.4.1.4](#_1_Final_dataset)).
2. The entire workflow is pictorially depicted in [2.4.1.5](#_Data_flow).
3. The information about the final dataset is detailed in [2.4.1.6](#_Details_of_the).
4. Refer table [2.4.1.7](#_Other_Visit_Wise) which shows the names of the datasets based on the data captured during the patient’s clinical visit. The yellow parts highlighted in the table shows Ayurveda specific datasets apart from the rest of the available hospital data.
5. The data encountered here is in the long format, each observation is one time point per patient per variable. Along with the disease and treatment information, many variables get collected at each visit. This data was saved into a few additional datasets, namely: patient\_section\_details, patient\_section\_values, and corresponding reference files are section\_master, section\_field\_options, section\_field\_desc, patient\_consultation\_field\_values. Each patient has multiple observations, with individual columns containing section number, visit information, date information, context (for which variable the value is captured) and the last column containing the actual response to the variable.
6. All these datasets combined provided us with 90+ datasets and 500+ variables Refer table [2.4.1.8](#_The_long_version) which shows example source datasets used for processing one of the final reference datasets.

## Clinical data understanding

Programming efforts are essential to make the data available for exploration and “analysis ready”. In our case, we have seen approximately 50 stages to create analyzable datasets as well as reference datasets from the pure source database. Various datasets containing important clinical information were reviewed either manually or programmatically. The following section outlines what was done as a part of clinical data review: vital sign dataset, lab measurement dataset, treatment dataset, review of clinically important variables across datasets. Patient profile report generation module was also checked. After reviewing the source datasets for clinical understanding, derived datasets were also reviewed.

## 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 are sufficient to produce the results, then these will be used as is. In case additional information is need then that will be derived as appropriate.

Reports using tableau software are created. Multiple types of data visualizations are used so that data is represented appropriately.

Preliminary analysis was carried out by Dr. Girish Tillu and he had found that the database contains a lot of patients in Metabolic area and (Rheumatic and Musculoskeletal disease) RMSD area. The analysis is split into 2 major sections in this chapter. Reports are created for the complete dataset and additional reports are created on a subset of patients’ metabolic area and RMSD area.

Following interactive reports are created and are analyzed for the complete set of patients to gain insights into patient demographic and patient specific factors: (1) Total number of patients treated – a tabular summary (Fig 4.3.1), (2) Patient analysis by country – a Country-wise visualization on the world map (Fig 4.3.2), (3) Age distribution by country and gender – 2 boxplot representations (Fig 4.3.3), (4) A tabular summary of Blood group distribution be gender (Fig 4.3.4), (5) Analysis of number of visits and types of visit (IP / OP) – a boxplot representation (Fig 4.3.5), (6) Number of diseases reported by gender – a descriptive summary statistics table (Fig 4.3.6)

Subsequent reports are created for metabolic and RMSD patients: (7) A bubble plot data tabulation for patients reporting RMSD and Metabolic diseases (Fig 4.3.7), (8) Disease distribution by Age and gender – a boxplot representation (Fig 4.3.8), (9) A tabular representation of Patient visit duration for Disease categories by Gender (Fig 4.3.9), (10) Seasonal Variations within Metabolic and RMSD disease areas by Indian rutus and gender (Fig 4.3.10). Pre and Post Disease Classification Analysis is carried out for Metabolic and RMSD disease areas to understand the disease trajectories (Fig 4.3.11) This data is generated from every day medical practice at the hospital. Hence the diseases are reported almost at random. The following analysis uses 1st occurrence of any disease as day 1 for an individual patient. Using this as a reference day “before period” and “after period” is 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. Algorithm to create the underlying data for analysis:

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

| **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 |

## Studying diagnostics and interventions

# Results

## Converting real life data into analyzable format

### Details of the database

The database had approximately 200 datasets. They covered various components of the hospital’s day-to-day functions right from the operational data to the patient level clinical information. The high level of classification of data types:

1. Operational datasets:
   1. Hospital charges – IP, OP
   2. Operation theater charges
   3. Inventory of equipment
   4. Doctor charges, etc.
2. Reference dictionaries
   1. Disease codes
   2. Ayurvedic services
   3. Medication names
   4. Master list of Laboratory tests
   5. Names of city, state, countries
3. Doctor details
   1. Doctor ID
   2. Relevant ward information
   3. Internal / Visiting / Part time / Full time
4. Patient information
   1. Patient details,
   2. Visit details
   3. Vital signs
   4. Registration details
   5. Discharge details
   6. Lab data details
   7. Diet details, etc.
5. Datasets related to managing access levels and other IT related contexts

For this study, the following data was not used to in accordance with the patient data protection and privacy, financial privacy as well as hospital management confidentiality thus avoiding any controversies:

1. Hospital monetary details
2. Doctor’s details
3. Patient details of sensitive nature – name, phone number, socio economic status, etc.

### Data Extracted from Hospital Database

In our study, we had different versions of data, details in the table below.

|  |  |  |  |
| --- | --- | --- | --- |
| Data version | Version 1 | Version 2 | Version 3 |
| Approach | CSV files provided by the Hospital IT support | Data extraction via the SQL DB connect | PDF files provided by the Hospital IT support |
| Date time frame | From start of the hospital to Oct 2016 | From start of the hospital to Oct 2017 | PDF file version of data for 15 In-Patients |
| Data domains | Lab  Vital signs  Diagnosis | All the available data in the hospital database | Specific patient visits case report forms |
| Type of extraction | Full extraction of available domains | Full extraction of all the available hospital data | Incremental extraction of the domains available in the PDF files |

## Clinical data understanding

This section outlines observations from structural review of the datasets: In a well-defined database for patients should have the primary key as Patient ID: mr\_no (in our case), but the underlying database considers unique visit for each patient as a primary key between tables (Patient\_ID). In general, a variable containing same information across tables should have the same name, but in our case, each table has a different variable, making it difficult to create logical links across tables. E.g. Consultation\_ID from doctor\_consultation and Patient\_ID from patient\_registration had the same information; Visit\_ID from mrd\_diagnosis and Patient\_ID from doctor\_consultation meant the same. The case report form allows for multiple diseases and multiple treatments to be recorded for each patient, this causes a “clinical logic” challenge – the potential 1-1 relation between a disease and a treatment is lost, this has to be derived outside of the database using expert understanding which would require investment of time and efforts from Ayurvedic vaidyas. There are multiple versions of the same table available in the database (as a programmer, it is well understood that older copies are retained in the system), but due to unavailability of the documentations, this increases the complexity.

The following segment outlines observations from the clinical data review of individual case report forms:

Vital sign dataset: The existing database has various vital signs parameters listed one below the other. The current structure has one record per patient per visit per parameter. For a lot of visits vital sign information is missing, or partially filled. There are certain records with implausible values for certain parameters such as height and weight having 0 value. Blood pressure values having character data.

Lab measurement dataset: Findings are similar to the Vital signs database. Along with the patient identifier information, only laboratory test name and laboratory measurements are present. In case a patient does the laboratory investigations outside of the hospital that data gets stored in a scanned image format. Apart from this the dataset does not contain the date of sample, reference ranges, laboratory parameter units, fasting status etc. A single test has multiple names. E.g. Alanine Aminotransferase is captured in the dataset in the following different ways:

Alanine Aminotransferase

Alanine Aminotransferase ALT (SGPT)(UV Kinetic)

Alanine Aminotransferase (SGPT)(UV Kinetic)

Alanine Aminotransferase ALT (SGPT)

Alanine Aminotransferase ALT (SGPT)(UV Kinetic)

S.G.PT ( UV kinetic)

SGPT ( UV kinetic)

ERYTHROCYTE SEDIMENTATION RATE is captured in the dataset in the following different ways:

ERYTHROCYTE SEDIMENTATION RATE

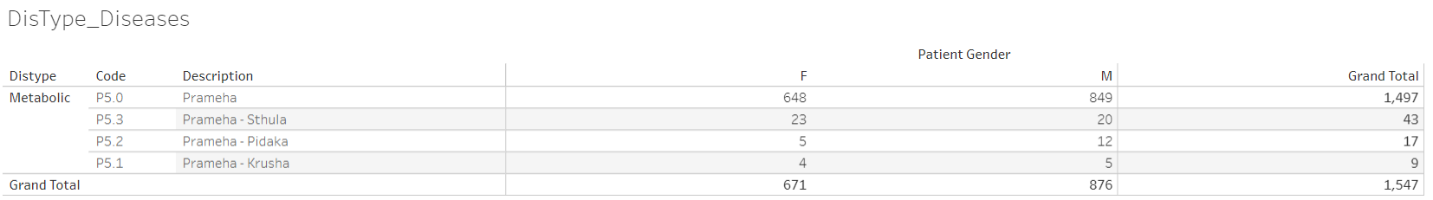
ERYTHROCYTE SEDIMENTATION RATE ( ESR)

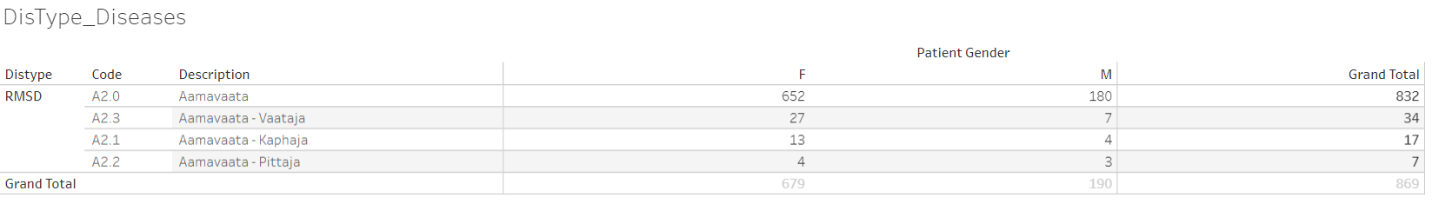
ERYTHROCYTE SEDIMENTATION RATE ( ESR)

Treatment dataset: the treatment or dosing or medication dataset does not get exported into a structured file for easy understanding and analysis. Which treatment is prescribed for which disease is not easily understandable based on the system generated report.

Medical coding and clinically important variables: the medical records for patients were captured differently by different doctors, nurses and other medical staff. Same information is found in more than 1 variable. Different acronyms are used inconsistently. Answer for more than 1 question is captured in 1 variable. Due to “free text nature” of variables simple questions like Yes / No have many different data values.

Classification and Sub-classification of the Doshas / Diseases: it is observed that the main disease classification by kapha, pitta, vata has disparity in numbers. The tables below show the variation in the counts of the diseases and their sub-classification.





Data version: 2011 to Oct 2017

<https://public.tableau.com/views/01SQL_Dis_Med_Ser/DisType_Diseases?:language=en&:display_count=y&:origin=viz_share_link>

Observations related to patient profile report generation module: The hospital database did not provide one holistic report of the patient as a whole including details from all the datasets.

