



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

1 Table of Contents

1 Table of Contents.....	2
List of Figures.....	4
List of tables	5
Preface	6
1 Introduction	8
1.1 Origins of Pharmaceutical Industry	8
1.2 Elaboration of Clinical Trials: Origin of RCT Blinded Trials	8
1.3 Modern Hospitals, Everyday Clinical Practice and Healthcare Environment	9
1.4 Real world evidence and observational studies.....	10
1.5 EHR across the globe	11
1.6 Role of Statistics, Analyst, Programmer	12
1.7 The Indian context.....	12
1.8 National level efforts AYUSH	13
1.9 Experts' view – hinting towards real world evidence.....	15
1.10 Potential opportunities for Real World Data analysis within Ayurveda	16
1.11 What this study aims to contribute to	17
1.11.1 Hospital management	17
1.11.2 Clinicians or treating doctors	17
1.11.3 Universities and students	18
1.11.4 Policy makers – AYUSH and relevant ministries, insurance sector:.....	18
1.12 Introduction to real life data.....	18
1.13 Introduction to Clinical data understanding.....	20
1.14 Introduction to study of demographics and patient characteristics	20
1.15 Introduction to study of diagnostics and interventions	21
2 Methods	22
2.1 Study design.....	22
2.2 Data analysis design	22
2.3 Converting real life clinical data into analyzable format.....	22
2.3.1 Data access.....	22
2.3.2 Data preparation.....	23
2.3.3 Data derivation	26
2.4 Clinical data understanding.....	33
2.4.1 Broad checks on the datasets.....	33
2.4.2 Contents checks.....	33
2.4.3 Visit pattern analysis	33
2.4.4 Patient disease and treatment journey view	33

2.5	Studying demographics and patient specific factors.....	34
2.6	Studying diagnostics and interventions.....	36
3	Results	40
3.1	Converting real life clinical data into analyzable format.....	40
3.1.1	Details of the database	40
3.1.2	Data Extracted from Hospital Database.....	40
3.2	Clinical data understanding.....	41
3.2.1	Broad checks on the datasets.....	41
3.2.2	Contents checks.....	43
3.2.3	Visit pattern analysis	44
3.2.4	Patient disease and treatment journey view	45
3.3	Studying demographics and patient specific factors.....	47
3.4	Studying diagnostics and interventions.....	54
4	Discussions	80
4.1	Converting real life clinical data into analyzable format.....	80
4.2	Clinical data understanding.....	80
4.3	Studying demographics and patient specific factors.....	87
4.4	Studying diagnostics and interventions.....	88
5	Conclusion	91
6	Appendix	98
6.1	Approval from the hospital management to carry out the retrospective study.....	98
6.2	Details of analysis dataset	99
6.3	All variables in the source database	108
7	References.....	109

List of Figures

Figure 1-1: RCTs on ClinTrials.gov	9
Figure 1-2: Screenshot of a dashboard from AYUSH website	14
Figure 1-3: Ayurvedic practitioners and patients from AYUSH dashboard.....	14
Figure 2-1: Pictorial representation of analysis	22
Figure 2-2: Flow diagram from data source to final usage by various usage types.....	25
Figure 2-3: A glimpse of data tables used to store source data from the database	27
Figure 2-4: Extraction of relevant data from Source database.....	28
Figure 2-5: Staged data converted into 6 datasets.....	29
Figure 2-6: Staged data.....	30
Figure 2-7: Final dataset with ~30 source variables and ~30 new derived variables	31
Figure 2-8: Data flow from source data to interpretable results	32
Figure 3-1: A snippet of disease table by gender	42
Figure 3-2: Variable classification by categories	43
Figure 3-3: Visit pattern analysis.....	44
Figure 3-4: Patient visit profile – Horizontal view	45
Figure 3-5: Patient visit profile – Vertical view	46
Figure 3-6: Total Number of Patients.....	47
Figure 3-7: Country-wise Visualization	47
Figure 3-8: Age distribution by country, age distribution by gender	48
Figure 3-9: Blood-group Distribution by gender	49
Figure 3-10: Number of Visits, and Visit Types.....	49
Figure 3-11: Descriptive summary statistics by number of Diseases by Age and Gender.....	50
Figure 3-12: Data tabulation for patients reporting RMSD and Metabolic diseases	51
Figure 3-13: Disease distribution by age and gender	52
Figure 3-14: Patient visit duration for Disease categories by Gender	52
Figure 3-15: Disease distribution by Seasonal Variations and gender	52
Figure 3-16: Pre and Post Disease Classification Analysis	53
Figure 3-17: ICD classification by Gender	60
Figure 3-18: Age distribution by ICD classification and Gender	61
Figure 3-19: Visit distribution by ICD classification and Gender	62
Figure 3-20: Duration distribution by ICD classification and Gender.....	62
Figure 3-21: Disease classification by Prakriti and Gender	63
Figure 3-22: Co-morbidity analysis approach 1 example 1: Vaatavyadhi	64
Figure 3-23: Co-morbidity analysis approach 1 example 2: Pandu	64
Figure 3-24: Co-morbidity analysis approach 1 example 3: Madhumeha	65
Figure 3-25: Co-morbidity analysis approach 2	65
Figure 3-26: Co-morbidity analysis approach 3: collapsible tree view	65
Figure 3-27: Patient Disease and Treatment administration by Study Day	66
Figure 3-28: Patient Disease by Study Day and Treatment administration by Study Day	67
Figure 3-29: Patient Cumulative Disease and Treatment administration by Visit	67
Figure 3-30: Area graph representation of diseases	67
Figure 3-31: Mosaic plot: Disease and treatment representation example 1: Prameha	68
Figure 3-32: Disease and treatment example 2: P5.0: Prameha and Oil: Kottamchukkadi.....	70
Figure 3-33: Disease and treatment example 3: P5.0: Prameha and Vati: Diabecon DS	71
Figure 3-34: Mosaic plot Disease and treatment representation example 4: Treatment: Oil: Kottamchukkadi.....	72
Figure 3-35: Cross tabulation of prescribed treatments and disease group by gender Example 1	73
Figure 3-36: Cross tabulation of prescribed treatments and disease group by gender Example 2	74
Figure 3-37: Circular view: Co-occurrences of disease – disease Example 1	75
Figure 3-38: Circular view: Co-occurrences of disease – treatment Example 2	75

Figure 3-39: Pre and Post distance analysis for disease: M2.0: Madhumeha	76
Figure 3-40: Pre and Post distance analysis for medicines given for diseases: P5.0, V2.23, V2.63	76
Figure 3-41: Radar plot	77
Figure 3-42: Dynamic bubble plot: Example 1: Disease: A6.0: Amavaata.....	79
Figure 4-1: Treatment principles defined in different texts.....	82
Figure 4-2: Data understanding from an observation – patient – disease to a clinical picture.....	85
Figure 5-1: Real World data - life cycle	93

List of tables

Table 2-1: Visit window table for Pre and post analysis	35
Table 3-1: Versions of data used for analysis.....	40
Table 3-2: Summary statistics and t-test for bhasma usage.....	74
Table 4-1: Proposed vital sign data structure	80
Table 4-2: Proposed idea for clinical coding timetable	83
Table 5-1: Different types of documents	95
Table 6-1: Details of the Reference Dataset “01adsl_met_rmsd”.....	99
Table 6-2: Metabolic and RMSD disease code and de-code	106
Table 6-3: All variables in the database	108

Preface

Various electronic media like computers, mobile devices, wearables, and other sensors collect and store huge amounts of health-related data. This explosion of data carries potential to better design and conduct clinical studies to answer questions previously thought infeasible.

Advancement of cutting-edge analytical capabilities is allowing researchers to analyze and comprehend this data at greater depths, permitting medical product development and approval at an accelerated speed [1]. Real world data is the information relating to patient health status and/or the delivery of health care routinely collected from a variety of sources like epidemiological studies, clinical practice, already published articles to answer questions previously thought infeasible.

Approval of Ibrance by US FDA for male breast cancer, a drug already approved for females, French health authorities allowing an RWE study of 600+ patients, over a period of 18 months, for a conditional re-imbursement scheme in COPD, are a couple of examples of approvals using RWD data. A study carried out by Clarivate Analytics, USA, reports 27 (non-exhaustive list), <5% of all approved drugs, examples of drug approvals by US FDA, EMA, Japan's PMDA and Health Canada, across broad spectrum of medicines between years 1998 and 2019. Real world data from Electronic Health Records and registries were used either as primary data, when non-comparative data were available to demonstrate tolerability and efficacy, or as a supportive data when validating findings. This provides increasing usage of "naturally reported data" in drug approvals in modern biomedicine. These examples provide evidence of novel use of data, which may have otherwise gone unused. The power available to society would have never been unearthed if not for this way of use [2].

Is Ayurvedic area dealing with the same type of challenge of not realizing the potential of available data? Just to give a glimpse of enormity of data: more than 10 crore number of patients have been reported on AYUSH website (As of May 2020), more than 140+ countries have population of less than 10 crores [3].

It is safe to assume that the conceptual developments in Ayurvedic knowledge base have taken place through everyday observations and basic laws of nature. These fundamentals have been adjusted to the relevant times as per the passage of time based on observations and experiences, where there are no artificial restrictions on usage of medicines, duration of treatment or type of patients to treat, which is next to impossible in a protocol driven clinical trial setting [4] [5].

Taking inspiration from respected Prof Patwardhan's quote, "Charaka would not have ignored modern technologies if they had been available during his time" [6], this study attempts to discover hidden wealth of Ayurveda related information in EHRs created at TDU hospital using modern methods of data sciences and statistical programming. Since 2011 to Oct2017, the hospital database contained data for >51,000 patients, >1,50,000 visits, > 900 disease types, >3,000 variations of medical procedures [7]. This study targets the methodological and learning framework covering short-, mid- and long-term influences in following categories:

- Hospital managements, clinicians, and patients

- Universities and learning institutes – clinical communication, researchers to build vital evidence-base
- Policy makers – AYUSH and relevant ministries
- Healthcare providers - Ayurveda Healthcare systems, General healthcare systems

1 Introduction

1.1 Origins of Pharmaceutical Industry

The origins of the pharmaceutical industry go back with the apothecaries and pharmacies that gave traditional therapies going back to the Middle Ages. The modern medicines regulation began only after revolutionary development in the 19th century life sciences, in chemistry, physiology and pharmacology, which put a robust foundation for the modern drug research and development [8]. Unfortunate events like deaths due to diethylene glycol poisoning in the US in 1930s, the thalidomide disaster in late 1950s, catalyzed the development of medicines regulation more than the evolution of a knowledge base throughout the world. The formation of ICH consortium was one such byproduct. The use of statistics to support R&D of new medicines grew multifold since the Kefauver-Harris Amendments (1962) and continues to grow [9]. These clearly stated that the Food and Drug Administration (FDA) would require “substantial evidence” of the impact of a drug in a clinical trial setting and proof of safety will not be sufficient for new drug approvals. In the USA and all over world, since 1970s, the value of medicine has been clearly exhibited by a longer life expectancy, a lower infant mortality rate, and the higher quality of life.

1.2 Elaboration of Clinical Trials: Origin of RCT Blinded Trials

Randomized Controlled Trial (RCT) is a classical research design of randomly allocating participants to one or the other treatments under investigation. Randomization is the fundamental characteristic of an RCT, and it describes the random distribution of participants to the study arms. RCTs aim towards supporting a conclusion that the difference in the outcomes among participants in study arms was exclusively caused by the intervention, as randomization equalizes the study group in all other factors. Thus, RCTs set the standard of excellence in health sciences research.

Although, randomization reduces the bias of participant assignment to the intervention and control group, it does not rule out the chances of biases due to investigator/caregiver or the patient or biases during the adjudication of the outcome variables. However, RCTs help in reliable description of causality.

Blinding assists in monitoring several types of biases that might unintentionally creep into the study. The two major biases, namely performance bias and the ascertainment bias that can be controlled using blinding. The four groups of people blinded in the trial are the study participants, the investigator(s), the outcome assessor(s), and the data analyst(s). Depending upon the number of individuals blinded, trials are categorized as open label, single blinded, double blinded, triple blinded, and quadruple blinded trial [10].

Physicians and clinical researchers have remained confident that RCTs deliver the most rigorous test of preventive, diagnostic, and therapeutic interventions. They are universally denoted as the “gold standard” of experimental medical analysis, as an undisputable starting point in diagnostic or therapeutic evaluation. When did RCTs become the “gold standard”? An article written by Alvan Feinstein and Ralph Horwitz in The New England Journal of Medicine (NEJM) in December 1982, is the first instance of RCTs being referred to as “gold standard” [11].

1982 appears late date for the first usage. They quoted that they were eager to be proven wrong on this 1982 date. But despite massive searches across many textbooks, conference materials, journals, and archival collections have not thrown up any earlier date [11]. While probing on details regarding RCTs on the ClinTrial.gov website the following data was seen on 24th April 2020 [12], [Link](#). Total trials in the ClinTrial.gov database = 3,36,905. The estimated % of RCTs based on the results range between 2.40% (only considering “Randomized control trial” and “randomized clinical trial” 1,862 + 6,234) to 19.31% (considering “Randomized control trial”, “Randomized” and “randomized clinical trial” 1,862 + 56,945 + 6,234), remaining 80% to 97% of the registered trials are not RCTs (Figure 1-1).

Figure 1-1: RCTs on ClinTrials.gov

Terms	Search Results*	Entire Database**
Synonyms		
Randomized control trial	1,862 studies	1,862 studies
trial	56,745 studies	117,642 studies
Clinical Trials	21,302 studies	42,784 studies
CLINTRIAL	17 studies	67 studies
control	50,668 studies	188,230 studies
Comparator	37,863 studies	123,329 studies
Controlled	31,962 studies	66,966 studies
Control Group	7,483 studies	24,005 studies
controlling	390 studies	1,772 studies
Randomized	56,945 studies	178,454 studies
Randomization	6,723 studies	12,452 studies
RANDOM	1,267 studies	3,526 studies
randomized clinical trial	6,234 studies	6,234 studies
clinical randomized trials	35 studies	35 studies
clinical trial	56,945 studies	118,708 studies
trial	56,743 studies	117,627 studies
Intervention Study	551 studies	1,599 studies
CLINTRIAL	17 studies	67 studies
clinical	32,693 studies	138,751 studies

This is a screenshot taken from the ClinTrial.gov website showing number of Randomized Clinical Trials

As of April 2020, Clintrial.gov database had 19% RCTs of all the studies registered in the database. This provides us with enough evidence that a lot of research work is carried outside of the RCT framework.

1.3 Modern Hospitals, Everyday Clinical Practice and Healthcare Environment

Health information provided by the hospital is vital for public health administrators. Such data, accumulated from all the relevant hospitals, are essential in formulating health planning for the district / state / country, by matching with population statistics, other demographic data, and the

health resources of the district / state / country. This medical information, containing not only data about the patient's illness but also the success or failure of therapy, is a precious source of clinical training for medical students, nurses, and related health care work force. This is a basis of new medicine and therapies. The development of new therapeutic drugs is built upon the careful observations of experienced physicians, and the comparison and analyses of the effects of previously administered drugs, gathered from the patient's data. This in turn leads to continued medical progress in uncovering a new entity of a disease or a new method of diagnosis [13].

One trend all modern hospitals have in common is that the amount of processed information generated per patient is constantly increasing e.g., EMR, patient charts, CT scans, X-rays, ECGs, pathology reports, wearable devices, sensor data, etc. When such information is properly compiled and aggregated, it provides data for efficient hospital administration. How each patient treated can yield important statistical data, while the number of patients handled, grouped by sex, age, and diagnosis, provides basic information required by administrators [14].

This information is processed manually or partially automatically in many instances. This inefficient method results in delays in reaching crucial treatment decisions, for instance when test results fail to get back to the physicians in time, or in delayed action by administrations when a prompt response is needed.

1.4 Real world evidence and observational studies

Real-world data are data captured in an observational manner, in a natural, uncontrolled setting – outside of traditional clinical trials. Ordinary clinical practice, producing a never-ending flow of results from everyday practice, can be viewed as infinite sequence of unsystematic observational studies. There must be a rational use of outcome data from patients, far more representative of the general population than those included in formal clinical trials [15]. Practice-based medicine is an important way to advance science [16]. For this to happen, the treating doctor should be a “clinician researcher investigator”. Doctors even in their busy practice, can see patterns, then formulate a research hypothesis, and then proceed to test it. Hence, a doctor can be good at good clinical practice (GCP) and good clinical research practice (GCRP). One can support the other and vice versa, and this way medicine advances [17].

Investigation of the literature across time points from year 2000 to year 2020 presents that well designed observational studies could be producing treatment effects like the RCTs:

Year 2000: 99 reports across 5 clinical topics covering observational studies provided comparable results like those of the randomized, controlled trials. There was no systematic over estimation of treatment effects seen [18].

Year 2007: RCTs provide one kind of knowledge but stop us from understanding other properties of a drug. When epidemiologic studies are carried out correctly can be both more conceptually demanding and more powerful than the average RCT, especially in assessing drug safety. Both kinds of research must be done with rigor and with humility [19].

Year 2014: There was little evidence found for significant effect estimate differences between observational studies and RCTs regardless of specific observational study design, heterogeneity,

or inclusion of studies of pharmacological interventions (Pooled odds ratio of 1.08). These results highlight that it is important for review authors to consider not only study design, but the level of heterogeneity in meta-analysis of RCTs or observational studies [20].

Year 2020: Cumulative evidence suggests that appropriately conducted RWD studies have the potential to support regulatory decisions in the absence of RCT data. Further work may be needed to better show the settings in which RWD analyses can robustly and consistently match the results of RCTs and the situations in which they cannot match. After careful consideration of the potential for bias, regulators can then determine when they would accept RWD in place of an RCT [21].

The data from RWE studies and Observational studies have a wealth of information which if processed accurately and summarized with in a structured manner can lead us to numerous assertions and confirmations. There have been many attempts to generate clinical evidence from primary health care by systematic utilization of patient records. But this has not been easily possible due to the deficient clinical data, in-accurate input “garbage in” leading to “garbage out”, insufficient follow-up, and very few fully completed case records with risk factors, co-morbidities, etc. Standardized protocols, checklists and assessment scales would contribute to good clinical practice and further assist clinical research.

Both clinical practice and research based on the principles of primary care would contribute substantially as follows:

Clinical practice (1) uses knowledge – supports confirmation, (2) focuses on individual patient – generates data, (3) has a short action span and immediate reward – quick decision, (4) it regards authority, follows custom, earns income, and encourages research. Whereas clinical research (1) Creates knowledge – supports future clinical practice, (2) focuses on groups – generates data, (3) has a long action span and delayed reward – leads to discovery and new pathways, (4) questions authority, challenges custom, earns reputation, and enriches practice.

1.5 EHR across the globe

In early 2000s countries like Canada, UK, New Zealand, Estonia, etc. started collecting data at national level. The key challenges experienced while implementing Electronic Medical Records (EMR) were: infrastructure creation, Policy & regulations, Standards & interoperability, and Research, development & education [22].

EMR data has been widely used for analysis and many papers have been published. These have generated supportive data for a variety of clinical outcomes, evaluation methods, and implementation of new technology or intervention along with awareness of unintended consequences; thus supporting the clinical decisions and aiding to improve the healthcare process or clinical outcomes.

Resource shortages, amplified by various socio-economic problems pose substantial difficulties for development of workable healthcare solutions which are global in nature. Healthcare development is one of most important aspects for the progress of both social and economic development of the world. More than 1/4th of the world population has unmet healthcare needs.

On World Health Day 2018, the WHO introduced the concept of Universal Health Coverage (UHC): healthcare for everyone and everywhere. However, this has not been completely effective due to the financial limitations within systems and factors like aged-population-related chronic diseases influence the availability, accessibility, and quality of care [23].

1.6 Role of Statistics, Analyst, Programmer

Statistics emerged as an extended stream from mathematics, operational research, and economics. Application of statistics in various fields of research like genomics, epidemiology, nutrition, biological science, biomedical research connoted the word “biostatistics”. Apart from biological sciences, statistics is applied in variety of other fields like market research, insurance, trades and stocks, banking etc. The 1990s presented explosive increase in applications of computationally demanding methods originated on statistical principles, due to the emphasis on translational exploration, individualised treatment, and the importance of data. Statistical modeling has become more complex due to the volume of data, computational requirements, and varied data sources. These developments led to creation of new roles like statistical programmer - statistical analyst, clinical programmer - clinical analyst [24]. Clinical data analysts (or clinical informatics analyst) are healthcare professionals responsible for confirming the validity of scientific experiments and data gathered. They apply their knowledge of data acquisition, management, analysis, and interpretation to healthcare data, providing actionable insights that doctors, clinical scientists, and others can use. They may be responsible for automating internal and external reports, creating executive-level dashboards, and presenting information to help various stake-holders understand the operational impact of the data. Data analysts ensure that processes and protocols are followed, thus improving overall care [25], [26], [27]. They provide data insights that drive clinical process improvement, such as reducing readmissions and hospital-acquired conditions. In addition, healthcare analysts help insurers, vendors, and others synthesize data that guides decision-making, population health management, cost containment, and quality improvement [28].

Clinical data analysts generate and provide the results of clinical business intelligence to management, stakeholders, and concerned Line functions. They coordinate with other relevant departments (e.g., clinical strategy, clinical operations) to determine the areas to be analyzed as well as the appropriate measures that should be taken to ensure data analysis proves useful. This role can combine strengths from multiple areas to build a powerful storyline on six dimensions of quality care: safe, effective, patient centered, timely, efficient, and equitable [29].

1.7 The Indian context

India has a mixed system of healthcare consisting of many government hospitals, the private hospitals, family doctors and private medical practices. We see this trend reflected in the actual health seeking behaviour of communities where people tend to combine medicine systems like Allopathy, Ayurveda, Siddha, Sowa Rigpa, Unani, Homeopathy and Yoga depending on the nature of the disease [30]. India, being one of the largest rural populations in the world, it becomes imperative to be both inclusive and optimal in the use of the available resources from any stream of the healthcare. Along with the global healthcare challenges, Indian healthcare faces specific issues like inadequate healthcare resources, insufficient funding, poor healthcare

infrastructure and rural–urban disparity. The national healthcare budget is a lowly single digit % of the Gross Domestic Product (GDP). The number of doctors, nurses and health workers are small in comparison to the high population in certain regions. The availability and utilization of technology is also at the lower end of the usage [31].

Integrating modern medicine and AYUSH, folk medicines as well as technological advances into the India healthcare system can generate new opportunities for healthcare. These efforts will be noteworthy where healthcare systems have limitations in terms of infrastructure, expertise, and human resources. This integration will not only allow utilization of the positives of each of the medical streams but will generate data-based evidence due to technological integration. Ongoing monitoring and updating the patient’s data together with consistent updates to physicians can help in identifying epidemiological patterns. Through this analysis, authorities or the government can employ specific precautions, or even start necessary facilities to avoid healthcare issues before the situation gets out of hand [32].

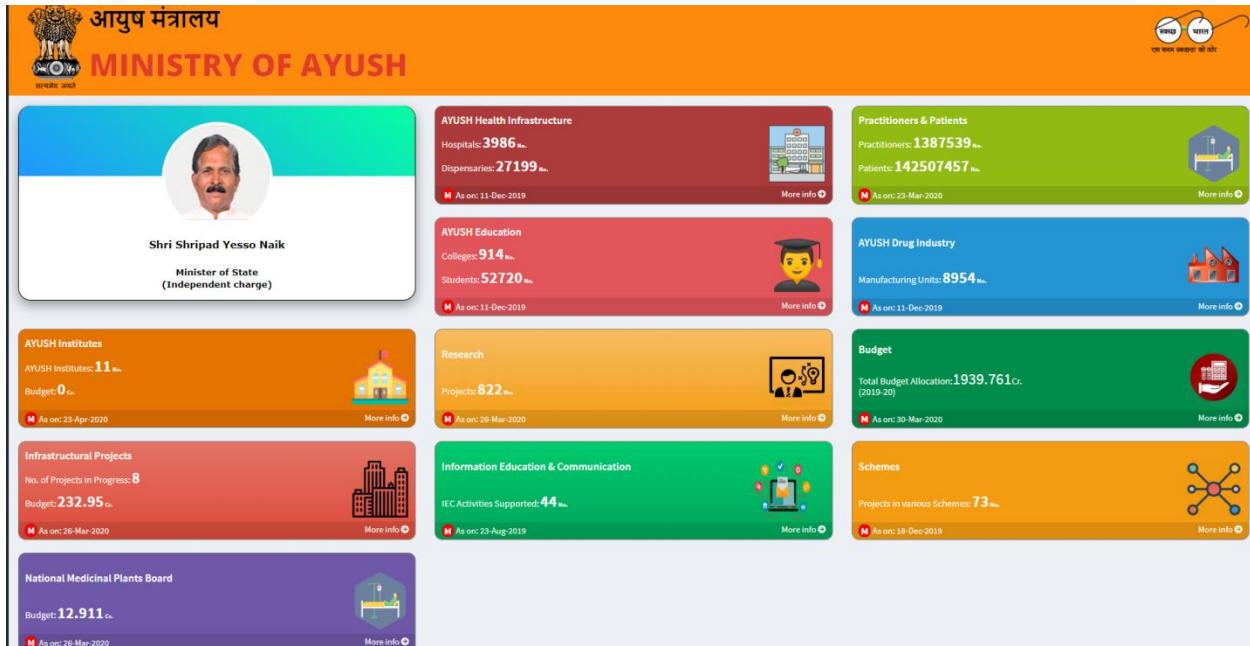
The applications of such assessed and analyzed data are substantial as the information brings out more and more knowledge about an area or a person. In under-resourced countries such as India, the resultant benefits are expanded further, suffering is reduced, hospital resources are saved and socio-economic improvements that lift a nation’s wellbeing are recognized. Moreover, India could achieve improved healthcare delivery, care audit, epidemiological study and quick response to epidemics and bring economic benefits to individuals by reducing the healthcare cost.

1.8 National level efforts AYUSH

This section covers efforts put in by the Government of India at the National level. The Ministry of AYUSH is a part of Government of India to promote and expand use of AYUSH systems of health care and medicine in India (Ayurveda, Yoga & Naturopathy, Unani, Siddha and Homoeopathy). To administer the different medicine systems encompassed by the Ministry of AYUSH, it has five research councils or departments, affiliated courses, and affiliated national institutes [3]. One of AYUSH affiliates, CCRAS has developed National AYUSH Morbidities and Standardized Terminologies E-Portal NAMASTE portal. It provides information about Standardized terminologies and Morbidity codes along with dedicated data entry module for updating morbidity statistics in consolidated form as well as on real time basis [33]. To publicize the merits of AYUSH systems across the globe, a web-based portal for Research publications was launched in 2011, which is being maintained by NIIMH Hyderabad [34]. Ayurveda Hospital Management Information System (A-HMIS) is a complete IT platform for all functions of health care delivery systems and patient care in AYUSH centers. THERAN (THE Research Application Nexus): HMIS is developed by Siddha Central Research Institute, Chennai under Central Council for Research in Siddha. These are 2 additional examples of ongoing work under Ministry of AYUSH [35]. Ministry of AYUSH has created AYUSH GRID to cohesively implement projects under Digital India Movement. It is an amalgamation of IT projects planned for advancement of AYUSH pan India [36]. The dashboard available on AYUSH homepage (accessed on 25th May 2020), reveals the following facts: the screenshot shows almost 10 crore patients treated at some or the other point (Figure 1-2, Figure 1-3). There are approximately 140+

countries with less than 10 crore population, which provides a perspective on the size of data available at the AYUSH level. It remains to be seen how to convert “data into information, information into knowledge and knowledge into wisdom” [37].

Figure 1-2: Screenshot of a dashboard from AYUSH website



Dashboard from AYUSH website covering: AYUSH institutes, Infrastructure projects, National medicinal plants board, Health infrastructure, Education and communication, Research, Practitioners and patients, Drug Industry, Budget, and Schemes

Figure 1-3: Ayurvedic practitioners and patients from AYUSH dashboard

Practitioners & Patients

Practitioners									
#	Facility	Ayurveda	Unani	Siddha	Yoga	Naturopathy	Homoeopathy	Sowa-Rigpa	Total
1	Registered Practitioner(IQ & NIQ)	443704	51110	9125	0	2485	293455	0	799879
2	Institutional Qualified Registered Practitioner	294162	38672	5685	0	2349	246792	0	587660

#	System	IPD	OPD				Total	
		CHC	CHC	Dispensaries	PHCs	IPD	OPD	
1	Ayurveda	1390950	16557395	71486719	8455545	1390950	96499659	
2	Unani	102884	2220231	8109909	1504284	102884	11834424	
3	Siddha	210016	10617115	1634022	16858888	210016	29110025	
4	Yoga	15741	1564861	50150	1282064	15741	2897075	
5	Naturopathy	18636	138187	210634	72176	18636	420997	
6	Sowa Rigpa	0	0	7050	0	0	7050	

Close

Screenshot from AYUSH website covering: Approximately 4.5 lakhs Ayurvedic practitioners and more than 10 crore treated patients

1.9 Experts' view – hinting towards real world evidence

There are some models proposed by various authors to handle complex and tricky situations arising in defining and understanding the action mechanism of Ayurvedic intervention. As described by Dr. Girish Tillu in his talk at TDU, huge observational data for Ayurvedic medicines covers more than 1,00,000 books and manuscripts, 57 authentic books (Drug and cosmetic act 1940), > 4500 diseases including subtypes and conditions (Ayusoft database), > 81,000 formulations (TKDL database), > 4,00,000 Practitioners (Planning Commission - 11th Plan) in India, infinite documents, references, experiential data, living tradition and knowledge in public domain [7]. Dravyaguna (Pharmacology), Bhaisajya Kalpana (Pharmaceutics), Nidana (Diagnosis) and Chikitsa (Management principles). This data points to a validated knowledge base.

Dr. D. B. Vaidya has explained the concept of reverse pharmacology to understand the action mechanism of Ayurvedic intervention. He says that there is huge amount of observational available and relatively low side effects have been reported in due course. This provides an opportunity to carry out large interventional Ayurvedic trials to assess safety, efficacy, and pharmacokinetic information. This approach should be economical and could be less time consuming compared to the hierarchical model used in western medicine. He further talks about integrating meticulously documented experiential and experimental observations [38].

Prof R. H. Singh has opined that lab-based research experiments within Ayurvedic area during the last 50 years have not been rewarding. On the other hand, literary experiments to make a few of the classical Ayurvedic texts accessible to masses have been extremely useful. This situation warrants newer strategies of scientific research without compromising on the fundamental principles of Ayurveda [4].

Prof. Bhushan Patwardhan writes that there are substantial similarities between the traditional systems like Ayurveda and modern medicines. Ayurveda emphasizes on health promotion, disease prevention, early diagnosis, and personalized treatment. The modern medicine system approach uses predictive, preventive, and personalized medicine (PPPM). In case of Ayurveda, the evidence can be drawn from two main sources: (1) Evidence based on historical and classical nature of clinical practice supported by credible and accepted documentation. (2) Evidence based on ongoing scientific research to support various theories, medicines and procedures used in Ayurvedic medicine [39] [40].

Dr. Ram Manohar has expressed that Ayurveda is based on 5000 years of clinical practice. Hence, practice-based clinical trials should complement natural ways to gain insights [41].

Dr. Baghel's interpretation is that one should think of Ayurveda being in the developmental phase like any other medical systems. Like many other scholars he thinks that Ayurveda is a pure science based on logical explanation, which is called *Darshana*. Ongoing research in Ayurveda should impact academics, pharmacy, and practice in a profound way to convert data into information, information into knowledge and knowledge into wisdom [42].

As per Prof. Darshan Shankar's analysis as of 2018, at the national level, Ayurveda receives a meagre 2.5% of the Central health Budget and at the State level, making it difficult to fund any meaningful research projects. Despite Ayurveda's strengths, it has some limitations in current scenario. To advance the science there is a need to embrace tools of information technology to organize its vast multifaceted data, in searchable formats. Meticulously documented clinical experiences interpreted through Ayurveda-biology will expand rejuvenation of healthcare in India.

All the thought leaders cited here point to the strengths of Ayurveda as well as the immediate needs. They have pointed out that the research must be of high quality, and it must be impactful. They have indicated the need for experimental as well as experiential research. They have already provided a few new solutions and have urged to the research community to find new ways of tackling problems [43].

1.10 Potential opportunities for Real World Data analysis within Ayurveda

The western medicines are developed using a method called as hierarchical method where it tries answering the questions with limited scope e.g., what is the efficacy of a particular drug, what is the safety profile of a drug? This method assumes a step wise approach and deals with the problem in successively conducted clinical trials of various types in a specific sequence. The pharmacology of the molecule is ascertained first at the very beginning. These studies are followed by cohort studies, Open-label randomized studies. The clinical trial testing usually concludes with the blinded, randomized, placebo-controlled trials (RCT). The RCTs offer most internal validity and reduce the bias. These studies could be complemented by then moving onto case studies, case series. This "one step at a time" approach has worked very well in the western medicine framework.

Ayurveda has been practiced for more than a millennium and is widely accepted in India as a worthy medical system. Over the last few decades' people across the world have gained knowledge and realized the importance of the age-old medical system and are constantly driven towards it. Although, having been in practice for ages Ayurveda still does not enjoy the recognition which the Western medicine does. Hence there needs to be a structured approach towards making this possible. The untapped potential of Ayurveda needs to be scientifically communicated globally for a wider reach for it to be utilized as a public health tool for promotion of health and prevention of diseases [44].

Ayurvedic vaidya usually use paper-based case report to record a patient's Ayurvedic parameters along with other details of medical consultation. These are not typically exchanged with other vaidyas. There is huge amount data available on paper if digitized could be a big revolutionary step. Increased use and interoperability with electronic medical records of digital Ayurvedic patient management systems are required. Based on a report published by AYUSH [37], there are 4.5 lakhs registered Ayurvedic practitioners. Even if 5% of doctors start using EMRs, i.e., 22,500 doctors and if data for 2 new patients is entered every day (~225 working days) for the whole year, 50 lakh unique patient data can be generated in a single year. Currently, this gold mine of data is not yet built.

1.11 What this study aims to contribute to

What this study aims to contribute to, and which stakeholder would be benefitted the most, the usage of tools and information generated through this exercise could be interchangeably used by interested audience:

1.11.1 Hospital management

To keep the hospital functioning smoothly, operational insights from the routine hospital data are important to improve management and efficiency of day-to-day activities. How many patients are present in the database? What are the characteristics of these patients? What is the gender distribution, age group distribution? Which countries, states, cities do they come from? How many times do they visit hospital? What is the number of In-Patients & Out-Patients? What kind of assessments are done at each visit? What is the duration of visits for a patient? Which diseases are getting treated?

The hospital database does not have a true “hard clinical endpoint” captured in the database, making it difficult to understand beneficial or adverse impacts on patients, so what kind of “endpoints” be created? A simple 1-10 scale-based scoring both by treating physician and the patient at each visit for each reported disease? Regular analysis of data would allow operational proficiencies, cost saving and profitability.

1.11.2 Clinicians or treating doctors

Clinicians can learn about the treatments prescribed and their variations as reported in the database: Clinical practice characteristics and thought process can be retrospectively created by looking at the data. Disease variations as seen in the database – What does this indicate? How many times the same disease is reported by a patient? Are there differences seen in the diseases by gender, season, age group? How many disease - disease combinations can be identified, some could be clinically meaningful, some may be clinically meaningless, and some could be rare combinations. What kind of diseases are reported before and/or after an onset an onset of a disease? How many doctors treat one patient through the course of their diseases at the hospital? How many different treatments are prescribed -herbo mineral, classical, ras-aushadhi, etc.? Which Ayurvedic procedures are ordered for which diseases? What treatment, and treatment - procedure combinations occur? Summary statistics for treatments, Frequency, duration of a treatment. Hospital management system usage has its own day-to-day challenges. How can experience of doctors using the Health Information system be viewed? Do doctors like data entry part of job? Do they see this as an additional burden or an integral part of the work? Do the doctors have time for real data entry while consulting patients? Data review and analysis tools developed for the thesis can enhance patient and doctor interactions.

1.11.3 Universities and students

Over the years, teaching methodology has not transformed even though there are quite a lot of advances in modern methodology. Can learning objectives of different kinds be tackled by making complex relationships easily visually available. Can interesting ways of explaining text and advice from Ayurveda which have contemporary relevance be created? E.g., occurrence of certain diseases in certain geography areas or season. **Add a relevant shlok** Very few large-scale studies like this have taken place in Ayurveda, so this study with large amounts of natural data can provide exploration opportunities to describe some findings, textually as well as diagrammatically. Tools developed here can be used as a supplementary material for any MD / PhD student.

Scientific literature generation by researchers: Hospitals in India or any part of the world focus on treatment and not on research publications -- can a team be put together for medical communication who are publishing papers as their primary job? If there is no known profile of patients visiting the Ayurvedic hospital and if this data is represented in the right form, then will it provide novel information? How can we measure the strengths and weaknesses of the practice? How to evaluate changes in results over time? Is it possible to build new hypothesis? Is prescribed treatment truly personalized? Is it possible trace back the treatment regimen followed as per the classical fundamentals? Is there a way to compare the demographics and patient characteristics from our hospital against the mainstream western hospital? Which are the rare diseases identified in the database and any clinically meaningful document could be written, like a case series? Can TDU initiate a registry for Ayurvedic data from other hospitals, private Ayurvedic doctors, yet another attempt to broaden evidence.

1.11.4 Policy makers – AYUSH and relevant ministries, insurance sector:

How can anyone use these data as “secondary use”? Can this data be used by insurance companies which has not been used so far? Do the approved labels of medicines and prescriptions in the database match each other? Metal based formulations are questioned by non Ayurvedic community, what insights can be drawn about the rasa-aushadhis? Which are these medicines? For what diseases are they given and for what duration? Before providing the metal-based treatment and after providing the metal-based treatment, is there any difference in duration seen in treatments? What is the % of patients prescribed these medicines and what is the % of duration of all the duration of treatment given to these patients?

One way of getting some of these answers is to convert real life data into analyzable format, then study the clinical data, study demographics and patient characteristics, and study diagnostics and interventions.

1.12 Introduction to real life data

With the passage of time, revolutions in technology have continually increased the creation of information and its exchange. With the advancement for communication from spoken to written, it became simpler to create texts, books thereby documentation; thus aiding transfer of knowledge from one person to another as well as from generation to generation without losing any data in translation. With increased and improved writing, compilation of articles, tables and records, there came a time where storing them became important, thus came in the libraries. The

ability to effortlessly widen accumulated data had to wait until the 15th century. Around 1439, Johannes Gutenberg developed the printing press, causing an astonishing growth in the sharing of information at an economical cost.

The 20th century generated a remarkable growth in the publication of scientific journals and monographs, most of which were not critically reviewed, as most physicians had no way to access to the existing medical information. Towards the late 20th century, the spread of computers and the internet providing immediate virtual access to diverse information has entirely changed the way knowledge is collected, stored, and circulated. The flow of information has been increasing at almost exponential levels. Today, data sets are measured in zettabytes (10^{21} bytes). Cost-effectively collected and stored data allows researchers across the world to successfully advance understanding of science and medicine [45].