The following section outlines observations from review of the derived datasets:

1. Complete disease and treatment information for each patient is available in a structured database format is generated
2. Longitudinal picture of a patient’s disease can be drawn easily
3. In-Patient and Out-Patient information is collated at one place
4. Disease and treatment information for related diseases is present at one place
5. Time between 2 visits to the hospital for a patient can be calculated to understand the treatment regimen and possible compliance
6. Easy filtering for a disease, treatment is possible
7. Complicated subsets and creation of cohorts is possible
8. The naming of the source datasets within the hospital database is quite logical but due to lack of documentation it was a puzzle to solve:
   * First dataset covers: patient\_details
   * After taking the details the patient is asked for: patient\_registration
   * Next logical step is of: doctor\_consultation
   * The treating doctor is able to diagnose the patient: mrd\_diagnosis
   * Patients are prescribed medicines: patient\_prescription, patient\_medicine\_prescriptions
   * If a patient is In-Patient then the information is stored in: ip\_prescription
   * Along with the medicines if there are services prescribed then they are documented in: services\_prescribed

## Studying demographics and patient specific factors

Related to complete set of patients:

While exploring the basic data, the following high-level picture appears: For the 5-year time frame from 2011 to 2016, the database contains ~40,000 unique patients (Fig 4.3.1), 90% of patients are from India and remaining 10% patients are from 50+ different countries (Fig 4.3.2). The proportion of male and female patient is ~50%. Median age for females is marginally higher than males across all visit types (Fig 4.3.3). Approximately 90% of patients are Out-Patients and 10% are In-Patients (Fig 4.3.1). Approximately12,000+ female patients and 14,000+ male patients have reported only a single disease (Fig 4.3.6), these patients could have come only once to the hospital and may not have come back at all after reporting the 1st disease. There are a few outliers observed having more than 10 disease conditions across the years. The maximum age of 108 years is a possible case of data issue. Similar anomalies are seen in a few other groups, e.g., patients reporting 23 diseases, is this accurate? This warrants additional data checks from operational and clinical perspective. Blood group is collected only for ~32,500 out of ~40,000 patients. There is missing data for almost 20% of patients. Blood group distribution is largely in line with the Indian blood group distribution (Fig 4.3.4).

Discussion for the metabolic and RMSD disease areas:

Out of ~40,000 patients, there are ~14,000 patients having reported at least 1 metabolic and/or 1 RMSD disease condition. It is quite evident that there are a lot more patients in the RMSD group compared to the metabolic group (Fig 4.3.7). Large number of patients are visiting the hospital only for 1 visit, ~62% patients are dropping off in first month of treatment. ~15% of patients having at least one RMSD disease are still visiting the hospital after 1 year of first ever visit to the hospital (Fig 4.3.9). Boxplot representation of age shows variability in age across disease type and gender (Fig 4.3.8). Presentation of disease burden by gender, Indian seasons (rutus) and disease category provides data about possible variations reported for different diseases (Fig 4.3.10). (1) Prameha, (2) Madhumeha, and (3) Sthaulya are the top three most frequently reported metabolic diseases where as (1) Vaatavyaadhi – Sandhigata Vaata, (2) Vaatavyaadhi, (3) Vaatavyaadhi – Gridhrasee, (4) Sthaanabhedana Shoola – Katee Shoola and (5) Sthaanabhedana Graha – Katee Graha are the top five most frequently reported RMSD diseases. Prameha and Madhumeha are reported more by males than females. There are more female patients with disease condition Sthaulya. In general, RMSD diseases are reported in more females than males. For RMSD disease group, 51 out of 97 diseases are reported in <= 10 patients. Metabolic diseases are not varying across seasons, while RMSD diseases have shown some seasonal variations. The before and after visualization of data allows to build a disease and medicinal trajectories. These should be useful for determining diagnostic and prognostic relationships.

## Studying diagnostics and interventions

# Discussions

## Converting real life data into analyzable format

## Clinical data understanding

Sections listed below offer possible technical solutions for the shortcomings identified:

Vital sign dataset: the vital signs database could have an alternative presentation of one record per patient per visit in addition to the existing presentation. The vital sign parameters can be presented as distinct columns, one each for each parameter.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Patient ID | Visit date | SBP | DBP | Pulse | Height | Weight |
| 1 | 01-Jan-2016 | xxx | xxx | Xxx | Xxx | Xxx |
| 1 | 15-Jan-2016 | xxx | xxx | Xxx | Xxx | Xxx |
| 1 | 31-Jan-2016 | xxx | xxx | Xxx | Xxx | Xxx |

This type of presentation would make the length of data smaller. Trends for the same patient over a period of time could be assessed faster. The parameter result values should be presented in numeric form, rather than character format. This will allow the data to be used for numeric calculations. In case of age and/or gender specific analysis; normal ranges can be applied in the database and these calculations could be done in the backend without affecting the end users, here doctors, nurses to name a few.

Lab dataset: When a lab test is done from other pathology the data from scanned reports is not translated into hospital dataset, this results in missing information in the database, as it is not retrievable for any analysis. A single test has multiple names making it very difficult to create summaries on laboratory parameters. Would it be possible to get the pathology lab data in electronic dataset format? A few suggestions on updating the existing version:

* A standard naming convention of lab tests should be created
* A standard units look up table should be built for possible conversion from one unit to another
* Lab results values should be saved as numeric and character (when some test results come out as ordinal scale measurements) variables
* In case of cancer patients, should the National Cancer Institute, USA, proposed NCI CTC grading variable be created for specific parameters? – this could be created in the backend database, more useful for analysis

Treatment dataset: This is one of the most important parts of data necessary for any type of analysis. A complex SQL data extraction code along with numerous merges and complex Cartesian products via the R-code had to be performed to arrive at easily readable dataset. E.g. MRD\_Diagnosis, Patient\_Prescription, Patient\_Medicine\_Prescriptions, IP\_Prescription datasets had to be merged to get a complete interventional view. Treatment database could be structured in a way that components can be collected and reported in a systematic manner:

* Name of the treatment(s)
* Treatment(s) prescribed for which disease
* Treatment(s) start date
* Treatment(s) end date
* Names of medications
* Type of medications (classical formulation, proprietary, etc.)
* Dosing information
* Route of administration (Treatment procedure, oral treatment, panchakarma, etc.)
* Dose increase or decrease

Due to the complex nature of the data, the structure would be one record per patient, per visit, per disease, per treatment assigned. E.g. if a patient has 2 disease conditions and 4 treatments are assigned then for that particular visit, there should be 8 records present in the database. There are numerous medicines prescribed. These medicines are classified into following broad categories:

1. Abhyanga
2. Aristham
3. Arka
4. Asavam
5. Avagha
6. Bhasma / Bhasma Cap / Bhasma Tab
7. Dhara
8. Ghritam and variations of Ghritam
9. Kashayam and variations
10. Kshar
11. Lehyam
12. Oil
13. Pichu
14. Rasayanam
15. Any additional classification which makes sense

Source variable for treatment will be classified into the following categories. This should allow recreation of treatment protocol as per Ayurvedic principles. Various Ayurvedic texts have defined standard treatment protocols for different ailments. Based on the table below extracted **from xyz paper**, can we propose a sequence of treatment for various conditions and build it in our database so that an automatic re-creation of the classical treatment text can be done? This will not only serve as a support to the practicing vaidyas but also serve as a validation tool for the given treatment regime.



Medical coding: Medical coding is a robust method to simplify the variation in the data by uniformly categorizing the medical terms appropriately. This step allows us to maintain high quality database. Coded medical data is a standardized form of data, globally approved, and can aid in future machine learning and automation. The most commonly used medical coding dictionaries for coding medical terms are MedDRA and WHO DDE [68] [69]. The other standardized medical dictionaries are:

* COSTART - Coding Symbols for Thesaurus of Adverse Reaction Terms
* ICD xx CM - International Classification of Diseases xx Revision Clinical Modification [70]
* MedDRA – Medical Dictionary for Regulatory Activities [68]
* WHO-ART – World Health Organization Adverse Reactions Terminology
* WHO-DDE– World Health Organization Drug Dictionary Enhanced
* ACD – Ayurvedic Classification of Diseases
* NCI – National Cancer Institute Code list [71]
* LOINC – Logical Observation Identifiers Names and Codes standards [72]
* Any other dictionaries, as recommended by AYUSH

Classification and Sub-classification of the Doshas / Diseases:Every scientist knows that research results are only as good as the data upon which the conclusions are formed. However, most scientists receive no training in methods for achieving, assessing, or controlling the quality of research data— topics central to clinical research informatics. An exhaustive list of all possible diseases should be created and a bibliography or checklist of disease classification and sub-classification should be maintained so that the physicians basis their judgment can classify the dosha appropriately and use the recommended disease classification term so that this inconsistency and disparity in reporting can be avoided. The problem should be split into operational and scientific component. Identify fields which would require coding: Diagnosis codes description, Compliant, Drug description. Operational steps to be taken as follows:

1. The existing data should be codified in a retrospective manner
2. Business guidance document should be prepared on how to work in future
3. Number of days should be predefined to have data coded from the time of patient visit
4. An automatic tracking mechanism should be defined to keep track of the status:

|  |  |  |
| --- | --- | --- |
| Number of days from patient visit to coded data | Number of records coded | Number of records yet to be coded |
| < 7 days | XX | XX |
| 7 days to 14 days | XX | XX |
| 14 days to 28 days | XX | XX |
| > 28 days | XX | XX |

Temporary staff should be allocated to complete the backlog.

Scientific questions regarding Ayurvedic medical terminology should be answered by doctors at hospital. A teamof3-4 doctors for a period of 4 months or as appropriate, contributing 20% of their time, would help complete the categorization process as described above or post graduate students, under the guidance of a senior vaidya can take on this responsibility.

Overall data standardization:Overall the hospital data is not captured in a standardized format as explained in the above points. Create standardized CRF pages for consistent and correct data capture. Already established standards like Clinical Data Interchange Standards Consortium (CDISC) or International Organization for Standardization (ISO) standards can be implemented [73] [74] Appropriate drop down menu lists with predefined inputs to be built into the system, with the help of experts, to ensure good quality data.

A patient profile report is a consolidation of all the data for a patient available in the database. This consolidated view at a patient level provides an easy access to the patient history. If this type of a report is electronically available for a patient then a patient’s case can be handled by any doctor available. The contents of a patient profile are outlined below:

* All the demographic characteristics of a patient: age, sex, race, religion, place of residence, etc.
* All the useful data for operational ease: policy number, health coverage status, in-patient, Out-patient, etc.
* Visit information: number of visits to the hospital, corresponding dates and day of visit. The day should be calculated based on the first visit date (visit date – first visit date + 1). This value must never be missing and must be positive.
* Background disease history: Is the patient disease history getting captured at first visit of each patient? Would it be useful to not down the background history in a systematic manner?
* Vital sign measurements: a tabular view of the collected vital sign measurements.
* Data collected for the diseases and diagnosis: details about the clinically relevant fields should be discussed. Some standard fields –
  + Complains as reported and coded either in ICD 10, or ACD, or Meddra dictionary (operational possibility to be checked)
  + Duration of disease or start date, end date
  + Data collected for Ayurvedic examination: variables outlined in *dash vidh pariksha* (if these variables are not captured currently then how to make provisions for the same?)
* Treatments administered:
  + Treatment start date
  + Treatment end date
  + Names of medications
  + Type of medications (classical formulation, proprietary, etc.)
  + Dosing information
  + Route of administration
* Details of lab results
* Outcomes
* Patient still ongoing or discontinued (need to create an algorithm to define this status)

The pictorial representation below, summarizes the cycle of understanding the hospital data so that a meaningful interpretation can be arrived at. This picture provides an overview of looking at the same data in different ways and how each section provides an additional view to the patient data thus converting the collected information into a meaningful story.