International Data Corporation (IDC) is one of the premier global providers of market intelligence, information technology, and a host of other areas. They predicted in a report issued in Dec 2018 that the world's cumulative data will grow from 33 zettabytes to a 175ZB by 2025, for a compounded annual growth rate of 61%. A zettabyte is a trillion gigabytes multiplied that by 175 times. This growth of data has been seen in every industry, in every corner of the world. The relentless increase in the quantity and flood of information denotes an important professional opportunity, but a challenge simultaneously for those in medicine and science [46].

As indicated by Toby Cosgrove MD, from Cleveland, medical information has doubled within every 73 days, in the year 2020, which used to take approximately 3.5 years in 2010. An estimated 8,00,000 papers are published in 5,600 medical journals every year. It is projected that 12,000 new articles and 300 randomized controlled trials are added to Medline each week, and that new medical articles appear at a rate of one every 26 seconds [47]. To be able to generate any kind of analysis and make accurate predictions, there is a need to access, connect various sources, collate, and consume all the data. Data can be produced, obtained, and stored in a numerous number of structures [48].

During an appointment at a hospital, diverse types of data are collected. Raw data are observations about individual patients created by the treating doctor at a hospital. These data may be in the form of measurements of patient's characteristics such as age, gender, height, weight, blood pressure, heart rate, etc. Raw data may also include description of the medical history, physical exam information, clinical laboratory results (e.g., serum lipid values, hemoglobin levels), whole exome or genome sequences, imaging results, ECGs, questionnaire data, or self-reported data (e.g., symptoms, quality of life).

Raw data, unprocessed source data, like unrefined gold buried deep in a mine is a precious resource. It is often: (1) Inconsistent, containing both relevant and irrelevant data, (2) Imprecise, containing incorrectly entered information or missing values, (3) Repetitive, containing duplicate data. To utilize the raw data to its fullest potential, it needs to be extracted, filtered through, understood, and transformed into analyzable format. One of the surveys carried out by Forbes estimates that data cleaning accounts for up to 80% of the development time and cost in data warehousing projects. Understanding the scope of data being analyzed and seeing the changes

made to the data can accelerate the entire process of going from “information to building wisdom” [49] [50] [51] [52].

1.13 Introduction to Clinical data understanding

The health care industry uses either a paper-based record keeping method and/or electronic health record (EHR) system to manage patient data. More and more organizations are using electronic data capture, but the practicing doctors in individual clinics may still be documenting observations on paper. The EHR has become an integral part of medical care, which transforms health care service quality and improves physicians’ satisfaction and facilitates patients’ decision. Accurate information from EHR enables physicians’ decision making and measures clinical validity, which in turn upgrades the quality of patient care. This functionality is crucial during diagnosis and therapy, which benefits medical and legal practices too [48].

Health authorities and top-level journals require the data to be submitted along with research papers. The processes outlined in the above sections provide the necessary “fit for purpose” data. The analyzable data set, is the result of many decisions made by varied people, as explained above. The errors, flaws, or biases in the processing of source data, will not necessarily be identified in the analyzable dataset. After the electronic data entry, new variables are generated to support further analysis. The final cleaned analyzable datasets consist of various components such as participant characteristics and primary outcome, pre-specified secondary and tertiary outcomes, adverse event data and exploratory data [53].

Physicians and other scientists are getting better at producing data. But we must become proficient—with or without the help of technology—at mining and managing the data in ways that will allow us to use it to maximum effect [54]. The full analyzable dataset is generally the most useful set of data to share, with large and likely important benefits to science and society. Secondly, the full analyzable dataset provides scientific validity to the outcome and ensures replication and repeatability. Further, meta-analysis increases the statistical power of detecting effects and maximizes the value of the outcome in the clinical knowledge base. Finally, analyzable data allows for further scientific discovery through additional secondary analyses, as well as the conduct of exploratory research to generate hypotheses for additional studies [55].

1.14 Introduction to study of demographics and patient characteristics

We have seen how the data has been collected and then converted into analyzable formats. Now, we will proceed with understanding the data contents and see if any of the questions raised earlier can be solved. Demography is the study of the population. It explains the composition, the distribution and the data trends seen in the population. Roles and functions for demography studies can be broadly defined as, (1) population projections, (2) inputs into government budget, (3) evidence-based policy, and (4) communication of vital statistics [56]. There is very little data on the profile of patients accessing traditional systems of medicine. A comparative study of profile of patients using an Ayurveda clinic and modern medicinal clinic will help in understanding of utilization of services and preference for health seeking behaviour.

1.15 Introduction to study of diagnostics and interventions

Diagnosis is one of the most important aspect in the process of treatment of the disease or condition. It is a patient-centered, cyclic process of gathering information, analyzing information, determining the health condition, and defining the type of intervention and continuously monitoring the progress till the desired state of functions/doshas is arrived at. Ayurvedic treatment involves removal of the causative factors. It assists in getting the functions/doshas into balance. The success of a treatment is possible only by timely and accurate diagnosis, a tailor-made intervention accompanied by an effective collaboration of the physician and the patient [57]. The term comorbidity refers to the coexistence of multiple diseases in relation to a primary disease in a patient. Patients report multiple diseases during their visits to the hospital. Some of these reported disorders are expected and some are unexpected. There are known as well as unknown disease combinations present due to biological linkages. Clinical and epidemiological studies indicate that disease comorbidities have a great impact on health status, selection of appropriate treatments and health system costs. Understanding comorbidities and their etiology is key to identify new preventive and therapeutic strategies.

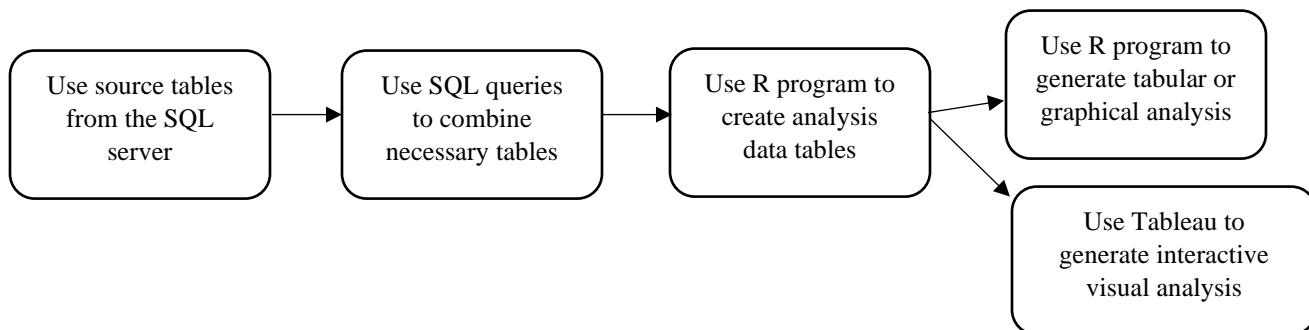
Finally, all these steps (1) accessing and understanding real life data, (2) converting that data into analyzable format, (3) understanding demographics and patient characteristics, and (4) understanding diagnostics and interventions should help in building transdisciplinary evidence to increase the scientific understanding outside of the community, then increase the confidence and thereby widening the user base.

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 >51,000 patients, >1,50,000 visits, > 900 variations of disease types, >3,000 variations of medical procedures. 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 [58] and R programs [59], python [60], Java [61], D3js [62] and tableau [63] 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 [64], (2) descriptive summary statistics [64], (3) data representation on world / country map [65], (4) boxplot representation [66], (5) barplot [67] and (6) dotplot representation [67], (7) radar plot representation [68], (8) individual patient level data listings – line by line data representation, (9) various types of bubble plots [67], [69], (10) circular data representation [70], (11) collapsible tree diagram [71], (12) mosaic plot [72], (13) butterfly plot [73], (14) area plot [67], (15) calendar plot [67]. 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.

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 PostgreSQL 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.

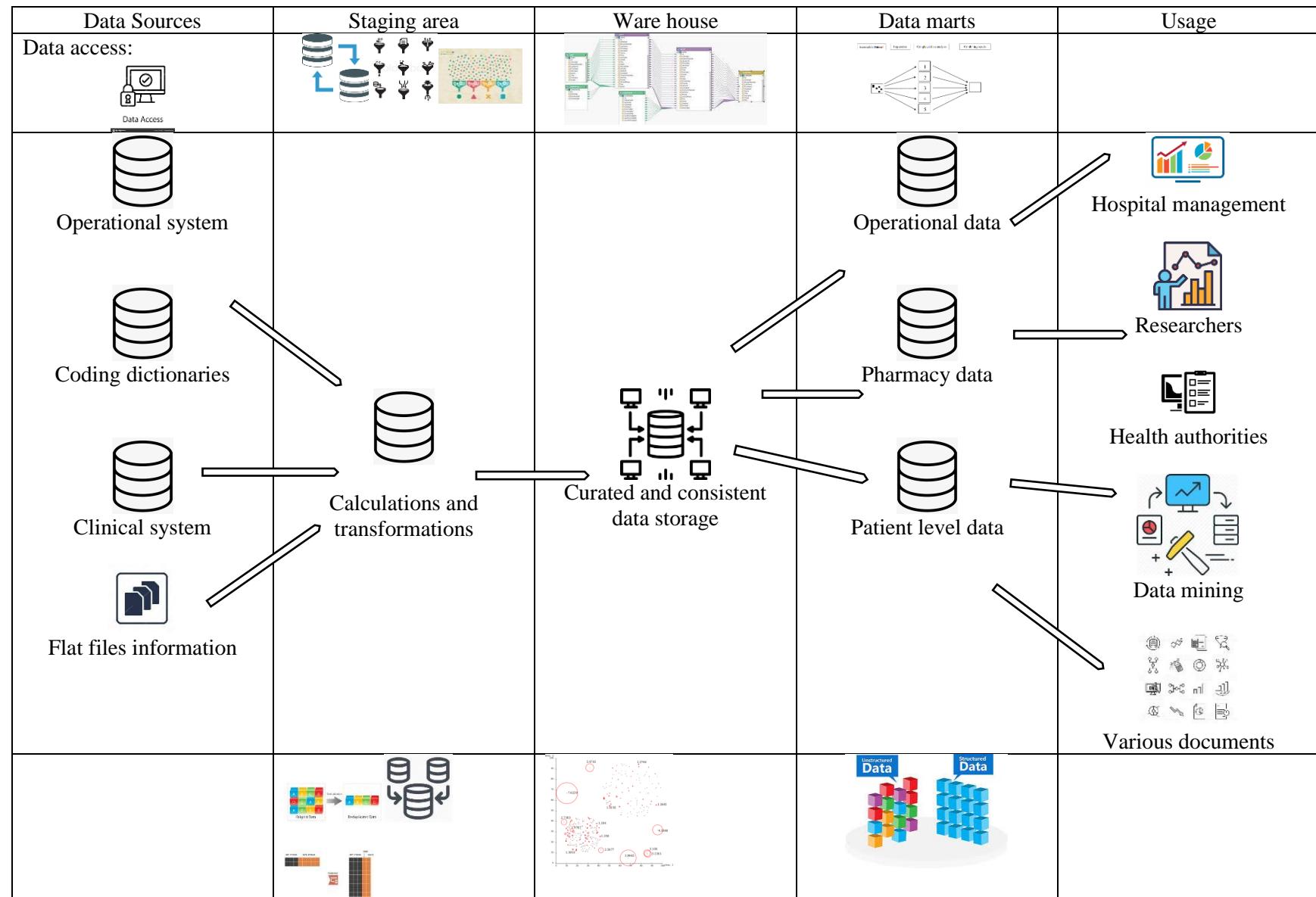
2.3.2 Data preparation

It is very important to think through the data preparation stage about the data holistically. Audience is critical while preparing data. Who will use the data and where and when, for what purpose determines how the data would 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 detail influences the data preparation process significantly, determining both the amount of effort and detail [50]. The following steps were followed in data preparation (Figure 2-2), the details can be found in the individual R, SQL, python programs:

- 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 resultant data [52].
- Appending: Combine two or more similar datasets into a single dataset [52].
- 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 [50] [52].
- 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 examines incoming data and compares it to data that is already stored in the system [50] [52].
- Transforming: This involves converting data from one structure (or no structure) to another to integrate it with a data warehouse or with different applications [50] [52].
- 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 [52].
- 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. [52].

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



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, 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. Let us go through the challenges experienced to assemble the “reference dataset” from the source data and practical explanation of the “data preparation” steps.

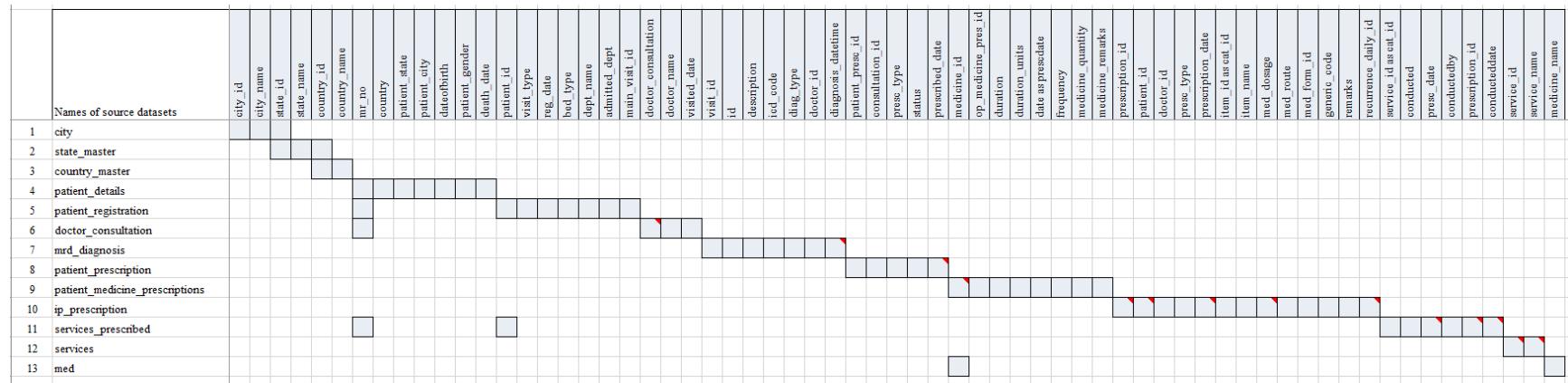
1. The database was manually explored using various SQL programming commands, the variables and observations were checked from numerous tables (Figure 2-3)
2. Patient information and key variables needed to be understood: unique patient ID is MR_NO, and unique ID for individual visit is PATIENT_NO (many 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 ~65 variables to generate the necessary snapshot of the source data. These were rearranged into 6 datasets and ~40 variables. 3 additional reference files were used for further processing. 1 final dataset having ~30 variables from source and ~30 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_origin	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	

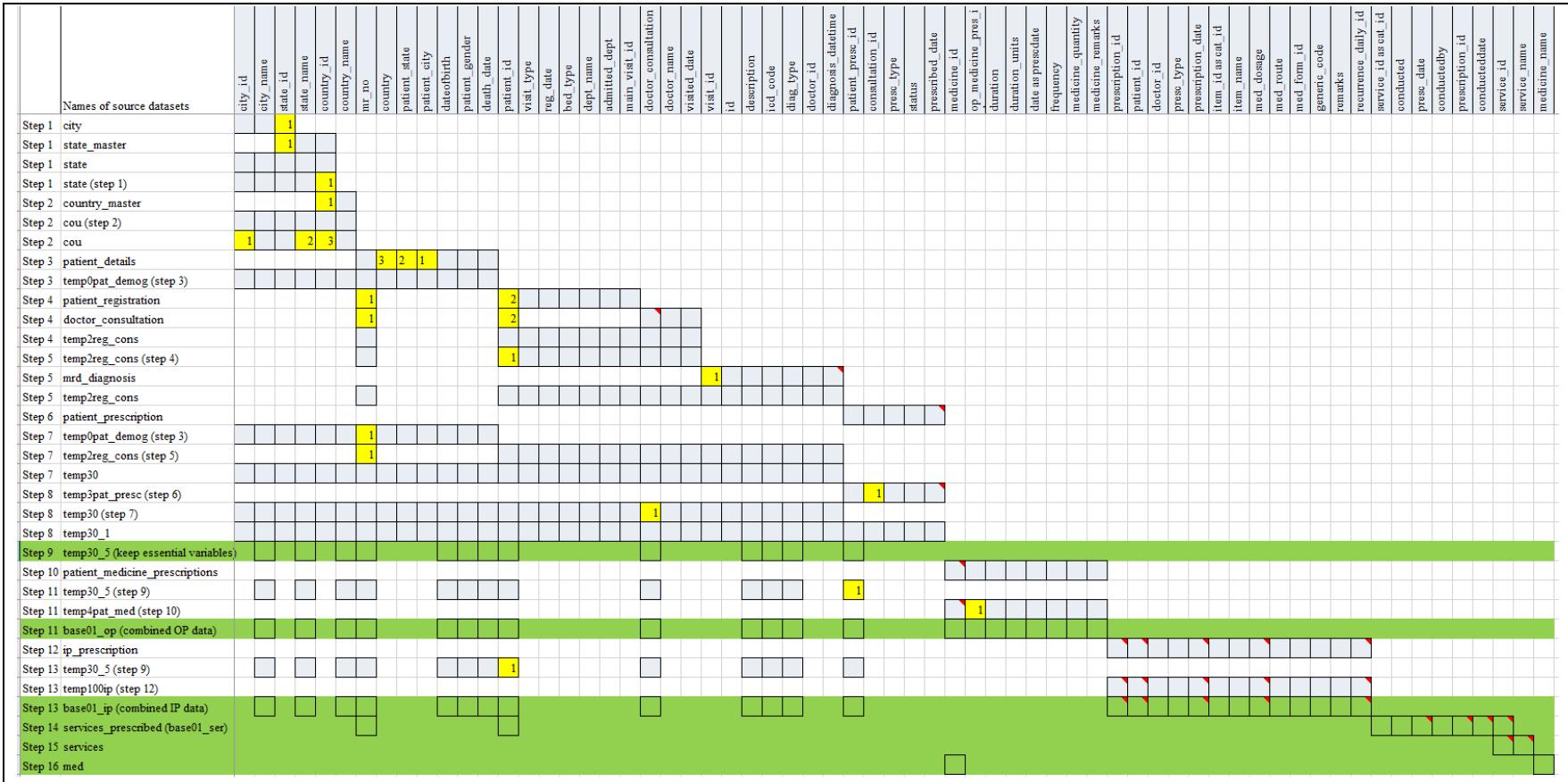
This table presents a glimpse of inventory of data tables in database. The cells marked in yellow are used for the generation of the analysis ready datasets.

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



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 ~65 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

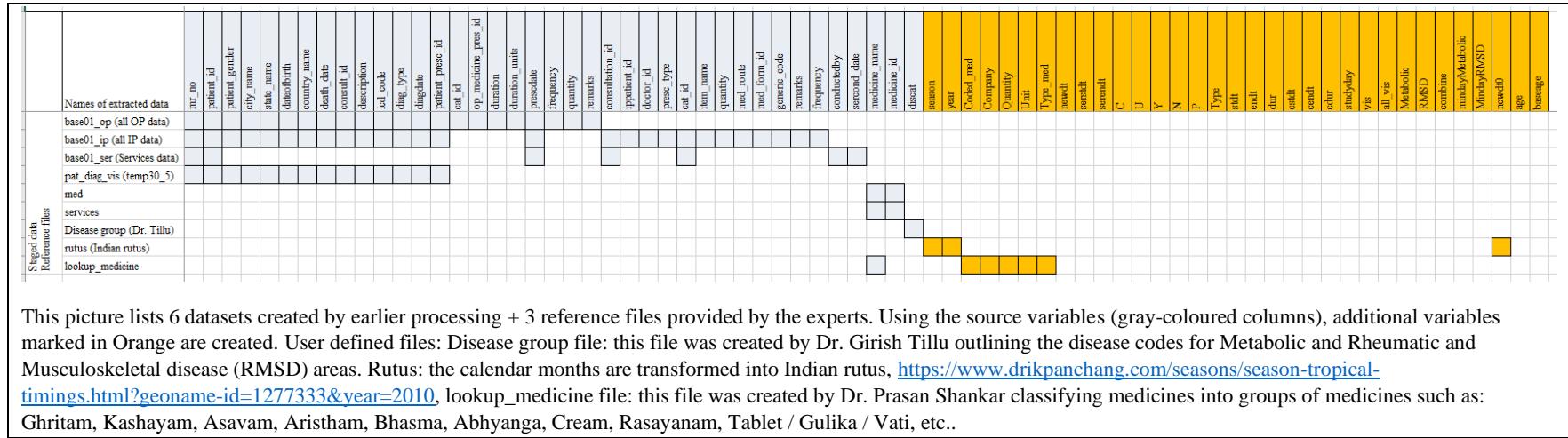
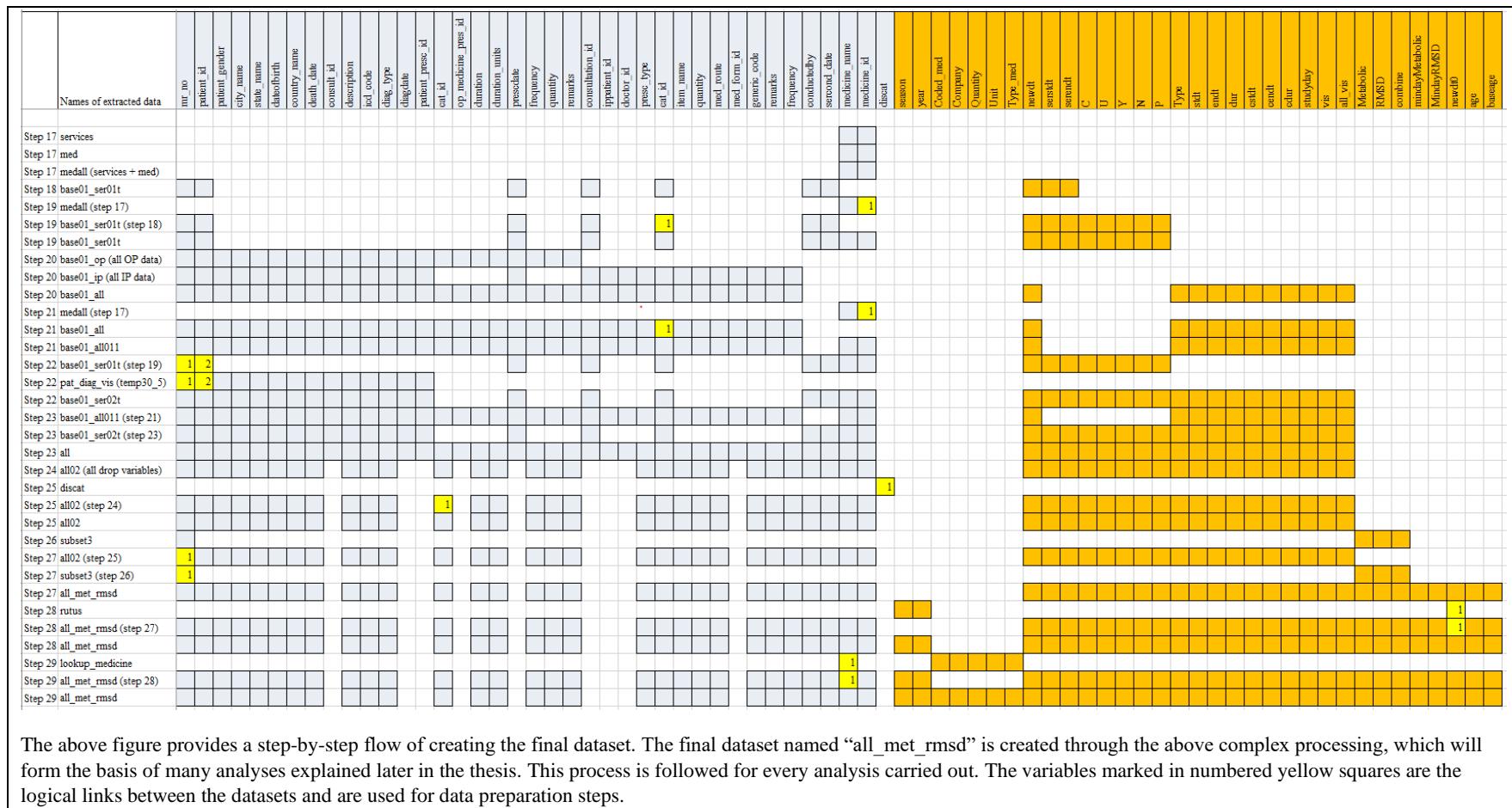
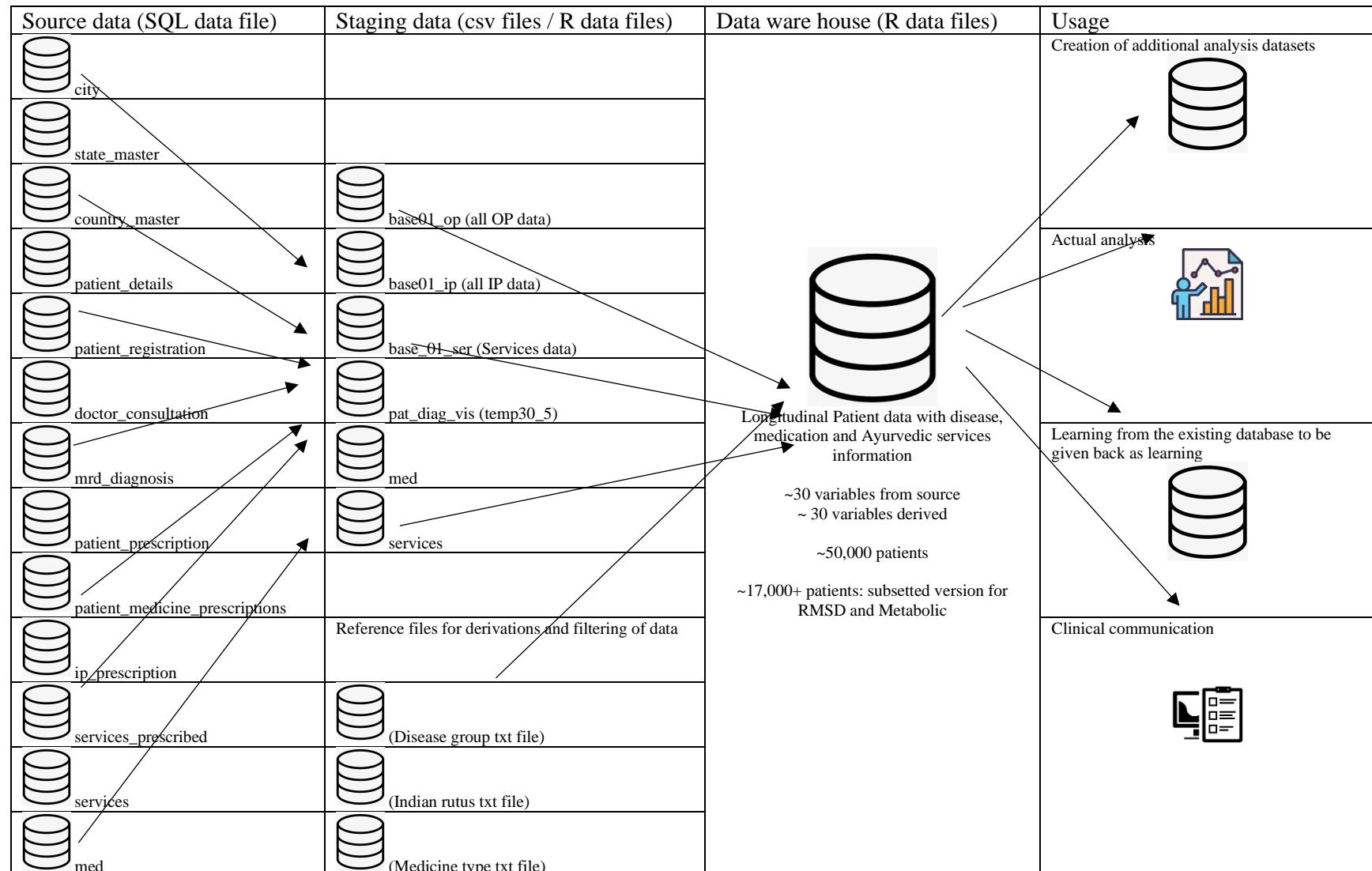


Figure 2-7: Final dataset with ~30 source variables and ~30 new derived variables



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



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, duplication, etc. across 90+ other 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. In our case, we have seen 1 example where ~50 stages to create analyzable datasets as well as reference datasets from the pure source database.

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. 2 longitudinal interactive views were created to display individual patient data. 1st 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.

2nd 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) Daysstt: 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 needed 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 had 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 states 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 [74] (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 1st 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. Algorithm 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) ACD and ICD 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, 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 3-34, 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 (Figure 3-37, Figure 3-38), (10) Distance metrics analysis for disease trajectories and medicine trajectories (Figure 3-39, Figure 3-40), and (11) Multi-dimensional data representation using radar plot displays (Figure 3-41), (12) Dynamic bubble plot visualization (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: 1st approach produced a bubble plot, boxplots, summary statistics tables and frequency tables for number of other diseases reported along with the primary disease. 2nd approach recreated the same analysis by visit window of each month, 3rd approach created disease trajectories visually displayed in a form of collapsible tree (Figure 3-22, Figure 3-23, Figure 3-24, Figure 3-25, Figure 3-26).

Algorithm for 1st 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. (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, bubble size was determined by the count of unique patient ID. For each disease how many other unique diseases were reported is calculated by gender. 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, could have occurred before or after the occurrence of reference disease). The tooltip shows min, median and max age, distinct counts of patients. A small table on the left side shows number of other diseases experienced (Figure 3-22, Figure 3-23, Figure 3-24)

Algorithm for 2nd approach: Same calculations for 1st algorithm were followed to create co-morbidities, 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 3rd approach: (1) Diseases experienced by each patient were sorted by date and only 1st instance of a disease was retained. (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 stored in Json file. Json file was used 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 “1st time disease reported”, any subsequent repetition was counted as “Repeat”. (2) When a treatment was prescribed very first time then that was counted “1st time treatment prescribed”, any subsequent repetition was counted as “Repeat”. (3) These 2 calculations were repeated throughout the complete duration for each patient.

Area graph representation of diseases was created to show variations related day-to-day, seasons, gender, and diseases. This analysis provides frequency count of patients for each disease by month 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 are 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 would be performed from an intervention's perspective. (1) TreeMapDisMed-Parameter sheet is used to filter a particular intervention, which is 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 is prescribed, (4) Medicine-list sheet shows a detailed list of diseases with the total patients prescribed with the intervention as well as total patients suffering from different diseases (Figure 3-34).

Cross tabulation of prescribed treatments and disease group by gender was generated. The interactive visualization was used to create a few examples for specific disease conditions or specific treatment. 1st example is created using Balaristham and 2nd 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.

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 small 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 the mathematical distances. There are numerous distance measures available in mathematics and statistics which allows in understanding the similarity and dis-similarity between objects [75].

Following assumptions were used to derive the disease trajectory: (1) Diseases experienced by each patient were sorted by date and only 1st 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. (4) The similarity measure was calculated for each disease trajectory, e.g., Jaccard distance was used as a distance measure for this display. (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. 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. 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 1st disease or 3rd disease or 5th 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 small representations of same kind next to each other.

Dynamic bubble plot visualization: explanation of an algorithm using Amavaata (A6.0) as an example: (1) Identified unique patients who have had Amavaata reported at least once, (2) Got all the other diseases and prescribed medicines for this subset of patients, (3) Created an input Json file to be passed into a D3 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 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/]

3 Results

3.1 Converting real life clinical data into analyzable format

3.1.1 Details of the database

The database had approximately 200 datasets (Figure 2-3). 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:
 - a. Hospital charges – IP, OP
 - b. Operation theater charges
 - c. Inventory of equipment
 - d. Doctor charges, etc.
2. Reference dictionaries
 - a. Disease codes
 - b. Ayurvedic services
 - c. Medication names
 - d. Master list of Laboratory tests
 - e. Names of city, state, countries
3. Doctor details
 - a. Doctor ID
 - b. Relevant ward information
 - c. Internal / Visiting / Part time / Full time
4. Patient information
 - a. Patient details,
 - b. Visit details
 - c. Vital signs
 - d. Registration details
 - e. Discharge details
 - f. Lab data details
 - g. 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.

3.1.2 Data Extracted from Hospital Database

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

Table 3-1: Versions of data used for analysis

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

3.2 Clinical data understanding

3.2.1 Broad checks on the datasets

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 allowed 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 had to be derived outside of the database using expert understanding which would require investment of time and efforts from Ayurvedic vaidyas. There were 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 increased 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 was missing, or partially filled. There were 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 were similar to the Vital signs database. Along with the patient identifier information, only laboratory test name and laboratory measurements were present. In case a patient had the laboratory investigations outside of the hospital that data got stored in a scanned image format. Apart from this the dataset did not contain the date of sample, reference ranges, laboratory parameter units, fasting status etc. A single lab test had multiple names. E.g., Alanine Aminotransferase was 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 was 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 was prescribed for which disease was 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 was found in more than 1 variable. Different acronyms were used inconsistently. Answer for more than 1 question was captured in 1 variable. Due to “free text nature” of variables simple questions like Yes / No had many different data values.

Classification and Sub-classification of the Doshas / Diseases: it was 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.

Figure 3-1: A snippet of disease table by gender

DisType_Diseases

Distype	Code	Description	Patient Gender		Grand Total
			F	M	
Metabolic	P5.0	Prameha	648	849	1,497
	P5.3	Prameha - Sthula	23	20	43
	P5.2	Prameha - Pidaka	5	12	17
	P5.1	Prameha - Krusha	4	5	9
Grand Total			671	876	1,547

DisType_Diseases

Distype	Code	Description	Patient Gender		Grand Total
			F	M	
RMSD	A2.0	Aamavaata	652	180	832
	A2.3	Aamavaata - Vaataja	27	7	34
	A2.1	Aamavaata - Kaphaja	13	4	17
	A2.2	Aamavaata - Pittaja	4	3	7
	Grand Total		679	190	869

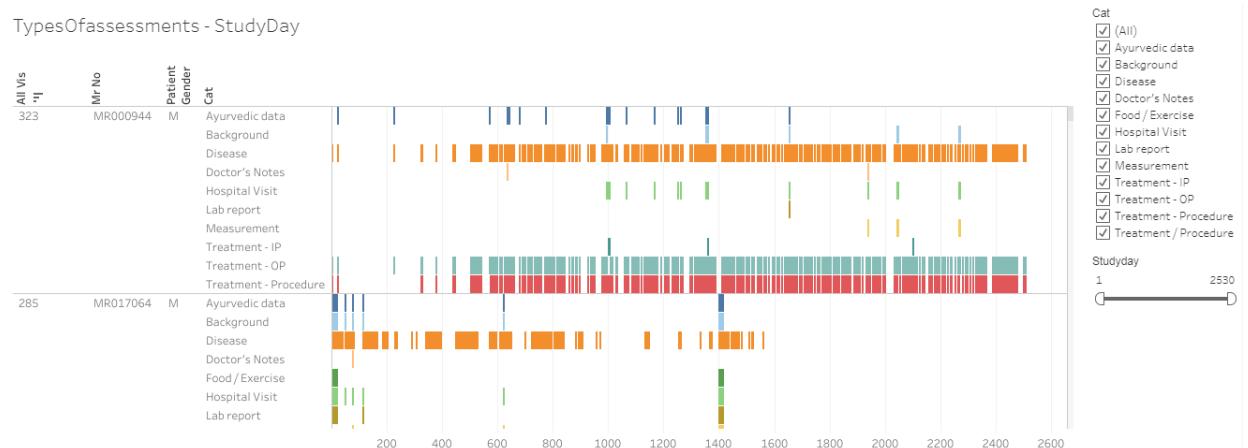
Frequency of Prameha (P5.0) and Aamavata by gender has been displayed. The frequency of patients is substantially lower for classification by Sthula, Pidaka and Krusha as well as Kapha, Vaata, and Pitta. Data version: 2011 to Oct 2017

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

3.2.2 Contents checks

The analysis carried out showed that for majority of the patients and for majority of the visits, the disease data and medication (Treatment /Procedure) were non-missing. Most of the other categories were not entered as consistently as they should have been, this was a basic interpretation generated. If this was the expected data collection pattern then these findings should not be considered.

Figure 3-2: Variable classification by categories



500+ variables are captured for each visit and each patient are classified 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 is any non-missing data present in a particular category then "Yes" is assigned, if the data is missing then "No" value is assigned. This data is presented as a listing for each patient for each visit (day). When the data is available it is presented as a color-coded bar and when it is missing then it is presented as a white blank space. Data version: 2011 to Oct 2017 https://public.tableau.com/views/03_typesOfassessment/TypesOfassessments- StudyDay?:language=en&:display_count=y&:origin=viz_share_link

2 example screenshots are shown as what was observed while content check was performed.

CRFname_variable number_variable label	Unique values	Unique patients	Variable classification
sec001_var008_Diabetes	1064	4124	Background

sec001_var008_Diabetes: represents CRF page number 1, labelled as “History of Present Illness” and variable number 8 “Diabetes”. This variable “Diabetes” has 1064 unique values entered by different doctors for different patients. Data is available for 4124 distinct patients.

CRFname_variable number_variable label	Unique values	Unique patients	Variable classification
sec001_var018_Associated Complaint with Onset & Duration	5102	4549	Disease

sec001_var018_Associated Complaint with Onset & Duration: presents CRF page 1 “History of Present Illness” and variable number 8 “Associated Complaint with Onset & Duration”. This variable “Associated Complaint with Onset & Duration” has 5102 different values entered by different doctors for different patients. Data is available for 4549 distinct patients.

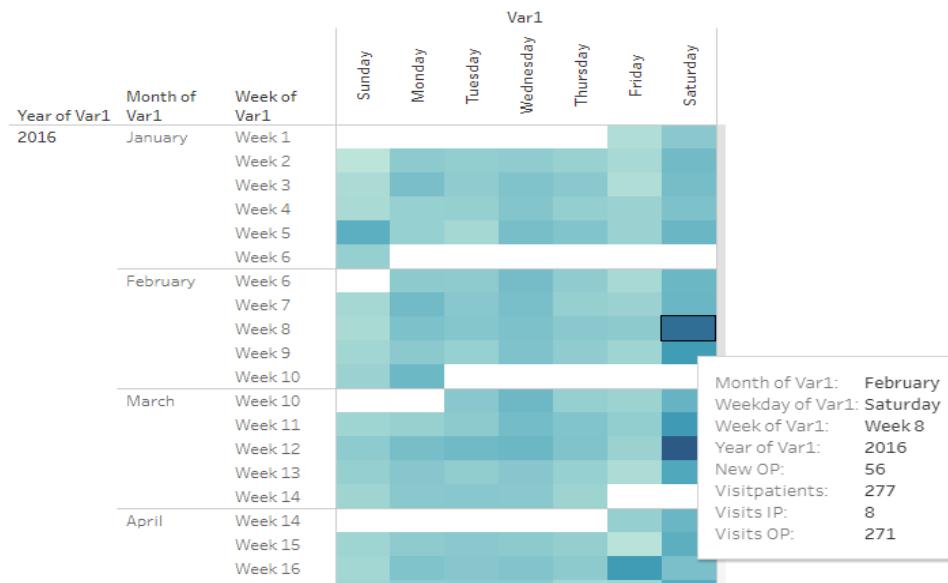
The large number of unique values show that the data entry rules are not consistent, and each doctor has different interpretation. Looking at the unique number of patients, the data has not been entered for all the patients, giving rise to missing data.