Improvements to the system architecture: The team sets up rules and guidelines for the implementation. Yet once the system is live, due to the lack of consistency in data entry methodology things begin to fall apart. User inputs the same data in different ways. New staff comes on board and has their own way of entering information. Inconsistent data creates inaccurate reports. Hence robust documentation and streamlined training and onboarding of the new data entry operators is a must. Building and implementing a design policy is the first step towards reinforcing the build rules. It provides documentation for EMR analysts to follow which will reduce inconsistencies and improve the EMR functionality. Below are some of the aspects which if kept in check can avoid inconsistency in data

* Capitalization: Monitor the use of caps. Document and suggest exactly when to use caps and otherwise. In case all caps are used, ensure appropriate warning message or suggestion is generated to alert the operator.
* Abbreviations: Prepare a bibliography of all allowed abbreviations and their meanings. If possible, create inbuilt check lists in the building forms to avoid incorrect abbreviations.
* Workflows: Periodically evaluate and monitor the workflows to check the effectiveness. During a system change a workflow audit is extremely important, since non-working workflows undermines’ the system functionality as the user may create smart workarounds skipping the important steps. Along with Workflow audits it is a best practice to have planned system audits.
* Naming conventions: This is the most important step. Following the appropriate naming conventions helps save time, money and future efforts. It is easier to onboard new employees and also eases future searches or any kind of analysis.
* Data Quality check plan: It is best to create data validation programs during data base setup which can be run periodically to check for the correctness of data entry.
* Well defined database maintenance plan: It is important to have a well-planned and periodically scheduled database maintenance program.
* Train regularly: Regular training and refresher programs is a must to ensure that the end users are up to the mark with the processes and systems so that a healthy database can be maintained.

If any organization is able to follow these proposed solutions, then possible outcomes could be as follows:

* The data will be closer to analysis ready format
* The database will be useful in publishing case studies, case series, etc. in very short period of time. This will help us gain more visibility in scientific world
* More empirical data will be available at our disposal
* The database could become a model database for other Ayurvedic institutions to follow

An initiation of new role for data related updates: For this type of work, it is essential to have someone who uses IT systems, uses a statistician’s mind, provides insights into building database, has scientific rigor, and uses multiple technologies and has ability to connect dots to solve this big puzzle. Thereby, converting an Observation to Summaries to Stories and Clinical insights without necessarily being a doctor. This expertise could be labelled as “statistical programmer”, “clinical analyst” and “clinical programmer” or now in our case “Ayur analyst”.

## Studying demographics and patient specific factors

A pictorial representation of data on world map is a convenient way to summarize large amounts of data. This form of data representation will help any public health official. If the individual state and city information is available, then additional drill down illustration is also possible – this supplementary graphic will allow us to identify the distribution of patients and diseases from different parts of India. More details related to diseases, treatments, additional demographic characteristics could be added to the visual analysis to efficiently recover key information as and when needed. This can form the basis of public health policies framed either by government or by private companies. The In-Patient and Out-Patient distribution suggests that the route of administration is simple and easily understood by the patients and the caregivers. The diseases may not be life threatening or fatal (Fig 4.3.3, Fig 4.3.5). Are these patients largely coming in for “2nd opinion”? Or if this data is to be looked at positively, are they getting benefitted and hence are not coming back for consultation beyond the first reported disease? On the other hand, a few patients could be having a lot of faith in Ayurvedic treatment, for them to continue with treatment, they could have found the underlying treatment effective. Blood group distribution for many patients is a great source of knowledge. Even though this does not help in day-to-day treatment options, there is undoubted epidemiological value in this presentation. There are obvious mistakes in documenting the blood groups observed via this tabulation – another secondary use of this tabulation is to build data quality related efficiencies. While finding data inconsistencies was not a primary objective of this analysis, there is this secondary usage available to the scientific community.

The empirical evidence generated by such fundamental data will be very useful for the hospital management, public health officials, treating physicians. This kind of tabulation plays a key role in evidence generation and synthesis. Is there a similar analysis available for another Ayurvedic hospital, or any other private or public hospital in public domain? This can be used to understand the use and misuse of the limited medical sources across the geographies.

Lesser duration of patient and hospital association may mean either the patients are benefitted by the treatment or are not happy and hence discontinue the treatment. Longer duration of association may mean that the patient is receiving benefit and hence is coming for regular follow-ups for the same condition, or the disease condition could be chronic in nature. These analyses provide a useful macro level representation of data for public health policies for these non-communicable diseases. Data driven approach of optimally utilizing resources suggest strengthening the RMSD disease treating facilities from pharmacy to Vaidyas to patient. The low rate of reporting of some of the diseases may explain the natural variations or may reveal inconsistent labelling of the diseases. Boxplot representation of age provides the distribution of diseases across age and grouped by gender. It also gives a comparative view of multiple diseases thus providing an information on the disease prevalence in the age category as well as gender.

## Studying diagnostics and interventions

# Conclusion

The introduction of this thesis outlines the need to undertake the study, by providing historic perspectives on medicine, pharmacy, development of hospitals throughout the world, internet era, and the Indian context relating to modern medicine as well as Ayurveda. The thoughts from the thought leaders in the field of Ayurveda are profound and they call for the modern methods, new approaches, innovative strategies to be attempted to take the science forward. There is a reflection on the type of evidence generated through different controlled experiments (RCTs) and life experiments (Observational, Experiential) – some evidence about how these multiple approaches could yield similar results. This chapter askes a few questions from diverse points of views and seeks answers – some of these answers are hidden in the everyday Ayurvedic clinical practice – which is still largely untapped, prompting the following: how data analysis of electronically captured data help in advancing understanding?

Next section of this thesis elaborated on the technical details of a database. We saw a few technical details about the hospital database: how may source tables, how are they stored in ~200+ tables, out of which ~20 to 25 how tables are used to generate datasets useful for the analysis. Subsequently we saw a few flowcharts outlining ~50+ steps to go from Live source –Staging data – transformed data - ~30+ source variables + ~30+ derived variables in 01adsl\_met\_rmsd dataset: patient level data covering Treatment and disease information. This forms a solid basis of all possible operational and clinical analysis going forward.

Clinical data understanding showed how individual observations can be transformed into meaningful patient narratives. This section explained how the usage of operational and clinical part of the data can benefit varied stake holders, emphasized the need to convert “a thought from a doctor’s mind” into “actionable and consistent data point” in the database for future use. It laid down the foundation for “understanding demographics and patient characteristics”.

The demographic and patient characteristic analysis provided good insights into varied components of the data which can feed into public health domain. It can be easily observed that the health and healthcare requirements of a population can be construed through the magnitude and the variability in the data. Basic science can use the relations developed through the above analysis for disease and generation of disease categorizations. Public health domain can be benefitted by disease surveillance and population health. This analysis showed how individual observations can be transformed into meaningful stories at hospital. It also provided actionable inputs to hospital management, practicing doctors and for research publications.

Day-to-day transactions at hospitals and clinics involve people from many backgrounds like, hospital administration, patients, doctors, nurses, pharmacists, pathologists, representative from insurance companies, lawyers, etc. These interactions generate a lot of information and are the primary data generators. Same set of people and a few additional professionals are the end users of the data e.g., scientists, statisticians, database developers, etc. This study has provided preliminary insights into various aspects of data generated during real time consultation at I-AIM:

* What kind of data has been collected so far for each patient visit
* What part of the data is related to the patient background characteristics, disease conditions, prescribed medicines
* What kinds of diseases are getting treated more frequently and treatments are prescribed for what kinds of diseases
* What are the strengths of the collected data and what are the areas of improvement going forward

It was understood that there is a need to have a profession of a “Statistical programmer” or a “clinical programmer” or an “Ayurdata expert” who can play a pivotal role in data journey from “Start to end” – “data collection – curation – analysis and reporting”.

This role should know Information technology requirements, data management techniques for generating quality data, in addition to knowing basic and advanced statistical concepts. The computational advances in the world of computer science could be leveraged via appropriate softwares. Using multiple technologies theoretical ideas can be converted into practical interactive visualizations and interactive analyses. These will help convert individual data observations into summaries then into stories thus enabling knowledge generation. We believe that this is a pioneering effort within ayurvedic data analysis area.

Ayurdata expert can contribute to the (1) clinical operations by increasing the operation efficiencies. (2) Research and development areas by generating reproducible research, automated documentation and audit trail creation, lastly by helping out in clinical decision support by providing Real world data, (3) Evidence based medicine area by creating empirical evidence, nature and context of diseases and, analysis of clinical practice data, (4) Public health area by contributing to Patterns of diseases and medicines, Disease surveillance, Data to actionable information. Ayurdata analyst can create new needs which are not considered as needs so far within Ayurveda. Complex computing and advanced visuals can be made available.

While studying both the structure and the content of the hospital database, it was observed, that standardization of database along with effective curation of the data could provide a gold mine of data which when summarized could lead us to a lot of supportive evidence.

Below is a summary of our observations and outcomes based on the retrospective view of the available hospital data

Proposals:

1. Building a standard data collection system / repository

* Develop standard data collection methodology
* Build electronic Case Report Forms for Ayurvedic patient data or use proposals from AYUSH ministry
* Adapt to standards of data entry
* Build a data-mart or data repository for all the Ayurvedic run trials for larger benefit of summarizing large data for effective evidence building

Possible outcomes:

1. Creation of visual dashboards
   1. To understand the most prevalent disease and clinical endpoints
   2. To regulate and monitor patient flow
2. Use the endpoints from the database to
   1. Identify cohorts of similar patients
   2. Build supportive clinical evidence and clinical effectiveness
   3. Create new hypothesis from new clinical trials
3. The Comorbidity outcomes can aid in
   1. Identifying the patterns and next possible disease along with the intervention
   2. Plan for seasonal variation
   3. Creation of patient supportive program for better disease management
4. The Pharmaco-Epi-Real World Demographics
   1. Support the policy makers, insurance companies, government

A variety of analysis and summarization of the hospital data was conducted with a view to derive meaningful outcomes which confirm the Ayurvedic principles. The picture [6.2.1](#_Real_World_Data) explains the different contexts that play a vital role in how the RWD shapes up and supports in providing key outcomes which in near future will aid in regulatory approvals, policy makers and the government and common man - the tax payer.

This thesis outlines many tools which can be used by various stakeholders. They are free and easy to use. They allow multi-dimensional display of data in a very short amount of space. The tools can create evidence for multiple stakeholders. Free softwares like, R, python, Java, tableau and many more have made it possible to harness the power of data in many ways and examples above show the same.

This is not an end but just a beginning of Ayurdata experts ...

# Appendix

## Approval from the hospital management to carry out the retrospective study

## 11.1 ANNEXURE 1 –NOC FROM IAIM MEDICAL DIRECTOR.

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

IAIM/2020/NOC/01 Date: 29.05.2020

LETTER OF PERMISSION AND NO OBJECTION CERTIFICATE

TO WHOM IT MAY CONCERN

This is to grant permission to Mr. Vinay Mahajan to conduct the research study “Review of hospital based Ayurvedic Electronic Health Records to gain real world knowledge - a retrospective data analysis” using I-AIM anonymized Electronic Health Records as per the protocol approved by the IEC.

I am assured that Mr. Vinay Mahajan will maintain confidentiality of the data.

Further, it is also agreed that any presentation and publication of the results arising from the study will be done after due permission from the authorities of IAIM and TDU.

Dr. Prasan Shankar

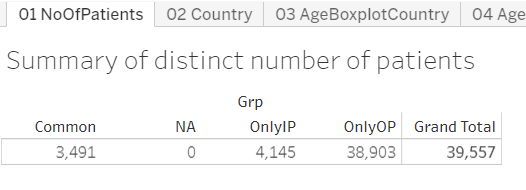
Medical director

IAIM

Bangalore

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

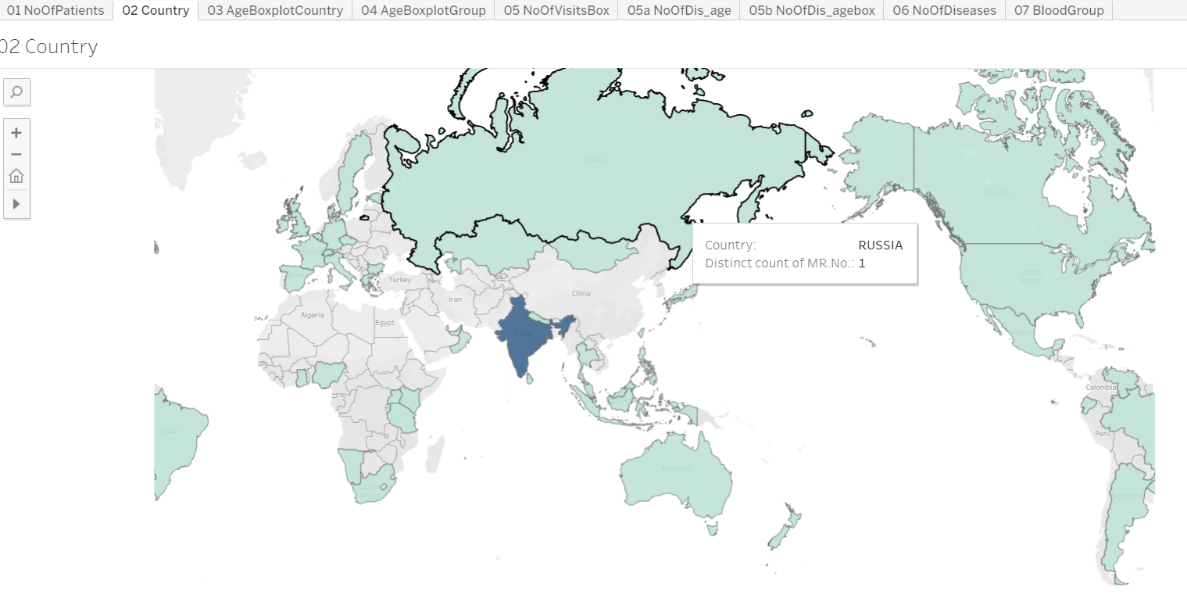
# Total Number of Patients



Data version: 2011 to Oct 2016

<https://public.tableau.com/views/04_patient_analysis_tablaeu/01NoOfPatients?:language=en&:display_count=y&:origin=viz_share_link>

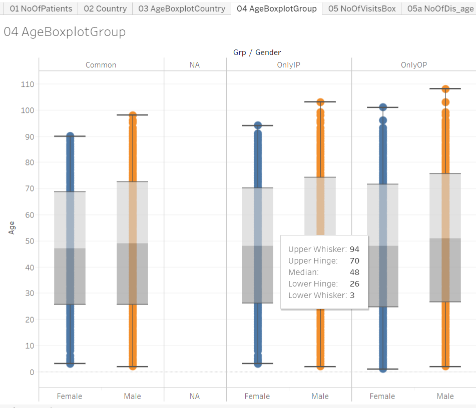
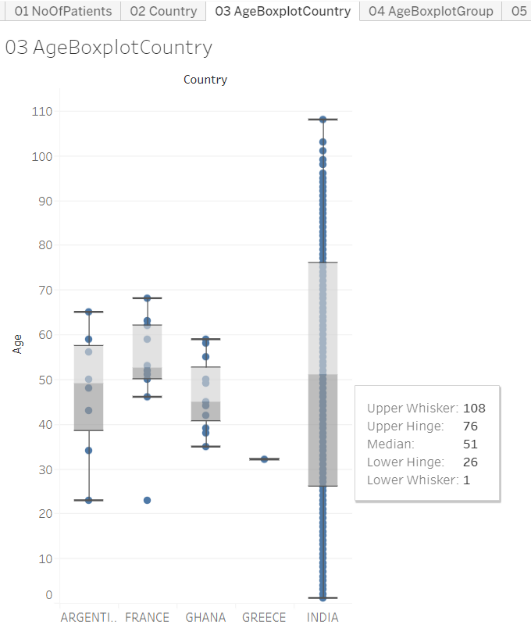
# Country-wise Visualization



Data version: 2011 to Oct 2016

<https://public.tableau.com/views/04_patient_analysis_tablaeu/02Country?:language=en&:display_count=y&:origin=viz_share_link>

# Age and Gender Distribution

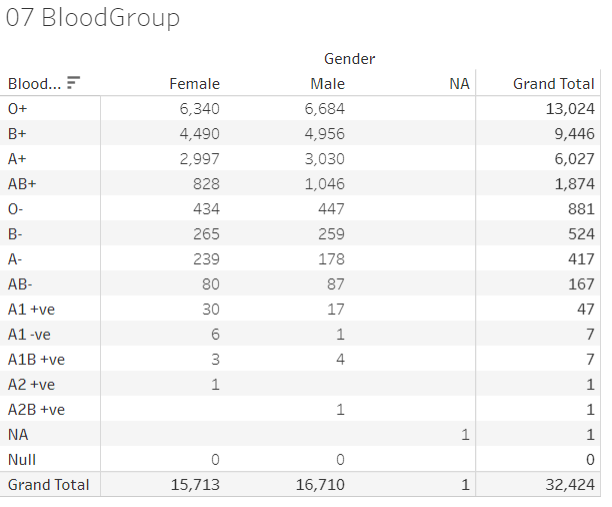


Data version: 2011 to Oct 2016

<https://public.tableau.com/views/04_patient_analysis_tablaeu/03AgeBoxplotCountry?:language=en&:display_count=y&:origin=viz_share_link>

<https://public.tableau.com/views/04_patient_analysis_tablaeu/04AgeBoxplotGroup?:language=en&:display_count=y&:origin=viz_share_link>

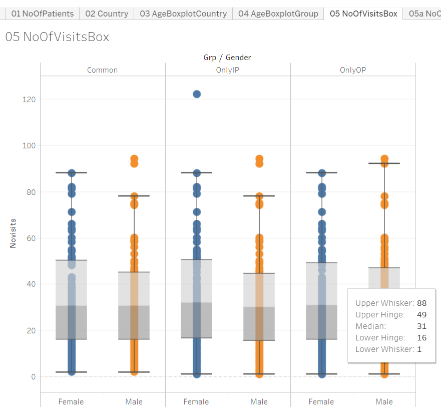
# Blood-group Distribution



Data version: 2011 to Oct 2016

<https://public.tableau.com/views/04_patient_analysis_tablaeu/07BloodGroup_1?:language=en&:display_count=y&:origin=viz_share_link>

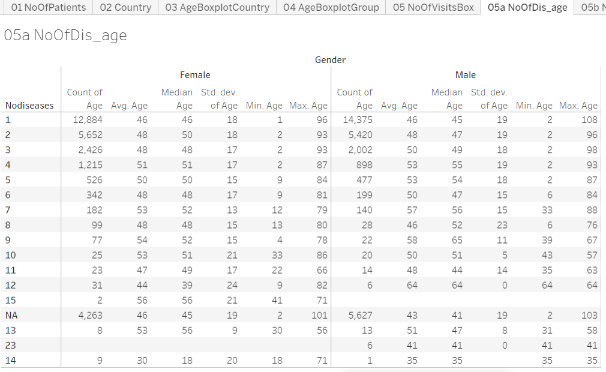
# Number of Visits, and Visit Types



Data version: 2011 to Oct 2016

<https://public.tableau.com/views/04_patient_analysis_tablaeu/05NoOfVisitsBox_1?:language=en&:display_count=y&:origin=viz_share_link>

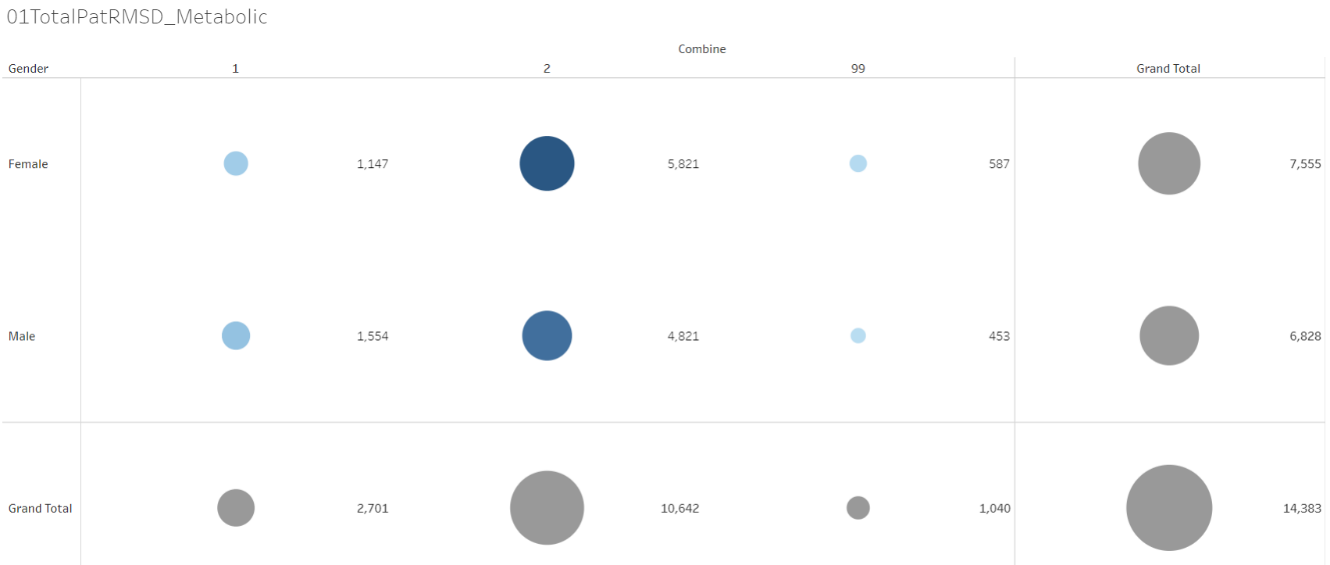
# Number of Diseases by Age and Gender



Data version: 2011 to Oct 2016

<https://public.tableau.com/views/04_patient_analysis_tablaeu/05aNoOfDis_age?:language=en&:display_count=y&:origin=viz_share_link>