3.2.3 Visit pattern analysis

Each of the cell displayed on a calendar display was coloured in shades of blue from light blue to dark blue showing increasing frequency count of number of patients. From 2011 to 2016, the number of patients visiting hospital on weekdays was less than the number of patients visiting on weekends. In-Patients were considerably less than Out-Patients. Overall number of patients coming to hospital have been increasing year on year. This information would help in employing staff across different departments from helpers, cleaners, nurses to doctors to adequately cover services for patients.

Figure 3-3: Visit pattern analysis

Sheet 1



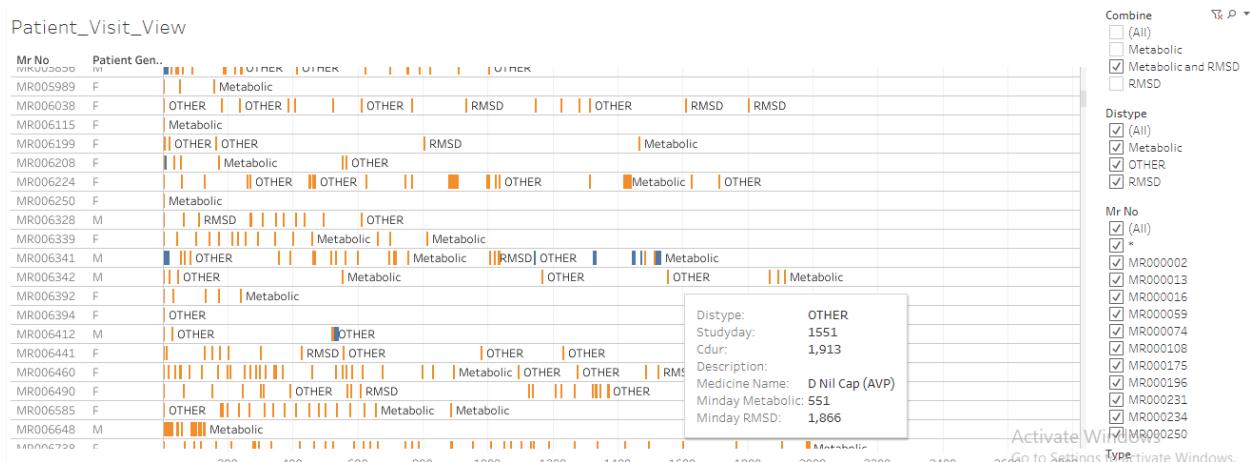
Frequency counts of 4 parameters, (1) new Out-Patients added on that day, (2) total number of patients visiting on that day, (3) total number of In-Patient visits on that day, and (4) total number of Out-Patient visits on that day are calculated for each day to understand the patient flow to hospital from year on year. Light blue to dark blue shows increasing frequency count of number of patients Data version: 2011 to Oct 2016

https://public.tableau.com/views/04_calendar_view/Sheet1?:language=en&:display_count=y&:origin=viz_share_link

3.2.4 Patient disease and treatment journey view

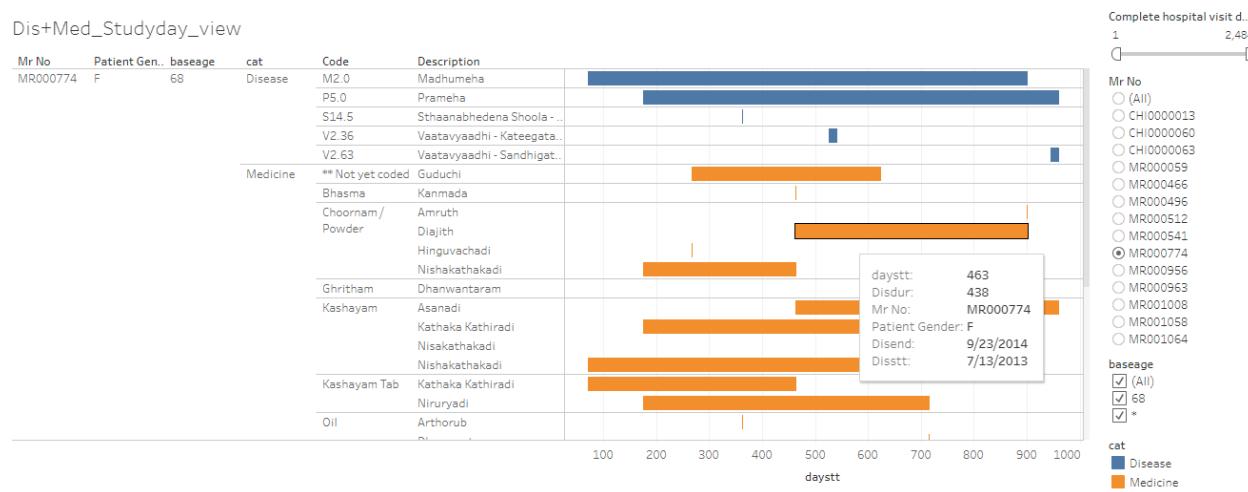
Multiple representations of data allow the end user to review data with different perspectives. Patient profile reports provide detailed view of individual patient's disease condition, prescribed medication, co-morbidities along with basic demographic information. Treating doctors and researchers will greatly benefit from this visual display. This representation provides the patient an understanding of the disease chronology as well as the prescribed medication and progress. Clear representation of co-morbidities and prescribed medicines is a simple intuitive way of sharing very important information with concerned parties involved. The horizontal view provides viewpoints across patients and vertical view gives in-depth viewpoints within a patient.

Figure 3-4: Patient visit profile – Horizontal view



Mr No: Patient ID, Patient gender, x-axis: duration of hospital visits, Orange bar: Out-patient visit, Blue bar: In-patient visit, Metabolic: when a metabolic disease is reported, RMSD: when a Rheumatic and Musculoskeletal disease is reported, OTHER: other diseases are reported, Tooltip has a lot of information relevant to each visit. Data version: 2011 to Oct 2017
https://public.tableau.com/views/01SQL_Dis_Med_Ser/Patient_Visit_View?:language=en&:display_count=y&:origin=viz_share_link

Figure 3-5: Patient visit profile – Vertical view



Mr No: Patient ID, Patient gender, baseage Age at the very first hospital visit, category: Disease and medicine, Code: ACD code, Description: disease description, x-axis: duration of disease and medicine, Tooltip has a lot of information relevant to each visit. Data version: 2011 to Oct 2017

https://public.tableau.com/views/01_Primary_madhumeha/DisMed_Studyday_view?:language=en&:display_count=y&:origin=viz_share_link

Observations related to patient profile report generation module which is used by the treating doctors: The following section outlines an improved version of patient profile:

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 can 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

3.3 Studying demographics and patient specific factors

Results related to complete set of patients: While exploring the basic data, the following high-level picture appeared: For the 5-year time frame from 2011 to 2016, the database contained ~40,000 unique patients (Figure 3-6), 90% of patients were from India and remaining 10% patients were from 50+ different countries (Figure 3-7). The proportion of male and female patient was ~50%. Median age for females was marginally higher than males across all visit types (Figure 3-8). Approximately 90% of patients were Out-Patients and 10% were In-Patients (Figure 3-1). Approximately 12,000+ female patients and 14,000+ male patients had reported only a single disease (Figure 3-11), these patients could have come only once to the hospital and may not have come back at all after reporting the 1st disease. There were a few outliers observed having more than 10 disease conditions across the years. The maximum age of 108 years was a possible case of data issue. Similar anomalies were 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 was missing data for almost 20% of patients. Blood group distribution was largely in line with the Indian blood group distribution (Figure 3-9).

Figure 3-6: Total Number of Patients

01 NumberOfPatients	02 Country	03 AgeBoxplotCountry	04 Age
---------------------	------------	----------------------	--------

Summary of distinct number of patients

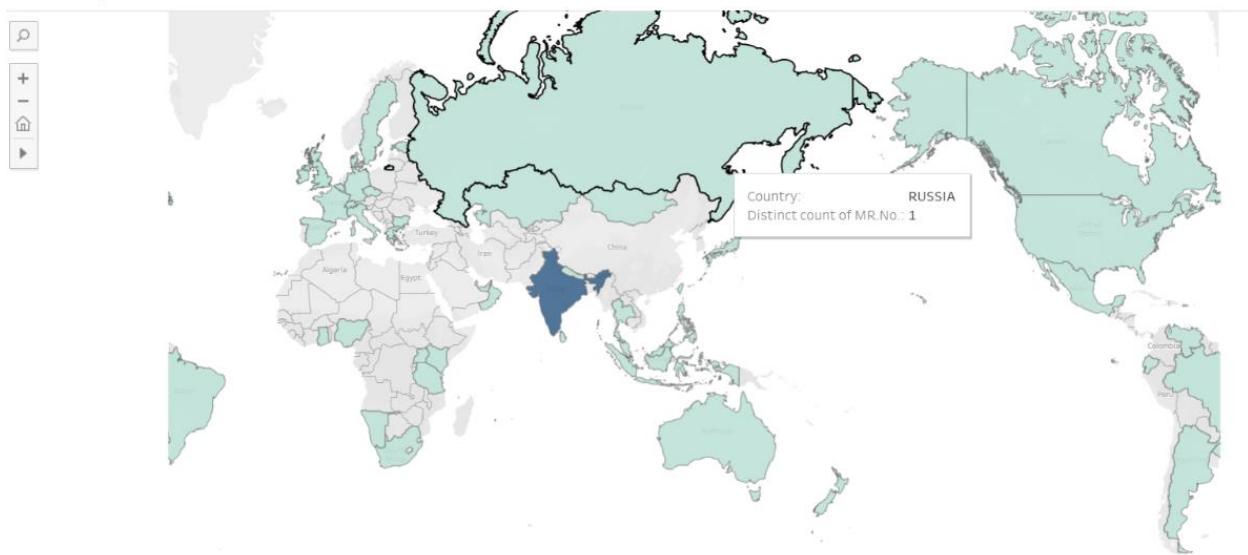
Grp				
Common	NA	OnlyIP	OnlyOP	Grand Total
3,491	0	4,145	38,903	39,557

Grad Total: Total number of patients used in the analysis, OnlyIP: patients having only In-Patient visits, OnlyOP: patients having only Out-Patient visits, Common: Patients having both type of visits. Data version: 2011 to Oct 2016

https://public.tableau.com/views/04_patient_analysis_tableau/01NumberOfPatients?:language=en&:display_count=y&:origin=viz_share_link

Figure 3-7: Country-wise Visualization

02 Country

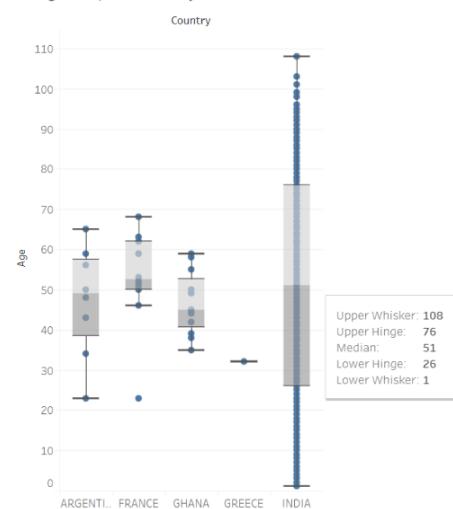


Unique number of patients are plotted on the world map, the map shows that at least 1 patient data is coming from 50+ countries, 95% or more patients are from India. Data version: 2011 to Oct 2016

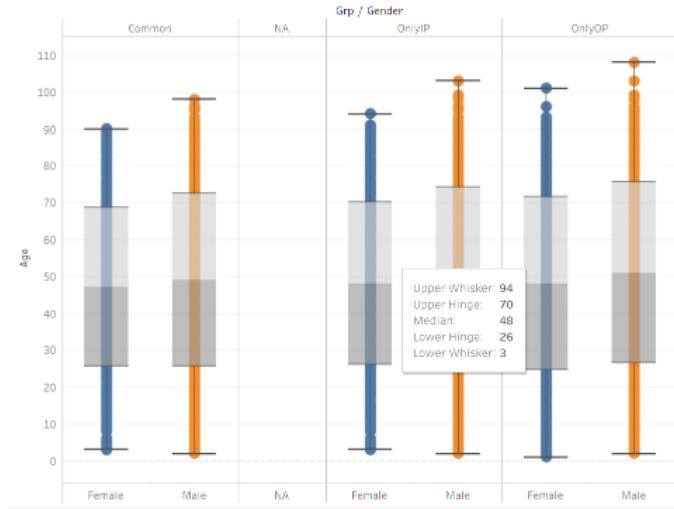
https://public.tableau.com/views/04_patient_analysis_tableau/02Country?:language=en&:display_count=y&:origin=viz_share_link

Figure 3-8: Age distribution by country, age distribution by gender

03 AgeBoxplotCountry



04 AgeBoxplotGroup



Boxplot representation: Age distribution is presented for each country and then by type of patient and gender, OnlyIP: patients having only In-Patient visits, OnlyOP: patients having only Out-Patient visits, Common: Patients having both type of visits Data version: 2011 to Oct 2016

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

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

Figure 3-9: Blood-group Distribution by gender

07 BloodGroup

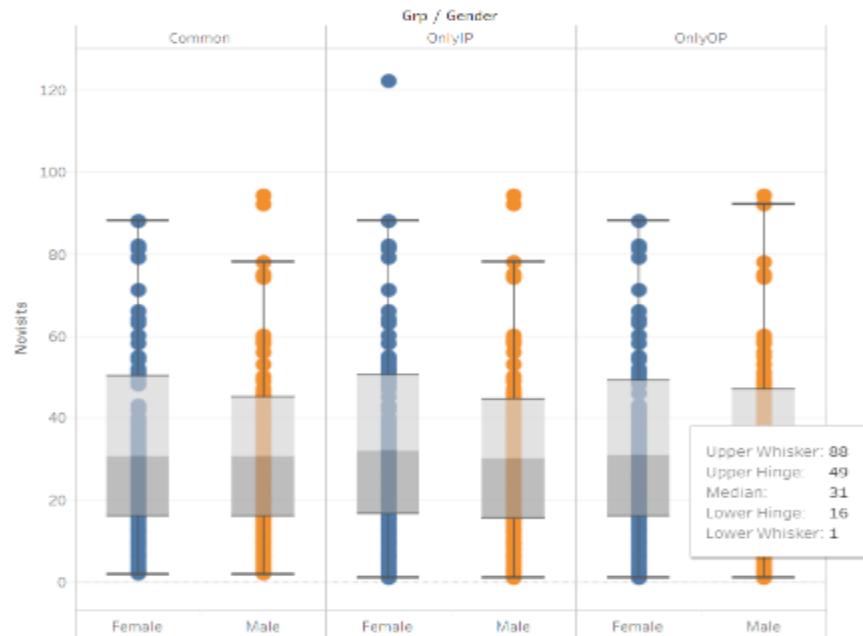
Blood...	Gender			Grand Total
	Female	Male	NA	
O+	6,340	6,684		13,024
B+	4,490	4,956		9,446
A+	2,997	3,030		6,027
AB+	828	1,046		1,874
O-	434	447		881
B-	265	259		524
A-	239	178		417
AB-	80	87		167
A1 +ve	30	17		47
A1 -ve	6	1		7
A1B +ve	3	4		7
A2 +ve	1			1
A2B +ve		1		1
NA			1	1
Null	0	0		0
Grand Total	15,713	16,710	1	32,424

Tabular frequency distribution table for Blood-group by gender. Data version: 2011 to Oct 2016

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

Figure 3-10: Number of Visits, and Visit Types

05 NoOfVisitsBox



Boxplot representation of number of Visits, and Visit Types, Data version: 2011 to Oct 2016

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

Figure 3-11: Descriptive summary statistics by number of Diseases by Age and Gender

05a NoOfDis_age

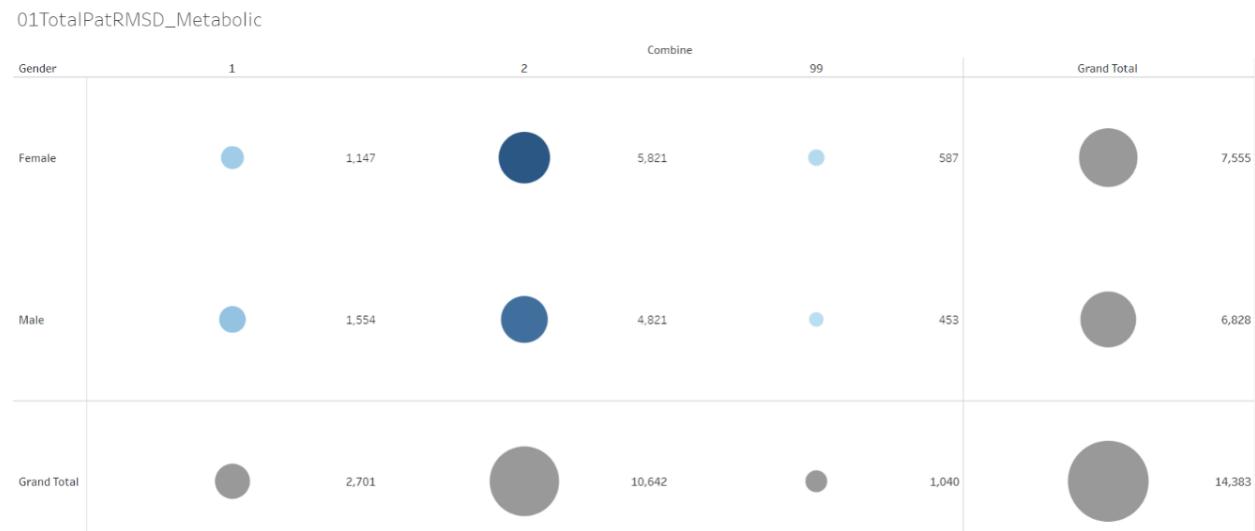
Nodiseases	Female						Male					
	Count of Age	Avg. Age	Median Age	Std. dev. of Age	Min. Age	Max. Age	Count of Age	Avg. Age	Median Age	Std. dev. of Age	Min. Age	Max. Age
1	12,884	46	46	18	1	96	14,375	46	45	19	2	108
2	5,652	48	50	18	2	93	5,420	48	47	19	2	96
3	2,426	48	48	17	2	93	2,002	50	49	18	2	98
4	1,215	51	51	17	2	87	898	53	55	19	2	93
5	526	50	50	15	9	84	477	53	54	18	2	87
6	342	48	48	17	9	81	199	50	47	15	6	84
7	182	53	52	13	12	79	140	57	56	15	33	88
8	99	48	48	15	13	80	28	46	52	23	6	76
9	77	54	52	15	4	78	22	58	65	11	39	67
10	25	53	51	21	33	86	20	50	51	5	43	57
11	23	47	49	17	22	66	14	48	44	14	35	63
12	31	44	39	24	9	82	6	64	64	0	64	64
15	2	56	56	21	41	71						
NA	4,263	46	45	19	2	101	5,627	43	41	19	2	103
13	8	53	56	9	30	56	13	51	47	8	31	58
23							6	41	41	0	41	41
14	9	30	18	20	18	71	1	35	35		35	35

Descriptive summary statistics by number of diseases reported, by gender. Noofdisases: Number of diseases reported in the database. Data version: 2011 to Oct 2016

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

Results for the metabolic and RMSD disease areas: Out of ~40,000 patients, there were ~14,000 patients having reported at least 1 metabolic and/or 1 RMSD disease condition. It was quite evident that there were a lot more patients in the RMSD group compared to the metabolic group (Figure 3-12). Large number of patients were visiting the hospital only for 1 visit, ~62% patients were dropping off in first month of treatment. ~15% of patients having at least one RMSD disease were still visiting the hospital after 1 year of first ever visit to the hospital (Fig 4.3.9). Boxplot representation of age showed variability in age across disease type and gender (Figure 3-13). Presentation of disease burden by gender, Indian seasons (ratus) and disease category provides data about possible variations reported for different diseases (Figure 3-15). (1) Prameha, (2) Madhumeha, and (3) Sthaulya were 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 were the top five most frequently reported RMSD diseases. Prameha and Madhumeha were reported more by males than females. There were more female patients with disease condition Sthaulya. In general, RMSD diseases were reported in more females than males. For RMSD disease group, 51 out of 97 diseases were reported in <= 10 patients. Metabolic diseases were not varying across seasons, while RMSD diseases had some seasonal variations (Figure 3-15). The before and after visualization of data allows to build a disease and medicinal trajectories (Figure 3-16). These should be useful for determining diagnostic and prognostic relationships.