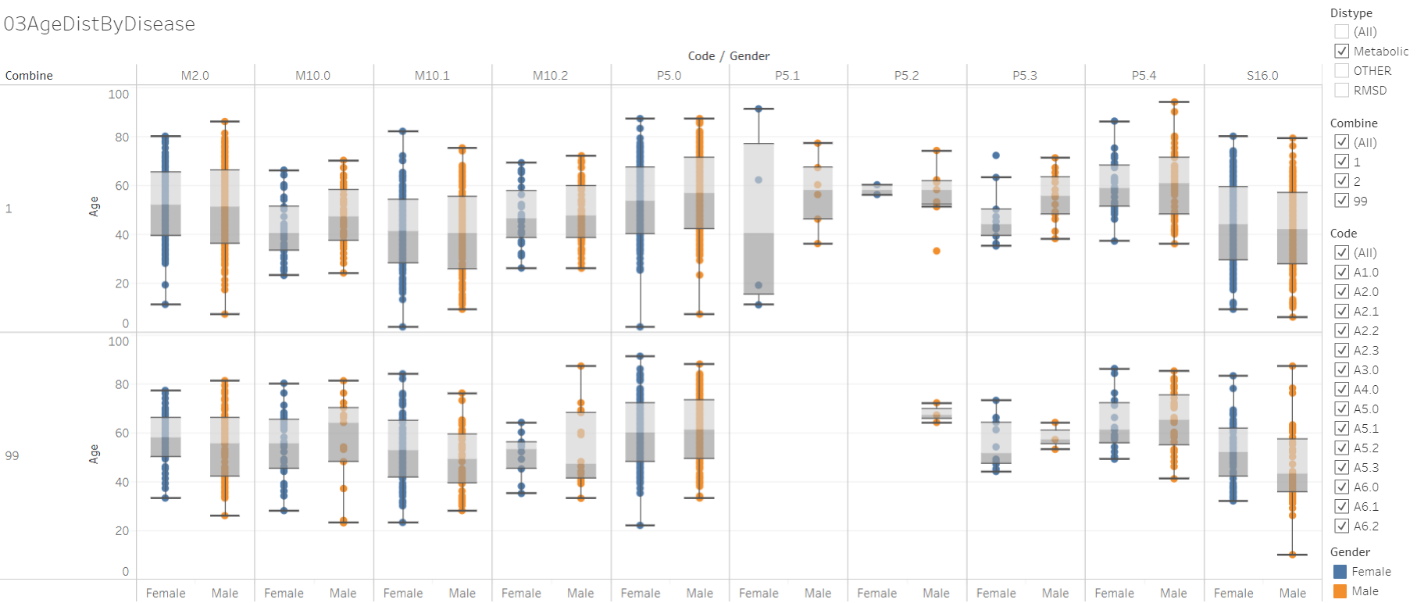
# Data tabulation for patients reporting RMSD and Metabolic diseases



Data version: 2011 to Oct 2016

<https://public.tableau.com/views/01RMSD_MET/01TotalPatRMSD_Metabolic?:display_count=y&:origin=viz_share_link>

# Disease distribution by Age and gender



Data version: 2011 to Oct 2016

<https://public.tableau.com/views/01RMSD_MET/03AgeDistByDisease?:language=en&:display_count=y&:origin=viz_share_link>

The columns in the above image are different diseases from Metabolic and RMSD categories. (Refer <https://public.tableau.com/views/00codelist/ListOfDiseases?:language=en&:display_count=y&publish=yes&:origin=viz_share_link> for codes and de-codes)

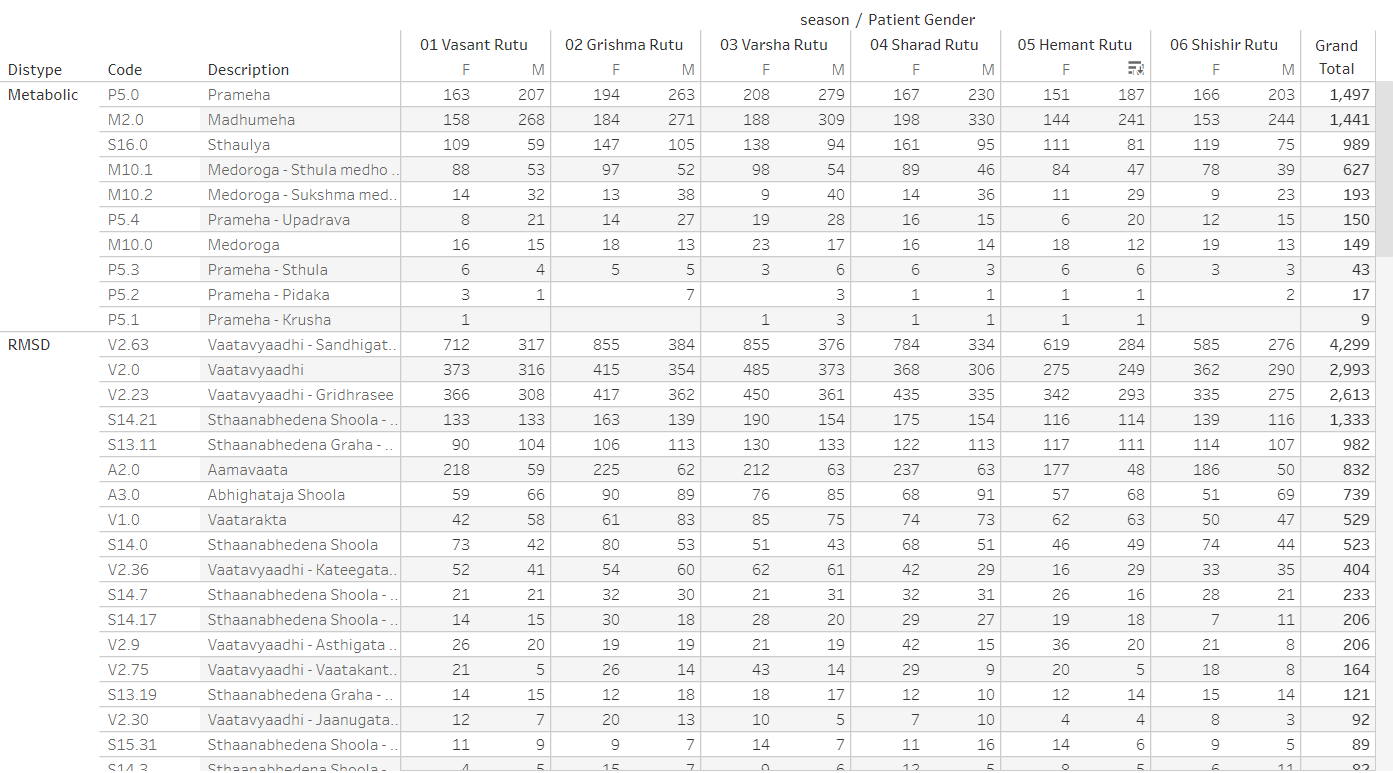
# Patient visit duration for Disease categories by Gender



Data version: 2011 to Oct 2016

<https://public.tableau.com/views/01RMSD_MET/08CumDisplayByDuration?:language=en&:display_count=y&:origin=viz_share_link>

# Seasonal Variations

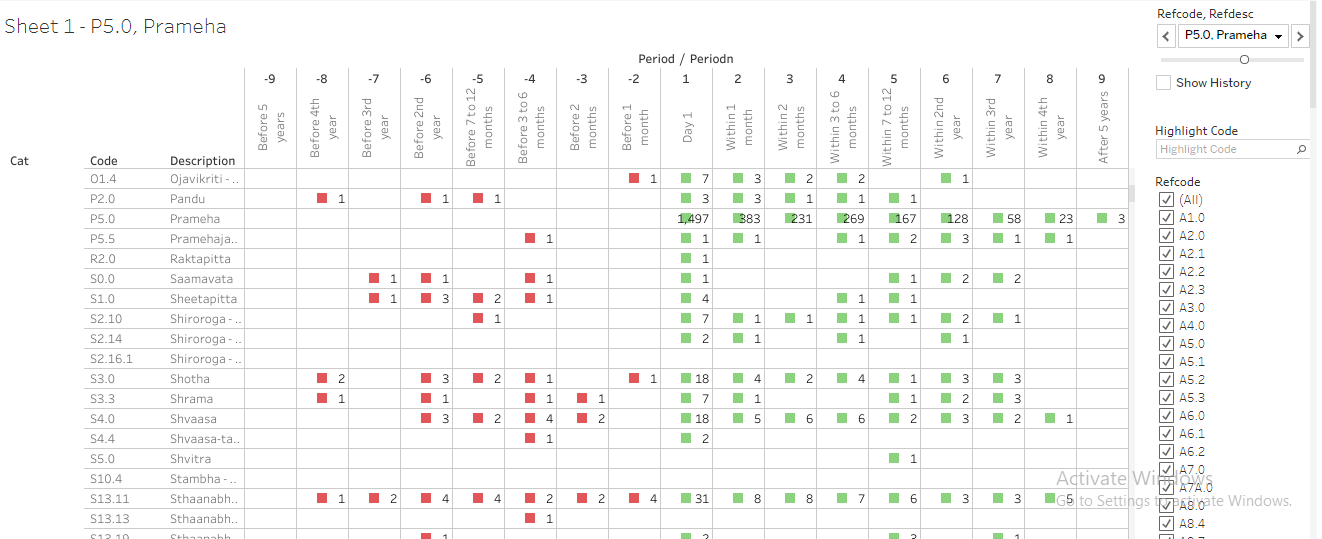


Data version: 2011 to Oct 2017

<https://public.tableau.com/views/01SQL_Dis_Med_Ser/MedicineByDay?:display_count=y&:origin=viz_share_link>

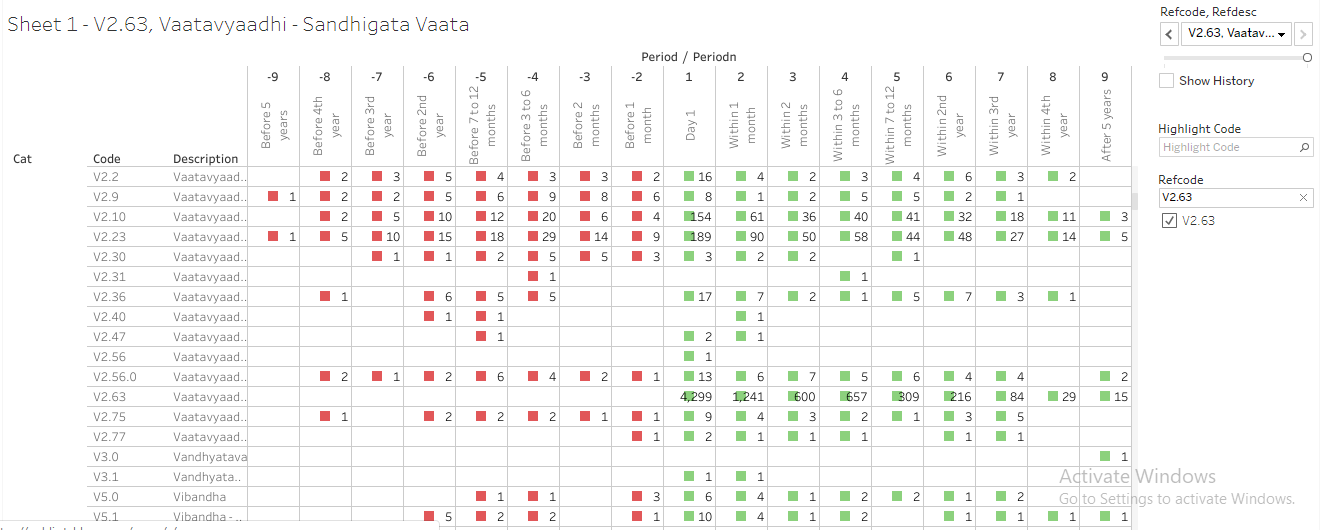
# Pre and Post Disease Classification Analysis

Example 1: Prameha



Data version: 2011 to Oct 2017

Example 2 Vaatavyadhi – Sandhigata Vaata



Data version: 2011 to Oct 2017

<https://public.tableau.com/views/085_dis_1st_time_refCal_NodesEdges/Sheet1?:display_count=y&:origin=viz_share_link>

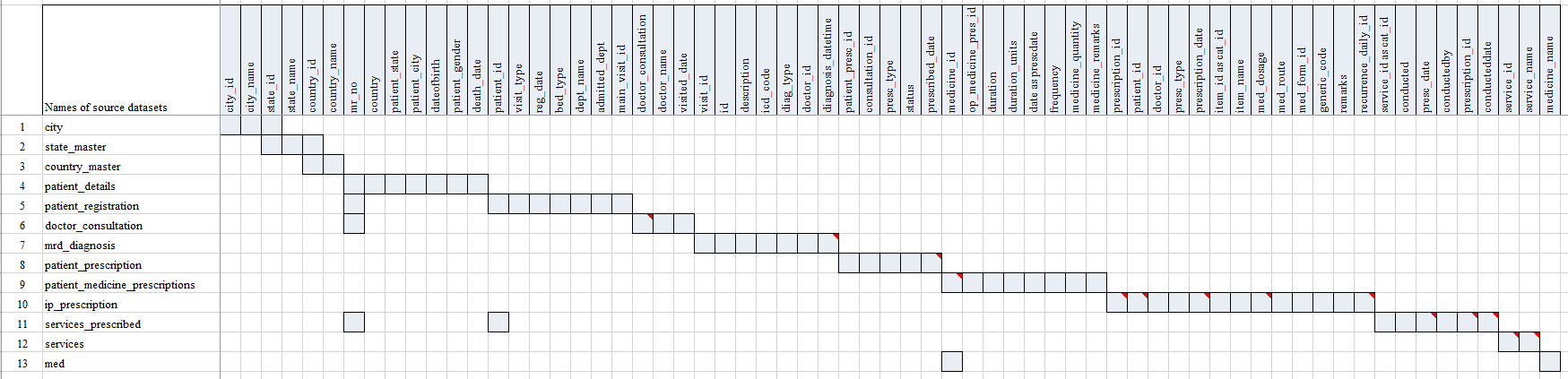
<https://public.tableau.com/views/085_dis_1st_time_refCal_NodesEdges/Sheet1?:display_count=y&:origin=viz_share_link>

Flow diagram from data source to final usage by various usage types

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Data Sources | Staging area | Ware house | Data marts | Usage |
| Data access:  Monitor Screen Check Mark Symbol Padlock Data Access Icon — Stock ... | Data Transfer Icon - StructuredVector black filter data icon set. ... | Stock vector | ColourboxThe Dirty on Data Cleansing & Appending |  | Chapter 4 Multiple Imputation | Book_MI.utf8.md |  |
| Operational system    Coding dictionaries    Clinical system    Flat files information | Calculations and transformations | Data Warehouse Icons - Download Free Vector Icons | Noun Project  Curated and consistent data storage | Operational data    Pharmacy data    Patient level data | Hospital management  Analytics - Free people icons  Researchers    Health authorities    Data mining  Modern Outline Style Data Analytics Icons Collection Stock ...  Various documents |
|  | Questions from Alteryx Training | InterWorks | Local outlier factor - Wikipedia |  |  |

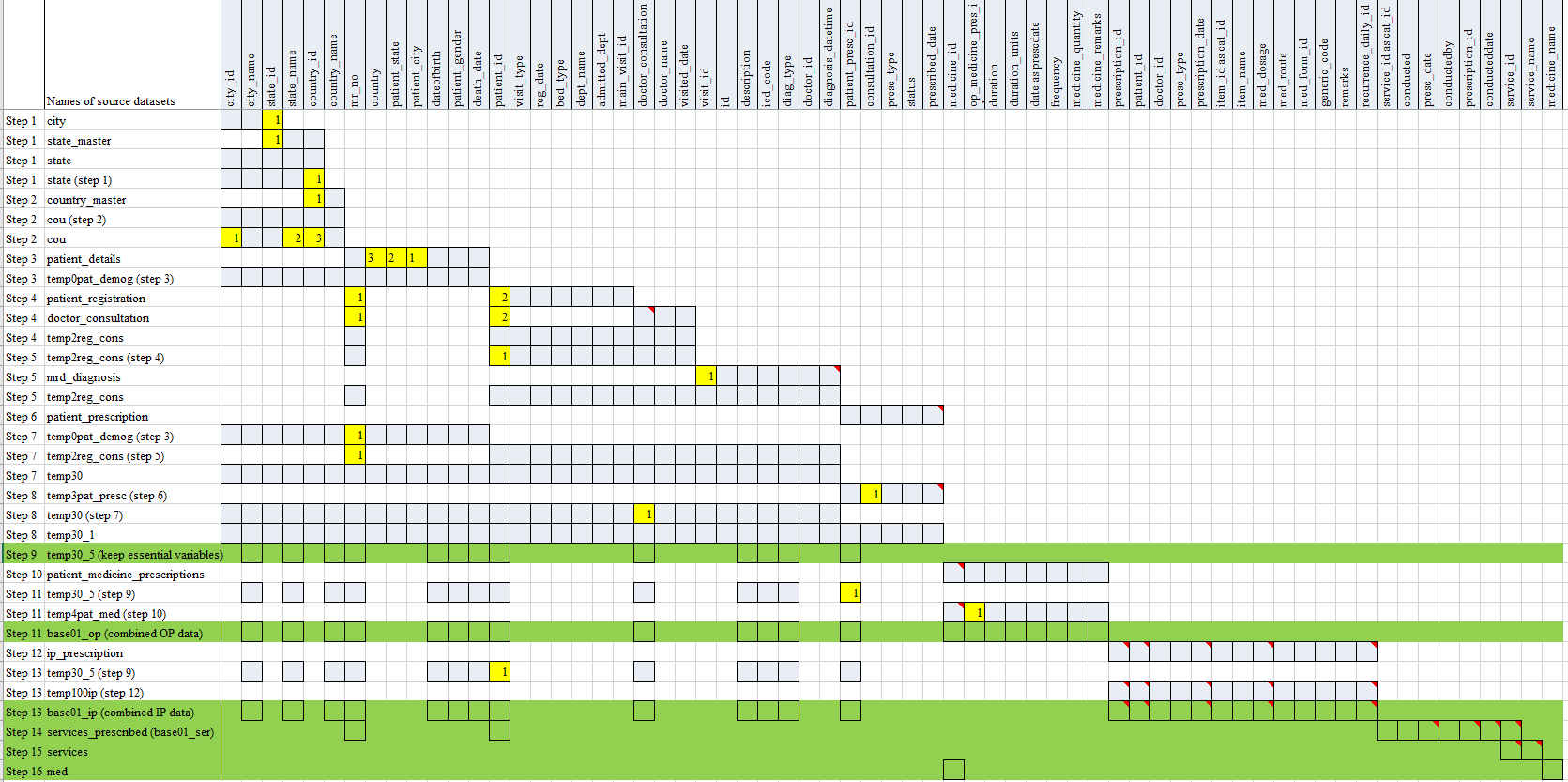
|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| action\_rights | diet\_prescribed | hospital\_technical | package\_componentdetail | patient\_registration | section\_field\_options | store\_item\_batch\_details | test\_details |
| admission | discharge\_format\_detail | 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\_charge\_adjustment | dyna\_package\_category\_limits | mrd\_casefile\_attributes | patient\_details\_patient\_phone\_country\_code | preauth\_prescription | stock\_issue\_details | store\_reorder\_levels | url\_action\_rights |
| 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 |  |

Extraction from Source database (13 datasets, ~65 variables):



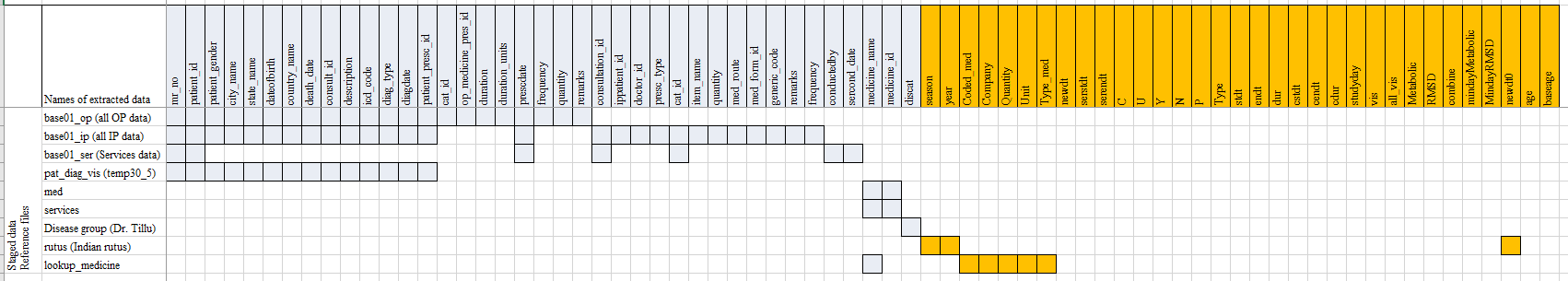
Each row in the above picture is one source dataset. Each column represents a variable. The gray coloured cell denotes the presence of the particular variable in the dataset.