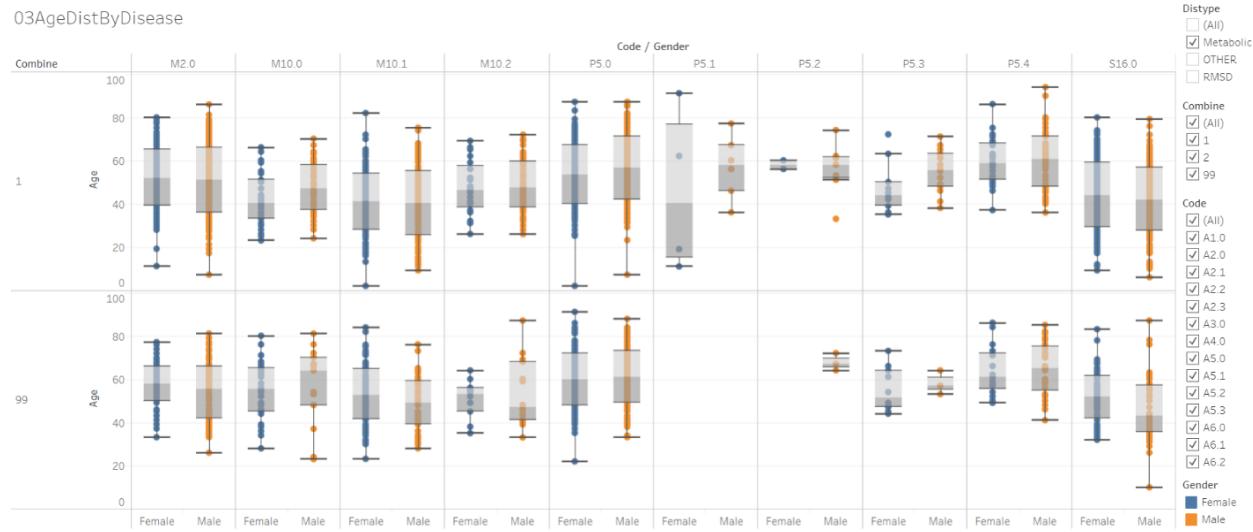
Figure 3-12: Data tabulation for patients reporting RMSD and Metabolic diseases



Bubble plot: 1 = Patients with at least 1 metabolic diseases, 2 = Patients with at least 1 Rheumatic, Musculoskeletal (RMSD) diseases, 99 = Patients with at least one disease from each of the groups, Data version: 2011 to Oct 2016

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

Figure 3-13: Disease distribution by age and gender



Boxplot representation of age by disease. Orange: male, Blue: female. Individual column represents a disease. 1 = Patients with at least 1 metabolic diseases, 2 = Patients with at least 1 Rheumatic, Musculoskeletal (RMSD) diseases, 99 = Patients with at least one disease from each of the groups, 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)

Figure 3-14: Patient visit duration for Disease categories by Gender

		Combine / Gender								Grand Total	
Durwr	Dur	Female	1	Male	Female	2	Male	Female	99	Male	
Null	Null		42	49		275	228		53	41	688
0	>=1 day		1,134	1,537		5,751	4,763		582	447	14,214
1	>=1 month		425	590		2,255	1,634		354	267	5,525
2	>=2 months		334	480		1,854	1,374		315	239	4,596
3	>=3 months		294	433		1,650	1,227		292	225	4,121
6	>=6 months		223	322		1,257	944		249	196	3,191
12	>=1 year		148	203		857	665		184	147	2,204
24	>=2 years		81	103		444	335		102	95	1,160
36	>=3 years		38	48		215	159		55	51	566
48	>=4 years		17	16		62	47		24	19	185
60	>=5 years		4	2		18	11		10	6	51
Grand Total			1,147	1,554		5,821	4,821		587	453	14,383

1 = Patients with at least 1 metabolic diseases, 2 = Patients with at least 1 Rheumatic, Musculoskeletal (RMSD) diseases, 99 = Patients with at least one disease from each of the groups, Data version: 2011 to Oct 2016

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

Figure 3-15: Disease distribution by Seasonal Variations and gender

Distype	Code	Description	season / Patient Gender												Grand Total
			01 Vasant Rutu		02 Grishma Rutu		03 Varsha Rutu		04 Sharad Rutu		05 Hemant Rutu		06 Shishir Rutu		
			F	M	F	M	F	M	F	M	F	M	F	M	
Metabolic	P5.0	Prameha	163	207	194	263	208	279	167	230	151	187	166	203	1,497
	M2.0	Madhumeha	158	268	184	271	188	309	198	330	144	241	153	244	1,441
	S16.0	Sthaulya	109	59	147	105	138	94	161	95	111	81	119	75	989
	M10.1	Medoroga - Sthula medho ..	88	53	97	52	98	54	89	46	84	47	78	39	627
	M10.2	Medoroga - Sukshma med..	14	32	13	38	9	40	14	36	11	29	9	23	193
	P5.4	Prameha - Upadrava	8	21	14	27	19	28	16	15	6	20	12	15	150
	M10.0	Medoroga	16	15	18	13	23	17	16	14	18	12	19	13	149
	P5.3	Prameha - Sthula	6	4	5	5	3	6	6	3	6	6	3	3	43
	P5.2	Prameha - Pidaka	3	1			7		3	1	1	1		2	17
	P5.1	Prameha - Krusha	1						1	3	1	1	1		9
RMSD	V2.63	Vaatavyaadhi - Sandhigat..	712	317	855	384	855	376	784	334	619	284	585	276	4,299
	V2.0	Vaatavyaadhi	373	316	415	354	485	373	368	306	275	249	362	290	2,993
	V2.23	Vaatavyaadhi - Gridhrasee	366	308	417	362	450	361	435	335	342	293	335	275	2,613
	S14.21	Sthaanabhedena Shoola - ..	133	133	163	139	190	154	175	154	116	114	139	116	1,333
	S13.11	Sthaanabhedena Graha - ..	90	104	106	113	130	133	122	113	117	111	114	107	982
	A2.0	Aamavaata	218	59	225	62	212	63	237	63	177	48	186	50	832
	A3.0	Abhighataj Shoola	59	66	90	89	76	85	68	91	57	68	51	69	739
	V1.0	Vaatarakta	42	58	61	83	85	75	74	73	62	63	50	47	529
	S14.0	Sthaanabhedena Shoola	73	42	80	53	51	43	68	51	46	49	74	44	523
	V2.36	Vaatavyaadhi - Kategata..	52	41	54	60	62	61	42	29	16	29	33	35	404
	S14.7	Sthaanabhedena Shoola - ..	21	21	32	30	21	31	32	31	26	16	28	21	233
	S14.17	Sthaanabhedena Shoola - ..	14	15	30	18	28	20	29	27	19	18	7	11	206
	V2.9	Vaatavyaadhi - Asthigata ..	26	20	19	19	21	19	42	15	36	20	21	8	206
	V2.75	Vaatavyaadhi - Vaatakant..	21	5	26	14	43	14	29	9	20	5	18	8	164
	S13.19	Sthaanabhedena Graha - ..	14	15	12	18	18	17	12	10	12	14	15	14	121
	V2.30	Vaatavyaadhi - Jaanugata..	12	7	20	13	10	5	7	10	4	4	8	3	92
	S15.31	Sthaanabhedena Shoola - ..	11	9	9	7	14	7	11	16	14	6	9	5	89
	S14.23	Sthaanabhedena Shoola - ..	4	6	15	7	0	6	12	5	0	6	6	11	92

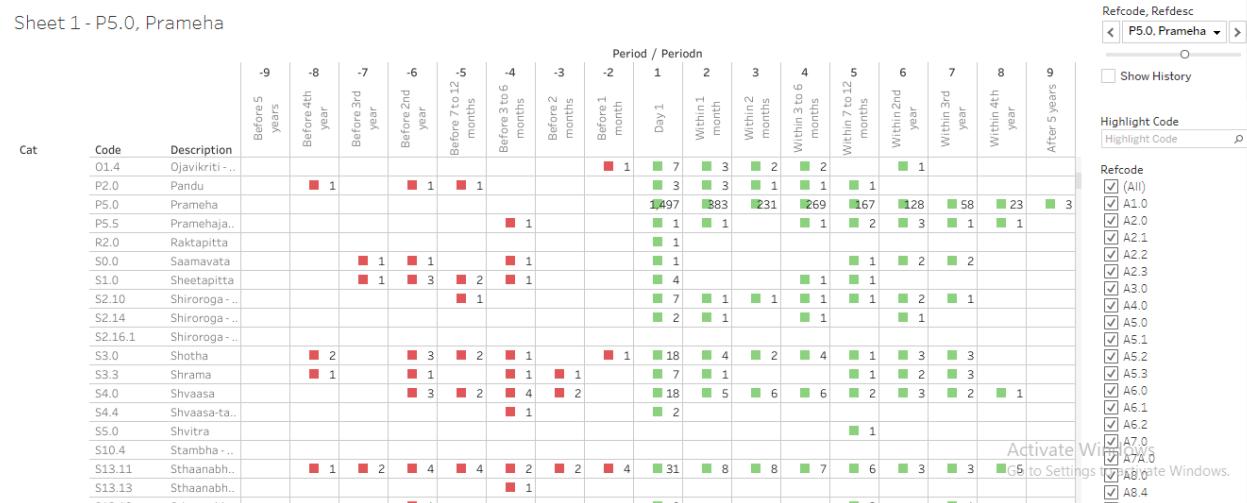
Distype: Metabolic and RMSD, Code: ACD disease code, Description: disease description, seasons are presented as: Vasant rutu, Grishma rutu, Varsha rutu, Sharad rutu, Hemant rutu, and Shishir rutu, Data version: 2011 to Oct 2017

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

Figure 3-16: Pre and Post Disease Classification Analysis

Example 1: Prameha

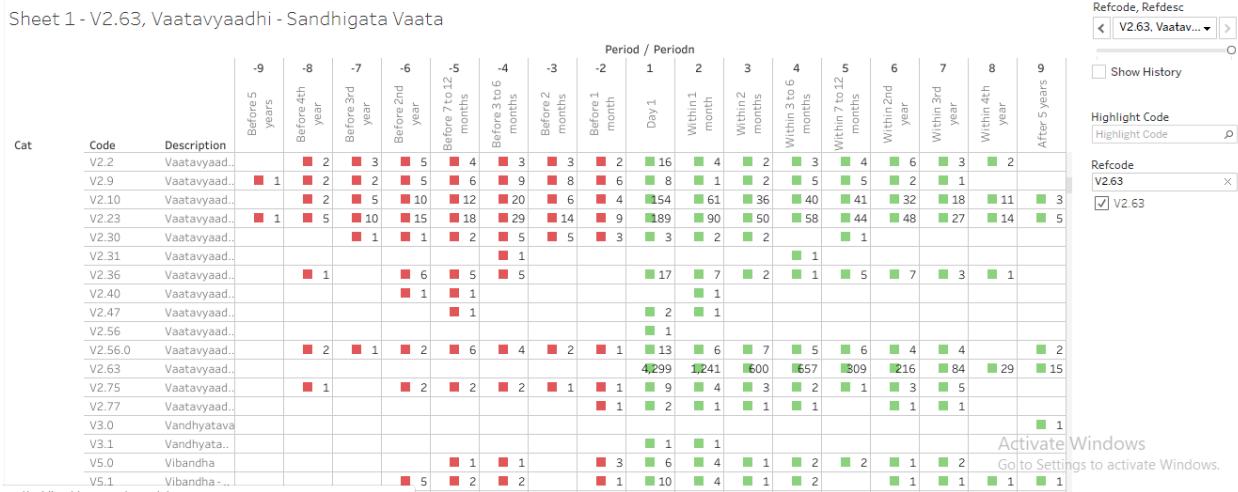
Sheet 1 - P5.0, Prameha



Cat: category of disease and medicine, Code: ACD code, prescribed medicine types, pre and post visit window w.r.to the 1st day of each of the reference diseases, 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

Example 2 Vaatavyadhi – Sandhigata Vaata



Cat: category of disease and medicine, Code: ACD code, prescribed medicine types, pre and post visit window w.r.to the 1st day of each of the reference diseases, 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

3.4 Studying diagnostics and interventions

Almost all the ICD categories were represented in the analysis. Some of the ICD classes had more patients than other categories. Age distribution showed natural variation. Visit distribution and duration for which patients were visiting hospital look like the earlier analysis. The ACD and ICD mapping exercise showed that the current hospital data demonstrates all the types of diseases being catered to at the hospital. (Figure 3-17, Figure 3-18, Figure 3-19, Figure 3-20)

The bar graph representation (Figure 3-21) provided a view of the spread of the patient population their diseases versus their prakriti classification seen at the hospital. The prakriti type as well as disease type can be filtered. It also provided a view of the combination of the gender and the prakriti manifesting into the kind of doshas and the most prevalent doshas for the combination.

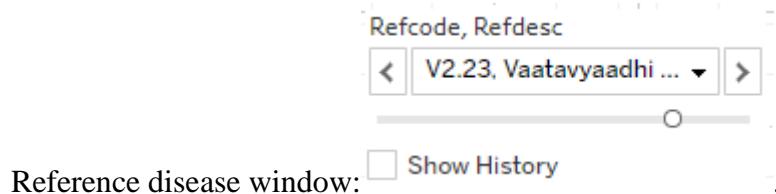
Observations from the 1st Co-morbidity approach (Figure 3-22, Figure 3-23, Figure 3-24): This display provided a comprehensive view of the disease clusters. Comorbidities were easily identified, some of them are clinically relevant, and some of them are not. Bubble size provided comparative view of number of patients reporting a specific disease. The age group distribution for each gender was available. Some diseases were reported more by males or by females, easy to spot on the graph. Box named “Number of other diseases” provided a contextual display about number of co-morbidities. Some diseases had higher number of co-morbidities, some had lower number. Variations were seen amongst gender as well. This analysis did not consider the before or after nature of time points, hence did not provide insights into the causal relationships between diseases. The 2nd co-morbidity approach provided views on the seasonal variations of diseases as well as seasonal co-morbidities (Figure 3-25). The 3rd co-morbidity approach provided the following: The collapsible tree showed progression of diseases as experienced by patients and reported in the database. The tree showed approximately 12,500 lines of data in very short space. Some diseases were experienced more by males than by females. Some diseases were only reported by one of the genders. Some diseases had many more branches than a few others. Some of the disease trajectories had very few numbers of patients. Some of the trajectories may be clinically meaningful and some may not be meaningful (Figure 3-26).

Treatment and disease analysis at individual patient level: This analysis was explained using example patient MR000335: For this patient there were 17 visits in the database, there were 17 distinct diseases reported and 45 distinct treatments, services prescribed. 4 out of 17 diseases were repeated and 10 out of 45 treatments had been repeated. When a new disease was reported, usually a new treatment or treatments re-reported, if there was only a new treatment added then it could indicate, the earlier treatment may not have worked, or it described the treatment regimen. If only new diseases were added and no new treatment was added, then the same treatment could work for multiple diseases. These visualizations allow the treating doctor insights into newer diseases getting reported as well as what newer treatments have been prescribed at what time points (Figure 3-27, Figure 3-28, Figure 3-29).

Area graph representation of diseases provided information about 800+ diseases, almost all diseases present in the database in very short space. Due to the data visualization scheme variations caused by day-to-day, seasons, gender, and diseases could be interpreted very easily. The interactive nature of visualization allowed for real time subset of diseases. One of the 4 diseases displayed has very few patients compared to other 3 diseases showing different nature of diseases (Figure 3-30).

Cross tabulation of prescribed treatments and disease group by gender was generated. 1st example is created using Balaristham. The source variable captured the quantity + unit + company name in the same variable, which did not allow for 100% accurate numerical calculations, but still provided a good idea. Only 30 patients having metabolic diseases were prescribed the medicine whereas 1,142 patients with RMSD were prescribed. 2nd example is created using bhasma: approximately 287 patients out of 17,406, 1.5% of patients are prescribed bhasmas (at least treatments having word “bhasma”) for various diseases (Figure 3-35, Figure 3-36). Summary statistics and hypothesis testing was conducted to conclude any impact of bhasma on visit duration. 514 patients were identified with at least 1 bhasma treatment. The mean duration of treatment before 1st bhasma treatment was 14.8 days, min - max duration was reported as (1, 111) whereas the mean duration of treatment after 1st bhasma treatment was 10.6 days, min - max duration was reported as (0, 89). The t-test at 5% significance level shows statistically significant difference between duration of treatment before bhasma treatment and duration of treatment after bhasma treatment.

Circular display, how to read the visualization? Figure 3-37 and Figure 3-38 show 2 examples of 2 combinations, the 1st example had many green bars, and the 2nd combination had very few green bars. Details about the display: For each reference disease 1 page was created. Each page was controlled by a combination of “Reference disease + disease”, “Reference disease + medicine”



HighlightCode

[A6.0:Amlapitta](#)

Reference disease or medicine window:

Tables displayed in the top part of the display: there were 9 columns created for each time point.

01 Day 1 Distinct count of Tab Code	02 1 month Distinct count of Tab Code	03 2 months Distinct count of Tab Med	04 3 to 6 months Distinct count of Tab Code	05 7 to 12 months Distinct count of Tab Med	06 2nd year Distinct count of Tab Code	07 3rd year Distinct count of Tab Med	08 4th year Distinct count of Tab Code	09 5 years Distinct count of Tab Med
---	---	---	---	---	--	---	--	--

2 columns were displayed in each time point to display “count of distinct number of diseases” and “count of distinct number of medicines”.

Count of distinct number of diseases:

Distinct count
of Tab Code

Count of distinct number of medicines:

Distinct count
of Tab Med

There were 3 rows for “Before period”, “Day 1” and “After period” with 2 lines in each period. Day 1 cell showed the start day of reference disease “V2.23: Vaatavyadhi – Gridhrasee”. This example showed 137 total number of distinct diseases reported and 664 total number of distinct medicines prescribed on day 1 for this combination of reference disease “V2.23: Vaatavyadhi – Gridhrasee” and disease “A6.0: Amlapitta”.

The first line in the Day 1 cell showed, 1 disease – which is “A6.0: Amlapitta” and 81 distinct medicines prescribed. These 81 different treatments could have been prescribed for “A6.0: Amlapitta”.

		01 Day 1	
		Distinct count of Tab Code	Distinct count of Tab Med
01	In		
Before	O.		
	T..		
02 Day 1	In	1.0	81.0
	O.	136.0	664.0
	T..	137.0	664.0

Cells in the “Before period” line provided the following information:

This example shows 1 month before “V2.23: Vaatavyadhi – Gridhrasee”, there were 61 distinct diseases and 298 distinct medicines reported.

The first line in the cell shows, 1 disease – which is “A6.0: Amlapitta” and 11 distinct medicines prescribed. These 11 different treatments could have been prescribed for “A6.0: Amlapitta”.

		01 Day 1	02 1 month
		Distinct count of Tab Code	Distinct count of Tab Med
01 Before	In	1.0	11.0
	O.	60.0	294.0
	T..	61.0	298.0

Cells in the “After period” line provide the following information:

This example shows 1 month before “V2.23: Vaatavyadhi – Gridhrasee”, there were 96 distinct diseases and 633 distinct medicines reported.