# Staged data converted into 6 datasets



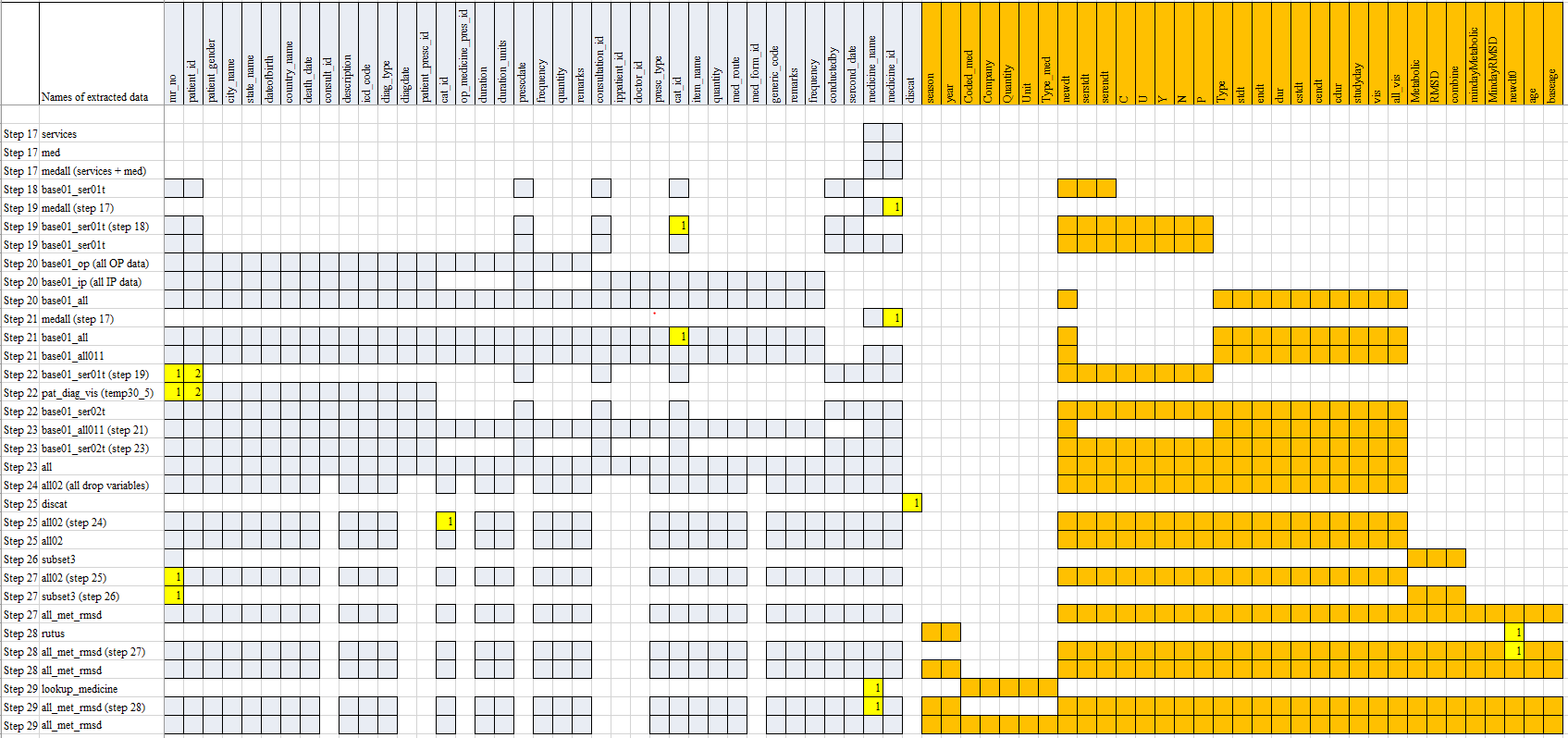
The source datasets have been merged step by step using the variables marked in yellow colour. The above picture shows 16 steps taken to generate 6 datasets, marked in Green for subsequent processing.

# Staged data (6 datasets, ~40 variables) + 3 reference files created based on inputs from experts



This picture lists 6 datasets created by earlier processing + 3 reference files provided by the experts. Using the source variables, additional variables marked in Orange are created.

# 1 Final dataset with ~30 source variables and ~30 new derived variables



The above picture 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.

# Data flow

|  |  |  |  |
| --- | --- | --- | --- |
| Source data (SQL data file) | Staging data (csv files / R data files) | Data ware house (R data files) | Usage |
| city |  | Longitudinal Patient data with disease, medication and Ayurvedic services information  ~30 variables from source  ~ 30 variables derived  ~50,000 patients  ~17,000+ patients: subsetted version for RMSD and Metabolic | Creation of additional analysis datasets |
| state\_master |  |
| country\_master | base01\_op (all OP data) |
| patient\_details | base01\_ip (all IP data) | Actual analysis  Analytics - Free people icons |
| patient\_registration | base\_01\_ser (Services data) |
| doctor\_consultation | pat\_diag\_vis (temp30\_5) |
| mrd\_diagnosis | med | Learning from the existing database to be given back as learning |
| patient\_prescription | services |
| patient\_medicine\_prescriptions |  |
| ip\_prescription | Reference files for derivations and filtering of data | Clinical communication |
| services\_prescribed | (Disease group txt file) |
| services | (Indian rutus txt file) |
| med | (Medicine type txt file) |

# Details of the Reference Dataset “01adsl\_met\_rmsd”

|  |  |  |
| --- | --- | --- |
| Variable name | Description | Derivation |
| mr\_no | Unique Patient ID | Source variable, no derivation needed  E.g. MR000001, MR040237, etc. |
| patient\_gender | Patient gender | Source variable, no derivation needed  E.g. M, F |
| patient\_id | Visit ID | Source variable, no derivation needed, the hospital database captures unique visit ID for each visit. |
| city\_name | City name | Source variable, no derivation needed |
| state\_name | State name | Source variable, no derivation needed |
| country\_name | Country name | Source variable, no derivation needed |
| dateofbirth | Date of birth | Source variable, no derivation needed, for some patients this is missing |
| newdt0 | Date of visit to hospital | Date of visit to hospital in numeric format  All the In-Patient visits, Out-Patient visits and Service related visits are combined from source datasets into a dataset, unique visit and date combinations are created. |
| newdt | Date of visit to hospital | Character version of newdt0 |
| vis | Visit | 1. Based on all the In-Patient visits, Out-Patient visits and Service related visits unique visit numbers are created. 2. Visit numbers are numeric values from 1 to n, based on current version of data; a patient has maximum number 323 visits. |
| all\_vis | All visits | This variable contains maximum number of visit for each patient. all\_vis = max(vis) grouped by each mr\_no |
| all\_ip | All IP visits | This variable contains maximum number of visits for each patient for IP type of visits. all\_vis = max(vis) grouped by each mr\_no and visit type is IP. |
| all\_op | All OP visits | This variable contains maximum number of visits for each patient for OP type of visits. all\_vis = max(vis) grouped by each mr\_no and visit type is OP. |
| studyday | Study day | studyday = 1 when the visit minimum visit or first visit for a patient, else studyday is calculated as newdt0 – min(newdt0) + 1.  Studyday is never missing and never less than 0 for the dataset created. |
| age | Age of patient at that visit | If date of birth is non-missing for a patient, then age is calculated as round( (anydate(newdt) - anydate(dateofbirth) + 1)/365.25, digits = 0 ) |
| baseage | Age of patient at the first visit | Age at vis = 1 for each patient is stored as base age |
| death\_date | Date of death | Source variable, no derivation needed |
| cstdt | Min Start date | cstdt = min(newdt) |
| cendt | End date | cendt = max(newdt) |
| cdur | Total duration in days | cdur = max(newdt) - min(newdt) + 1 |
| stdt\_IP | Start date of IP visits | Minimum visit date for IP visits for each patient |
| endt\_IP | End date of IP visits | Maximum visit date for IP visits for each patient |
| dur\_IP | Duration of IP visits | dur\_IP = endt\_IP – stdt\_IP + 1 |
| stdt\_OP | Start date of OP visits | Minimum visit date for OP visits for each patient |
| endt\_OP | End date of OP visits | Maximum visit date for OP visits for each patient |
| dur\_OP | Duration of OP visits | dur\_OP = endt\_OP – stdt\_OP + 1 |
| serstdt | Service Start date | Minimum visit date for Service visits for each patient |
| serendt | Service End date | Maximum visit date for Service visits for each patient |
| Code | Code | Source variable, no derivation needed, ACD code |
| description | Description | Source variable, no derivation needed, description |
| Type | Type of visit | This variable identifies a visit either as IP or OP based on visit classification |
| diag\_type | Diagnosis type | Source variable, no derivation needed:  Primary or Secondary |
| year | Year | Year part of the newdt variable |
| season | Indian seasons | Derivation of Indian seasons based on the date variable for each visit:  # Add Indian rutus as new variables  # <https://www.drikpanchang.com/seasons/season-tropical-timings.html?geoname-id=1277333&year=2010>   * 01 Vasant Rutu * 02 Grishma Rutu * 03 Varsha Rutu * 04 Sharad Rutu * 05 Hemant Rutu * 06 Shishir Rutu |
| C, N, P, U, X, Y | Values related to Services offered to patients | Source variable, no derivation needed:   * C- Cancelled * U - Condn. Unnecessary * Y -Conducted * N - Not Conducted * P - Partially Conducted |
| presc\_type |  | Source variable, no derivation needed |
| medicine\_name | Medicine name | Source variable, no derivation needed  Prescribed medicine names follow a certain predefined naming convention. Medicine name + Quantity + Producer’s name are the details recorded for each prescribed medicine. |
| item\_name | Source value of medicine name | Source variable, no derivation needed |
| quantity | Quantity of prescribed medicine | Source variable, no derivation needed |
| med\_route | Route of administration of prescribed medicine | Source variable, no derivation needed |
| generic\_code |  | Source variable, no derivation needed |
| remarks | Notes provided by doctors for medicines | Source variable, no derivation needed |
| frequency | Frequency of prescribed medicine | Source variable, no derivation needed |
| duration | Duration of prescribed medicine | Source variable, no derivation needed |
| duration\_units | Unit for duration of prescribed medicine | Source variable, no derivation needed |
| Coded\_med | Only name of medicine | Derived from medicine\_name |
| Company | Name of the company producing the drug | Derived from medicine\_name |
| Quantity | Quantity of prescribed medicine | Derived from medicine\_name |
| Unit | Unit of prescribed medicine | Derived from medicine\_name |
| Type\_med | Type of medicine | Derived based on medicine\_name. Classified into different kinds of medicines, e.g.   * Ghritam * Kashayam * Asavam * Aristham * Bhasma * Abhyanga * Cream * Rasayanam * Tablet / Gulika / Vati * … |
| cat\_id |  | Identification of categories |
| distype | Disease type | Disease type as OTHER, RMSD, Metabolic   1. If a disease code is present in Metabolic list then the value is Metabolic 2. If a disease code is present in RMSD list then the value is RMSD 3. Any other disease is classified as OTHER |
| Metabolic | Metabolic | If a patient has reported any Metabolic disease at least once then that patient is given value Metabolic = 1, else Metabolic =0  Metabolic disease group has 10 diseases (Refer [2.4.1.6.1](file:///C:\Users\mahajvi1\Downloads\ThesisPresentations\Word%20files\work-varsha\ThesisWorking-March2021.docx#_Metabolic_and_RMSD_1)) |
| RMSD | RMSD | If a patient has reported any RMSD disease at least once then that patient is given value RMSD = 1, else RMSD =0  RMSD disease group has 97 diseases (Refer [2.4.1.6.1](file:///C:\Users\mahajvi1\Downloads\ThesisPresentations\Word%20files\work-varsha\ThesisWorking-March2021.docx#_Metabolic_and_RMSD_1)) |
| combine | Metabolic  RMSD  Both | 1. If a patient is classified only as Metabolic diseased patient then combine = 1, 2. If a patient is classified only as RMSD diseased patient then combine = 2, 3. If a patient is classified as Metabolic as well as RMSD diseased patient then combine = 99 |
| Minday Metabolic | First day on which reported metabolic disease | First day on which any metabolic disease has been reported by a patient. |
| Minday RMSD | First day on which reported RMSD disease | First day on which any RMSD disease has been reported by a patient. |

# Metabolic and RMSD disease code and de-code

|  |  |  |
| --- | --- | --- |
| Code | Description | Distype |
| M10.0 | Medoroga | Metabolic |
| M10.1 | Medoroga - Sthula medho roga | Metabolic |
| M10.2 | Medoroga - Sukshma medho roga | Metabolic |
| M2.0 | Madhumeha | Metabolic |
| P5.0 | Prameha | Metabolic |
| P5.1 | Prameha - Krusha | Metabolic |
| P5.2 | Prameha - Pidaka | Metabolic |
| P5.3 | Prameha - Sthula | Metabolic |
| P5.4 | Prameha - Upadrava | Metabolic |
| S16.0 | Sthaulya | Metabolic |
| A2.0 | Aamavaata | RMSD |
| A2.1 | Aamavaata - Kaphaja | RMSD |
| A2.2 | Aamavaata - Pittaja | RMSD |
| A2.3 | Aamavaata - Vaataja | RMSD |
| A3.0 | Abhighataja Shoola | RMSD |
| S10.0 | Stambha | RMSD |
| S10.1 | Stambha - Baahu Stambha | RMSD |
| S10.10 | Stambha - Prishtha Stambha | RMSD |
| S10.12 | Stambha - Sandhi Stambha | RMSD |
| S10.13 | Stambha - Siraa Stambha | RMSD |
| S10.14 | Stambha - Uru Stambha | RMSD |
| S10.4 | Stambha - Greevaa Stambha | RMSD |
| S10.5 | Stambha - Hanu Stambha | RMSD |
| S10.6 | Stambha - Hridaya Stambha | RMSD |
| S13.0 | Sthaanabhedena Graha | RMSD |
| S13.1 | Sthaanabhedena Graha - Anga Graha | RMSD |
| S13.11 | Sthaanabhedena Graha - Katee Graha | RMSD |
| S13.13 | Sthaanabhedena Graha - Manyaa Graha | RMSD |
| S13.14 | Sthaanabhedena Graha - Marma Graha | RMSD |
| S13.17 | Sthaanabhedena Graha - Paada Graha | RMSD |
| S13.18 | Sthaanabhedena Graha - Paarshva Graha | RMSD |
| S13.19 | Sthaanabhedena Graha - Prishtha Graha | RMSD |
| S13.20 | Sthaanabhedena Graha - Shiro Graha | RMSD |
| S13.22 | Sthaanabhedena Graha - Uro Graha | RMSD |
| S13.23 | Sthaanabhedena Graha - Vaak Graha | RMSD |
| S13.3 | Sthaanabhedena Graha - Gala Graha | RMSD |
| S13.5 | Sthaanabhedena Graha - Hanu Graha | RMSD |
| S13.6 | Sthaanabhedena Graha - Hrid Graha | RMSD |
| S13.7 | Sthaanabhedena Graha - Jaanugraha | RMSD |
| S13.8 | Sthaanabhedena Graha - Janghaa Graha | RMSD |
| S14.0 | Sthaanabhedena Shoola | RMSD |
| S14.11 | Sthaanabhedena Shoola - Guda Shoola | RMSD |
| S14.13 | Sthaanabhedena Shoola - Gulpha Shoola | RMSD |
| S14.14 | Sthaanabhedena Shoola - Hanu Shoola | RMSD |
| S14.15 | Sthaanabhedena Shoola - Hasta Shoola | RMSD |
| S14.16 | Sthaanabhedena Shoola - Hrid Shoola | RMSD |
| S14.17 | Sthaanabhedena Shoola - Jaanu Shoola | RMSD |
| S14.18 | Sthaanabhedena Shoola - Janghaa Shoola | RMSD |
| S14.19 | Sthaanabhedena Shoola - Kantha Shoola | RMSD |
| S14.21 | Sthaanabhedena Shoola - Katee Shoola | RMSD |
| S14.23 | Sthaanabhedena Shoola - Kukshi Shoola | RMSD |
| S14.24 | Sthaanabhedena Shoola - Manyaa Shoola | RMSD |
| S14.3 | Sthaanabhedena Shoola - Amsa Shoola | RMSD |
| S14.4 | Sthaanabhedena Shoola - Anga Shoola | RMSD |
| S14.5 | Sthaanabhedena Shoola - Anguli Shoola | RMSD |
| S14.6 | Sthaanabhedena Shoola - Asthi Shoola | RMSD |
| S14.7 | Sthaanabhedena Shoola - Baahu Shoola | RMSD |
| S15.28 | Sthaanabhedena Shoola - Nakha Shoola | RMSD |
| S15.31 | Sthaanabhedena Shoola - Paada Shoola | RMSD |
| S15.32 | Sthaanabhedena Shoola - Paarshni Shoola | RMSD |
| S15.34 | Sthaanabhedena Shoola - Parva Shoola | RMSD |
| S15.36 | Sthaanabhedena Shoola - Prishtha Shoola | RMSD |
| S15.41 | Sthaanabhedena Shoola - Sakthi Shoola | RMSD |
| S15.42 | Sthaanabhedena Shoola - Sandhi Shoola | RMSD |
| S15.43 | Sthaanabhedena Shoola - Skandha Shoola | RMSD |
| S15.44 | Sthaanabhedena Shoola - Snaayu Shoola | RMSD |
| S15.45 | Sthaanabhedena Shoola - Sphik Shoola | RMSD |
| S15.46 | Sthaanabhedena Shoola - Stanaanta Shoola | RMSD |
| S15.47 | Sthaanabhedena Shoola - Trika Shoola | RMSD |
| S15.48 | Sthaanabhedena Shoola - Urah Shoola | RMSD |
| S1A.0 | Shoola | RMSD |
| V1.0 | Vaatarakta | RMSD |
| V1.1 | Vaatarakta - Dvandvaja | RMSD |
| V1.2 | Vaatarakta - Gambheera | RMSD |
| V1.3 | Vaatarakta - Kapha Vaataja | RMSD |
| V1.4 | Vaatarakta - Kaphaadhika Vaatarakta | RMSD |
| V1.5 | Vaatarakta - Pittaadhika Vaatarakta | RMSD |
| V1.7 | Vaatarakta - Uttaana | RMSD |
| V1.8 | Vaatarakta - Vaata Kaphaja | RMSD |
| V1.9 | Vaatarakta - Vaataadhika Vaatarakta | RMSD |
| V2.0 | Vaatavyaadhi | RMSD |
| V2.12 | Vaatavyaadhi - Stabdhagaatra | RMSD |
| V2.16 | Vaatavyaadhi - Baahugata Vaata | RMSD |
| V2.23 | Vaatavyaadhi - Gridhrasee | RMSD |
| V2.30 | Vaatavyaadhi - Jaanugata Vaata | RMSD |
| V2.31 | Vaatavyaadhi - Janghaagata Vaata | RMSD |
| V2.36 | Vaatavyaadhi - Kateegata Vaata | RMSD |
| V2.42 | Vaatavyaadhi - Maamsagata Vaata | RMSD |
| V2.43 | Vaatavyaadhi - Maamsamedogata Vaata | RMSD |
| V2.44 | Vaatavyaadhi - Majjaagata Vaata | RMSD |
| V2.45 | Vaatavyaadhi - Majjaasthigata Vaata | RMSD |
| V2.46 | Vaatavyaadhi - Manyaagata Vaata | RMSD |
| V2.47 | Vaatavyaadhi - Manyaastambha | RMSD |
| V2.48 | Vaatavyaadhi - Medogata Vaata | RMSD |
| V2.61 | Vaatavyaadhi - Prishthagata Vaata | RMSD |
| V2.63 | Vaatavyaadhi - Sandhigata Vaata | RMSD |
| V2.64 | Vaatavyaadhi - Sarvaangagata Vaata | RMSD |
| V2.65 | Vaatavyaadhi - Shaakhaagata Vaata | RMSD |
| V2.68 | Vaatavyaadhi - Siraagata Vaata | RMSD |
| V2.69 | Vaatavyaadhi - Siraagraha | RMSD |
| V2.70 | Vaatavyaadhi - Snaayugata Vaata | RMSD |
| V2.72 | Vaatavyaadhi - Trikgata Vaata | RMSD |
| V2.73 | Vaatavyaadhi - Tvaggata Vaata | RMSD |
| V2.74 | Vaatavyaadhi - Urugata Vaata | RMSD |
| V2.75 | Vaatavyaadhi - Vaatakantaka | RMSD |
| V2.77 | Vaatavyaadhi - Vishvaachee | RMSD |
| V2.9 | Vaatavyaadhi - Asthigata Vaata | RMSD |

Data understanding from an observation – patient – disease to a clinical picture



Real World Data – life cycle

|  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Has the study focused on an intervention? | | Is the intervention an approved one? | | | Is it a comparative study? | | | Has treatment been assigned by study protocol? | | | Is data available in existing sources? | |
| Causal diagram: Specify causal relations & supporting evidence among treatment, outcome(s), & other variables to control confounding | | | | | | | | | | | | |
|  |  | |  |  | | All Data Sources |  | |  |  | |  |
|  |  | | Data protection | Different data types | | IT infrastructure | Data quality | | Information governance |  | |  |
|  | Valid sample size | | Clinical outcome | Disease registry | | Patient registry | Patient charts | | Sensor data, Mobile App | Low recall bias | |  |
|  | Medical practitioner bias | | Patient reported outcome | Real World Data  Individual Patient Data (pragmatic trials, cohort trials, observational)  Effectiveness in wider population  Stakeholders: Regulatory Authorities, Policy Makers, Government and Payers | | | | | Longitudinal data | Co-morbidities and Cost effectiveness | |  |
| Methodology context | Low adherence | | Quality of life outcome | Health surveys | Preference of other medicine | | Clinical context |
|  | Confounding and Population homogeneity | | Economic outcome | Hospital EHRs | Real life data, clarity of treatment impact and AEs | |  |
|  | Un-blinded treatment and Treatment switch | | Primary / Secondary data | Retrospective / Prospective study | | Big data, large sample size | Social media | | Individual practice | More data available on drug and life style interaction | |  |
|  |  | | Operational challenges | Comparable data | | Patient level data access | GDPR and Anonymization | | Incomplete data |  | |  |
|  |  | |  |  | | Data context |  | |  |  | |  |

# References