The first line in the cell shows, 1 disease – which is “A6.0: Amlapitta” and 49 distinct medicines prescribed. These 49 different treatments could have been prescribed for “A6.0: Amlapitta”.

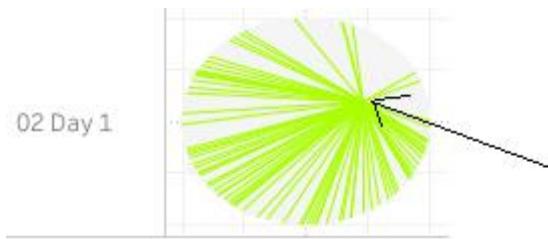
		01 Day 1	02 1 month
		Distinct count of Tab Code	Distinct count of Tab Med
03 After	In	1.0	49.0
	O.	96.0	633.0
	T..	97.0	635.0

The bottom section follows the same structure as the top section. The following table provided distinct number of diseases reported and distinct number of medicines prescribed for this particular combination. There were 308 diseases and 1,357 medicines reported for this combination of reference disease “V2.23: Vaatavyadhi – Gridhrasee” and disease “A6.0: Amlapitta”.

Distinct diseases / medicines	
Disease	Medicine
308	1,357

Explanation about the circular view:

The starting point marked the position of the other disease in this case, “A6.0: Amlapitta”, The green colored spokes going from point of origin were different treatments prescribed. These were showing 664 distinct medicines prescribed on day 1.



Hovering tooltip provided details about the disease, treatment name and count of number of



patients:

The inner circle displayed the diseases. And the outer circle displayed the treatments.

Distance based Pre and post analysis: in the example for the disease M2.0 (Madhumeha), disease trajectory distances were plotted (Figure 3-39). (1) Madhumeha was reported by 1,441 patients at least once. (2) Of these 567 patients reported only Madhumeha and no other disease thereafter. For such patients, the disease trajectory calculation was not possible, hence these patients were removed from the analysis. (3) The following table showed details of patient count. The disease trajectory calculation was based on 874 patients comprising of 530 males and 344 females

patient_gender	numtype	ncount
M	Only 1 dis	377
F	Only 1 dis	190
M	> 1 dis	530
F	> 1 dis	344

Out of 530 male patients: (1) 201 patients had diseases reported before the 1st reported instance of M2.0, (2) 460 patients had at least one other disease reported other than M2.0 on or after the 1st reported instance of M2.0, (3) Disease trajectory for 70 patients could not be calculated since the next reported disease was M2.0.

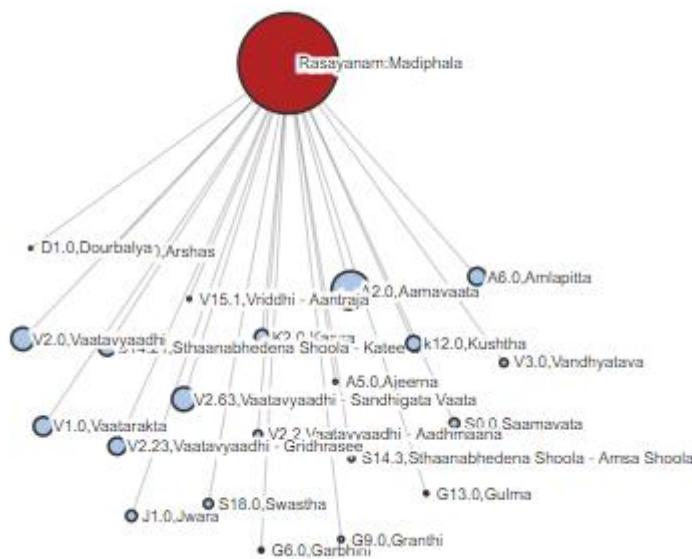
Out of 344 female patients: (1) 140 patients had diseases reported before the 1st reported instance of M2.0, (2) 291 patients had at least one disease other than M2.0 reported on or after the 1st reported instance of M2.0, (3) Disease trajectory for 53 patients could not be calculated since the next reported disease was M2.0.

patient_gender	numtype	refday2	ncount
M	> 1 dis	Before	201
M	> 1 dis	After	460
F	> 1 dis	Before	140
F	> 1 dis	After	291

The trajectories were calculated for these patients and displayed for the before and after period. More number of patients had disease trajectories in the “after onset” section. More than 73% of males lie in the score >0.25 and around 36% of them lie in the score >0.5 which could confirm that there were similar diseases experienced by the patients post the onset of the reference disease. Similarly, around 64% of females lie in the score >0.25 and around 24% of them lie in the score >0.5 which could confirm that there were similar diseases experienced by the patients post the onset of the reference disease. A few more examples of similar kind were shown for diseases: P5.0: Prameha, V2.23: Vaatavyaadhi - Gridhrasee, V2.63: Vaatavyaadhi - Sandhigata Vaata (Figure 3-40).

The radar plot shows multi-dimensional data in a short space, 7 different parameters were shown on 7 vertices. Different shapes suggest that there were underlying differences to the data structure.

Dynamic bubble plot visualization showed intricate relationships and details in a limited space, one more analysis method to see multi-dimensional data (Figure 3-42). Using data for Amavaata (A6.0), in the snapshot below, we could see the patients with Aamvata who received Rasayanam Madiphala and in addition what were the other diseases that these patients reported.



Another view which showed the further detailed view of Aamvata patients having received Rasayanam Madiphala and further having reported Stanbheedana Kathee Shoola disease and

further its treatment. Further, the big bubble here closer to Stanbheedana Kathee Shoola disease was Aamdosha which showed that Stanbheedana Kathee Shoola was also one of the diseases reported by Aamdosha patients.

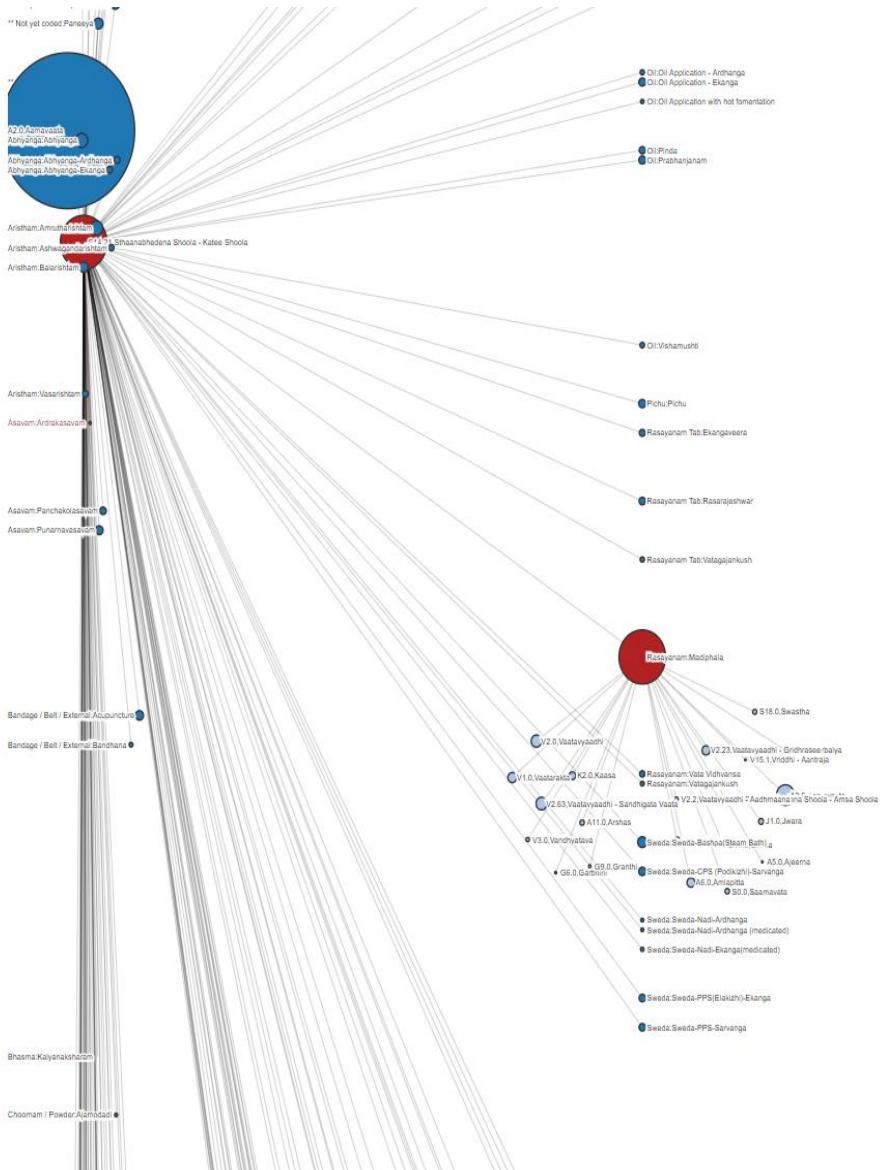


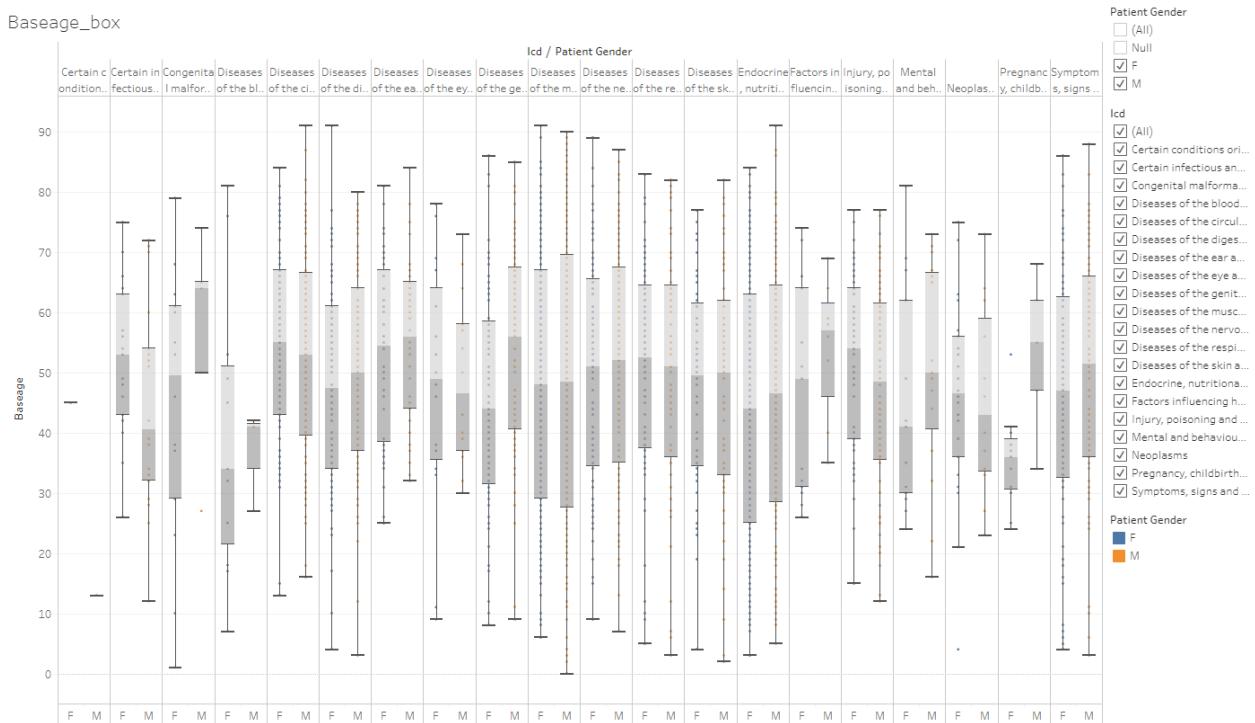
Figure 3-17: ICD classification by Gender

ICDFreq

Icd	High	Grand Total	Patient Gender		
			Null	F	M
Grand Total		8,027	1	4,455	3,571
Certain conditions originating in the perinatal period	Total	2		1	1
P91		2		1	1
Certain infectious and parasitic diseases	Total	90		42	48
A01		9		6	3
A06		1		1	
A15		7		3	4
A18		3			3
A30		1		1	
A41		1		1	
A46		1		1	
A81		2		2	
A90		2		2	
A92		5		4	1
B01		1		1	
B02		5			5
B07		8		5	3
B08		2		1	1
B19		6			6
B34		1			1
B35		32		12	20
B37		4		4	
B74		1		1	
B82		1			1

ICD classification and ICD classification high level categories, frequency counts by gender. Data version: 2011 to Oct 2017
https://public.tableau.com/views/Allopathic_diag/ICDFreq?:language=en&:display_count=y&:origin=viz_share_link

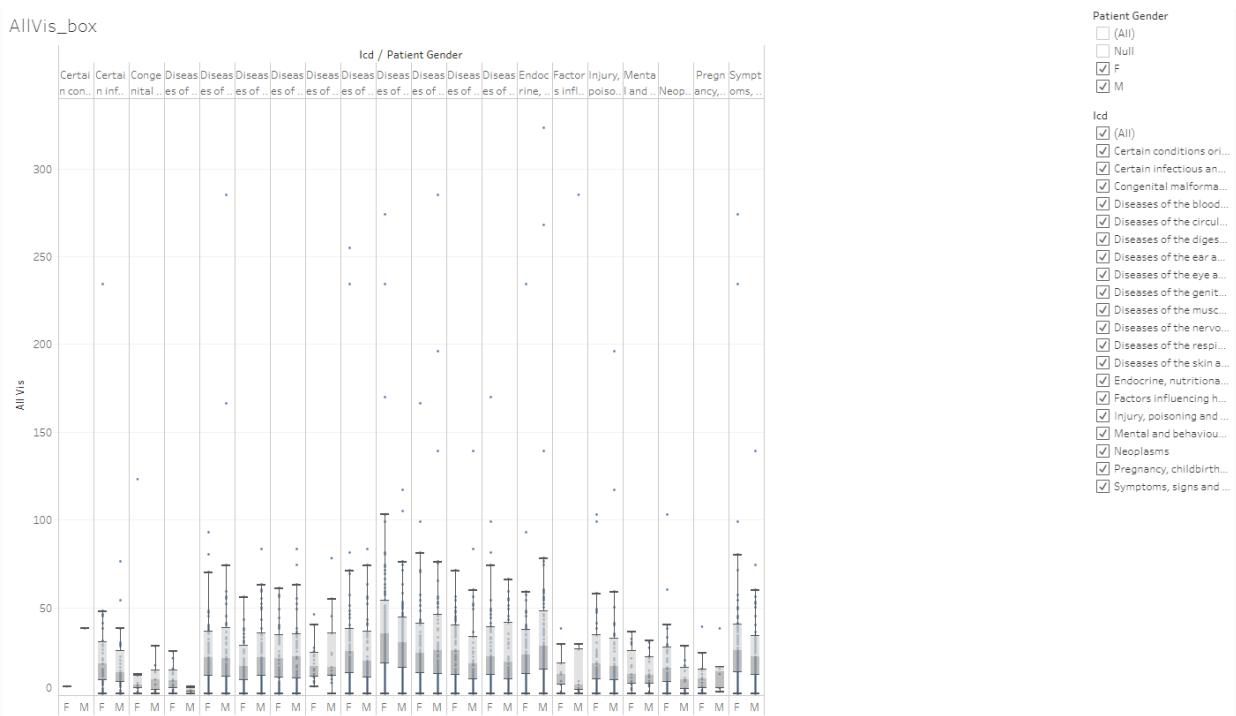
Figure 3-18: Age distribution by ICD classification and Gender



Boxplot representation of age by ICD classification and gender, Data version: 2011 to Oct 2017

https://public.tableau.com/views/Allopathic_diag/Baseage_box?:language=en&:display_count=y&:origin=viz_share_link

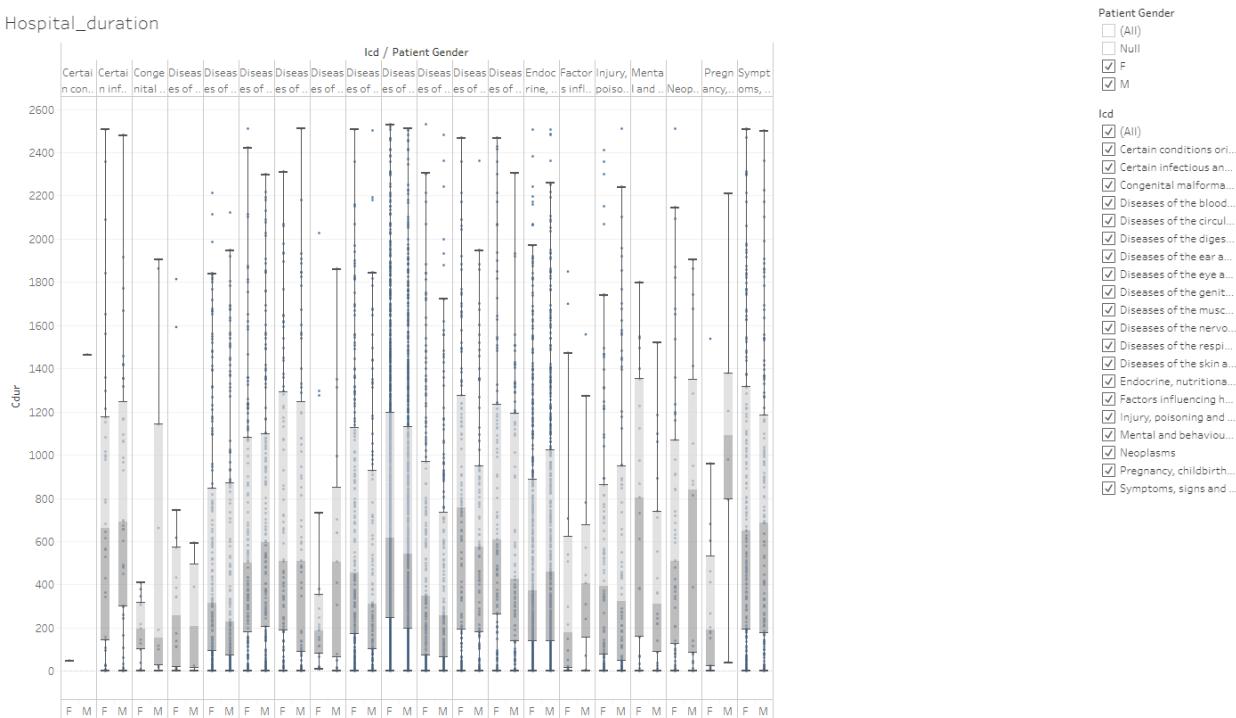
Figure 3-19: Visit distribution by ICD classification and Gender



Boxplot representation of hospital number of visits by ICD classification and gender, Data version: 2011 to Oct 2017

https://public.tableau.com/views/Allopathic_diag/AllVis_box?:language=en&:display_count=y&:origin=viz_share_link

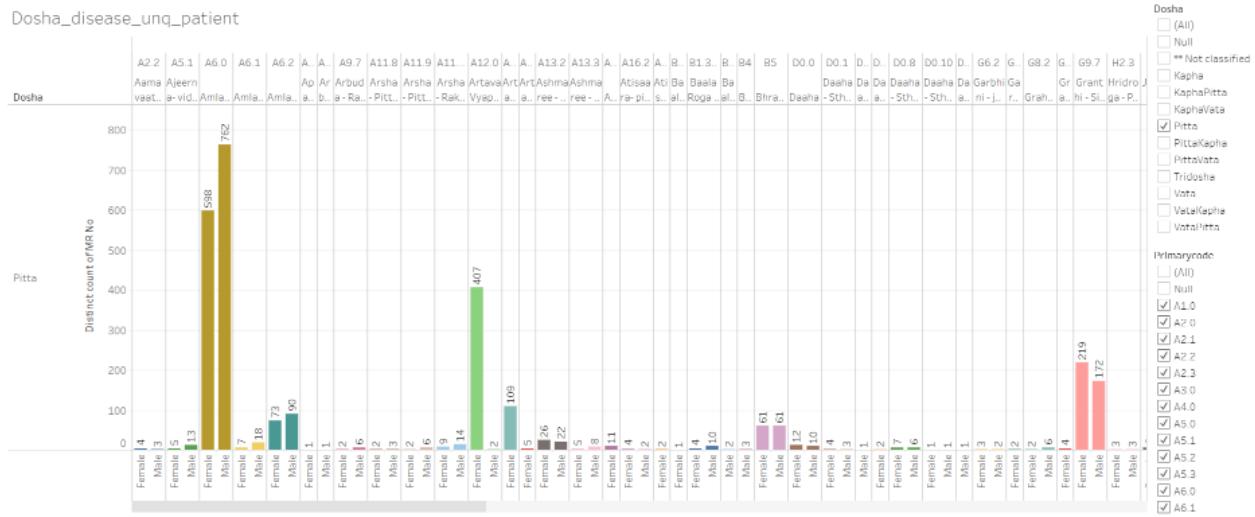
Figure 3-20: Duration distribution by ICD classification and Gender



Boxplot representation of hospital visit duration by ICD classification and gender, Data version: 2011 to Oct 2017

https://public.tableau.com/views/Allopathic_diag/Hospital_duration?:language=en&:display_count=y&:origin=viz_share_link

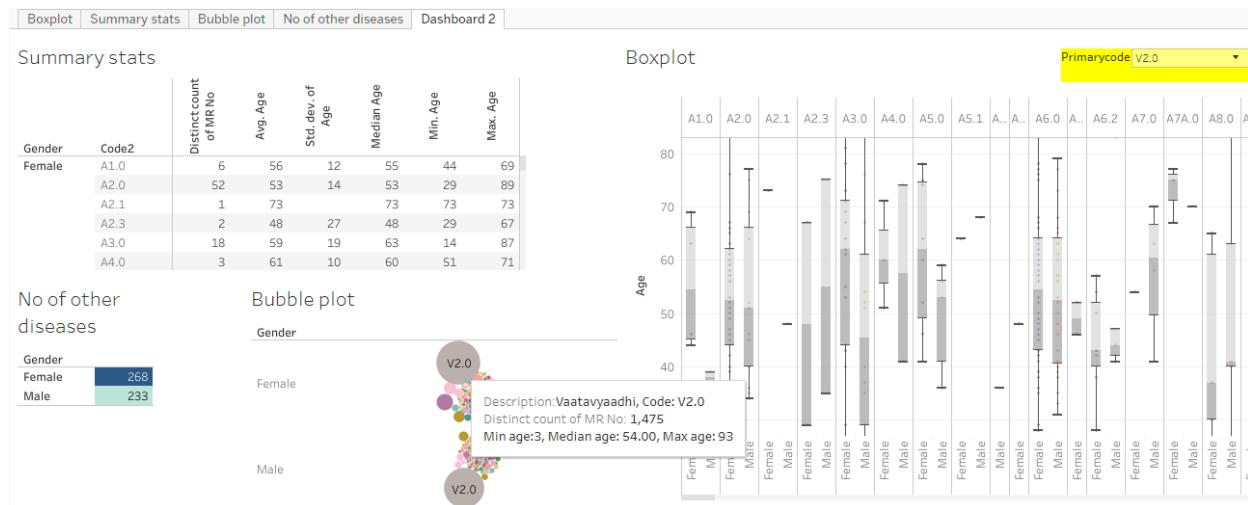
Figure 3-21: Disease classification by Prakriti and Gender



Dosha: Prakriti type, x-axis: male and female grouped by individual disease, y-axis: frequency counts of unique patients, Data version: 2011 to Oct 2016

https://public.tableau.com/views/Disease_by_dosha_type/Dosha_disease_unq_patient?:display_count=y&:origin=viz_share_link

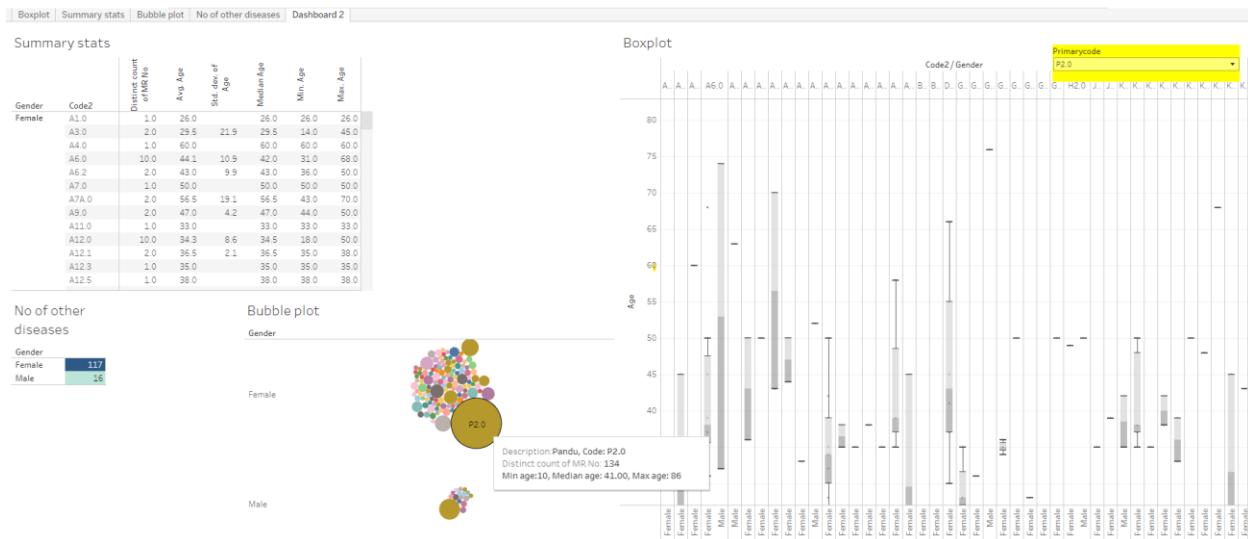
Figure 3-22: Co-morbidity analysis approach 1 example 1: Vaatavyadhi



Summary stats section: descriptive statistics details by gender and other diseases reported, No of other disease: distinct number of other diseases reported by patients who had reported the primary disease, Bubble plot: frequency count of distinct patients by disease, Boxplot: age distribution by disease and gender, Data version: 2011 to Oct 2016

https://public.tableau.com/views/Primary_disease_and_all_other_diseases/Dashboard2?:display_count=y&:origin=viz_share_link

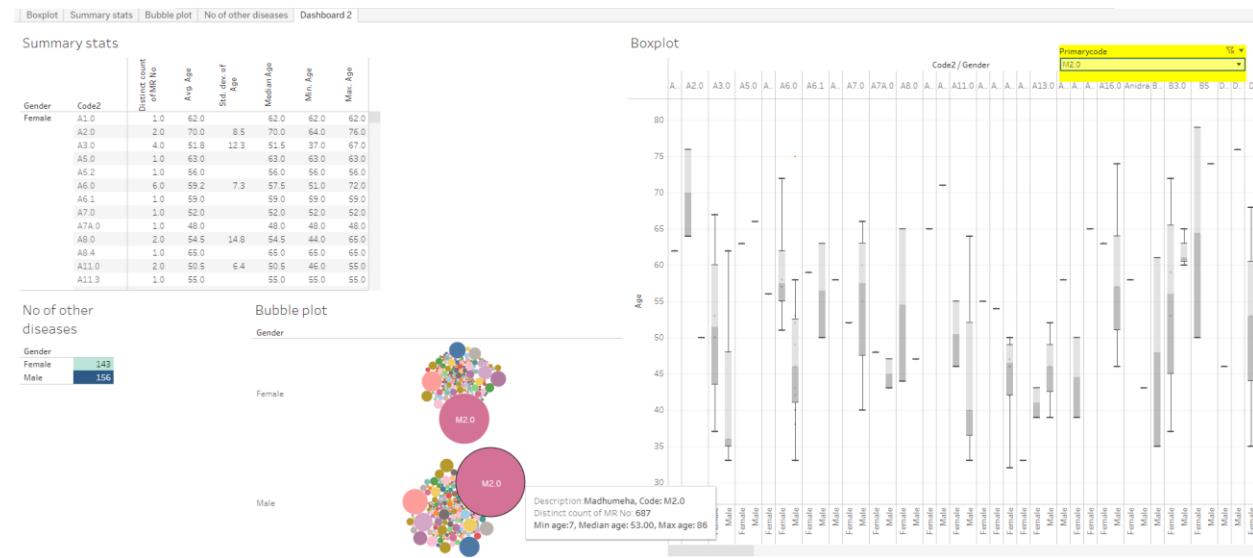
Figure 3-23: Co-morbidity analysis approach 1 example 2: Pandu



Summary stats section: descriptive statistics details by gender and other diseases reported, No of other disease: distinct number of other diseases reported by patients who had reported the primary disease, Bubble plot: frequency count of distinct patients by disease, Boxplot: age distribution by disease and gender, Data version: 2011 to Oct 2016

https://public.tableau.com/shared/C27BJZNZG?:display_count=n&:origin=viz_share_link

Figure 3-24: Co-morbidity analysis approach 1 example 3: Madhumeha



Summary stats section: descriptive statistics details by gender and other diseases reported, No of other disease: distinct number of other diseases reported by patients who had reported the primary disease, Bubble plot: frequency count of distinct patients by disease, Boxplot: age distribution by disease and gender, Data version: 2011 to Oct 2016

https://public.tableau.com/shared/6SQBQ6QRZ?:display_count=n&origin=viz_share_link

Figure 3-25: Co-morbidity analysis approach 2



Upper section: bubble plots for the reference disease and other diseases reported, bubble size is based on number of distinct patients. Lower sections: unique number of other diseases reported for the reference disease. 1, 2, ..., 12: January to December month, Data version: 2011 to Oct 2016

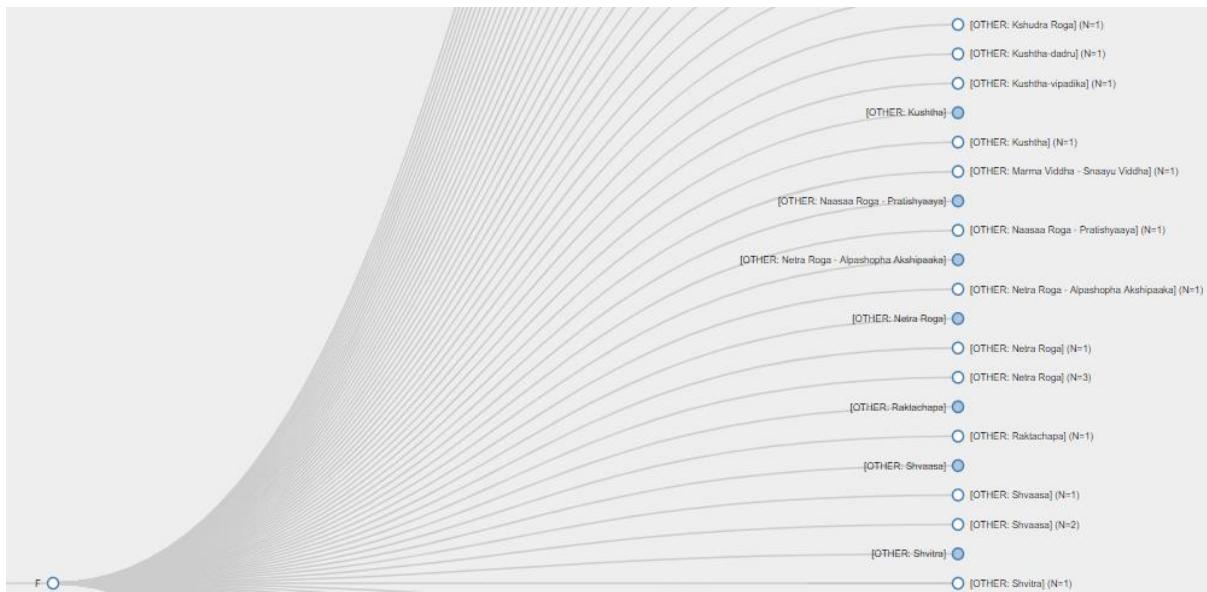
https://public.tableau.com/views/PrimDis_otherDis_ByMonth/Dashboard1?:language=en&:display_count=y&origin=viz_share_link

Figure 3-26: Co-morbidity analysis approach 3: collapsible tree view

Initial view of the tree



After clicking on F (Female), the collapsible tree opens up



An example of a disease experienced only by one gender



Figure 3-27: Patient Disease and Treatment administration by Study Day

Patient_Dis_Med_counts					
Mr No	Grpday	Studyday	Cat / Newold2		Medicine
			Disease	1sttim..	
MR000335	1	1		1	4
	2	24		1	1
	3	69		1	1
	4	521	1		1
	5	565		1	2
	6	572	5		10
	7	646	1		3
	8	1093	1		1

Patient_1st_Repeat					
Mr No	Grpday	Cat	Type med	Coded med	
MR000335	1	Disease	** Not yet coded	** Not yet coded	
		Medicine	Asavam	Lodhrasavam	
			Kashayam	Varanadi	
				Vasaguloochayadi	
			Tablet/Gulika/Vati	Kanchanara Guggulu	
	2	Disease	** Not yet coded	** Not yet coded	
		Medicine	Asavam	Lohasavam	

Mr No
MR000335
 MR000335

Newold2
█ 1st dose
█ 1st time disease
█ Repeat

Activate Windows
Go to Settings to activate Windows.

The report displays individual patient data. Upper left part: Mr No: Patient ID, study day, Disease reported 1st time and repeated, medicine reported 1st time and repeated, Lower part of the report displays individual patient data for each day and distinguishes: 1st dose, 1st disease and Repeat reporting for the same. Data version: 2011 to Oct 2017

https://public.tableau.com/views/080_medicine_dis_repeat_prop/Dashboard1?:language=en&:retry=yes&:display_count=y&:origin=viz_share_link

Figure 3-28: Patient Disease by Study Day and Treatment administration by Study Day

The dashboard displays two tables side-by-side. The left table, titled 'Disease', shows the number of diseases reported at a particular visit. The right table, titled 'Medicine', shows the number of prescribed treatments reported at a particular visit. Both tables have columns for Mr No, Grpday, Studyday, and Newold2Dis or Newold2.

Mr No	Grpday	Studyday	Newold2Dis	
			1st time disease	Repeat
MR000335	1	1	1	
	2	24		1
	3	69		1
	4	521	1	
	5	565		1
	6	572	1	
	7	646		1
	8	1093	1	
	9	1388		1
	10	1418		1
	11	1481	1	
	12	1528		1
	13	1628	1	
	14	2025	2	
	15	2488		1
	16	2503	1	
	17	2508		2

Mr No	Grpday	Studyday	Newold2	
			1st dose	Repeat
MR000335	1	1	4	
	2	24	1	2
	3	69		1
	4	521		1
	5	565	2	1
	6	572		10
	7	646	3	
	8	1093	1	2
	9	1388		3
	10	1418	2	2
	11	1481		4
	12	1528		3
	13	1628	2	1
	14	2025	6	1
	15	2488		1
	16	2503	2	
	17	2508	2	1

The report displays individual patient data. Panel on the left hand: number of diseases reported at a particular visit, panel on the right hand: number of prescribed treatments reported at a particular visit, Data version: 2011 to Oct 2017

https://public.tableau.com/views/080_medicine_dis_all_met_rmsd_prop/Dashboard1?:language=en&:display_count=y&:origin=viz_share_link

Figure 3-29: Patient Cumulative Disease and Treatment administration by Visit

The dashboard displays two tables side-by-side. The left table, titled 'Patient_dis_med_Cumulative', shows absolute values of diseases and prescribed treatments till particular visits. The right table, titled '%Patient_dis_med_Cumulative', shows the percentage values of diseases and prescribed treatments till particular visits. Both tables have columns for Mr No, Cumday2, and Cat / Newold2.

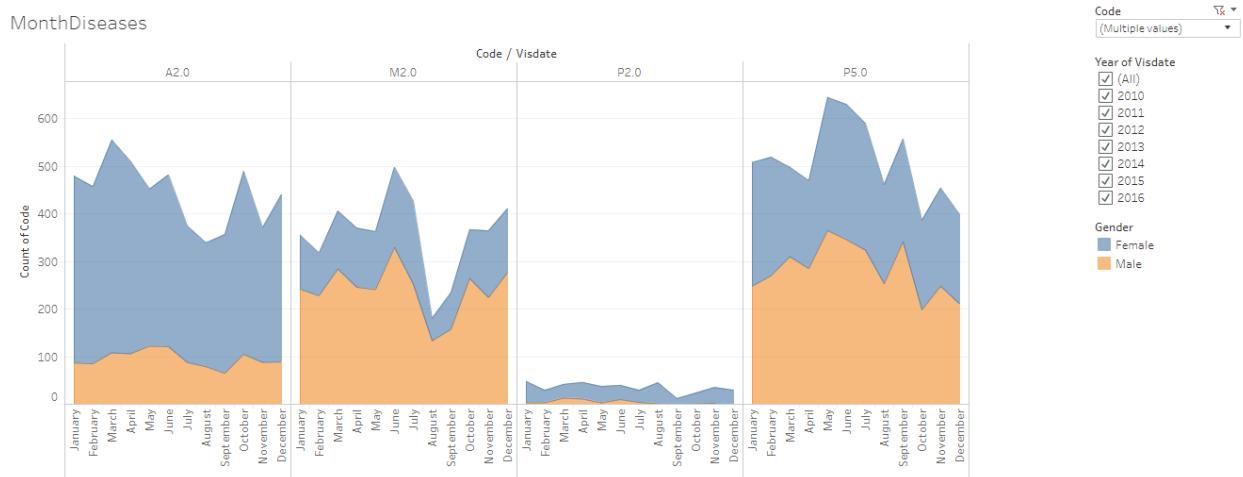
Mr No	Cumday2	Cat / Newold2	
		Disease	Medicine
MR000335	Till visit 1	1	4
	Till visit 2	1	5
	Till visit 3	1	6
	Till visit 4	2	6
	Till visit 5	2	8
	Till visit 6	7	18
	Till visit 7	8	21
	Till visit 8	9	22
	Till visit 9	10	25
	Till visit 10	10	27
	Till visit 11	11	31
	Till visit 12	12	34
	Till visit 13	13	36
	Till visit 14	15	42
	Till visit 15	16	43
	Till visit 16	17	45

Mr No	Cumday2	Cat / Newold2	
		Disease	Medicine
MR000335	Till visit 1	100.00%	100.00%
	Till visit 2	50.00%	50.00%
	Till visit 3	33.33%	66.67%
	Till visit 4	50.00%	50.00%
	Till visit 5	40.00%	60.00%
	Till visit 6	70.00%	30.00%
	Till visit 7	72.73%	27.27%
	Till visit 8	75.00%	25.00%
	Till visit 9	76.92%	23.08%
	Till visit 10	71.43%	28.57%
	Till visit 11	73.33%	26.67%
	Till visit 12	75.00%	25.00%
	Till visit 13	76.47%	23.53%
	Till visit 14	78.95%	21.05%
	Till visit 15	80.00%	20.00%
	Till visit 16	80.95%	19.05%

The report displays individual patient data. Panel on the left hand: absolute values of diseases and prescribed treatments till particular visits, panel on the right hand side: % values of diseases and prescribed treatments till particular visits, Data version: 2011 to Oct 2017

https://public.tableau.com/views/080_medicine_dis_repeat_prop_cumulative/Dashboard1?:language=en&:display_count=y&:origin=viz_share_link

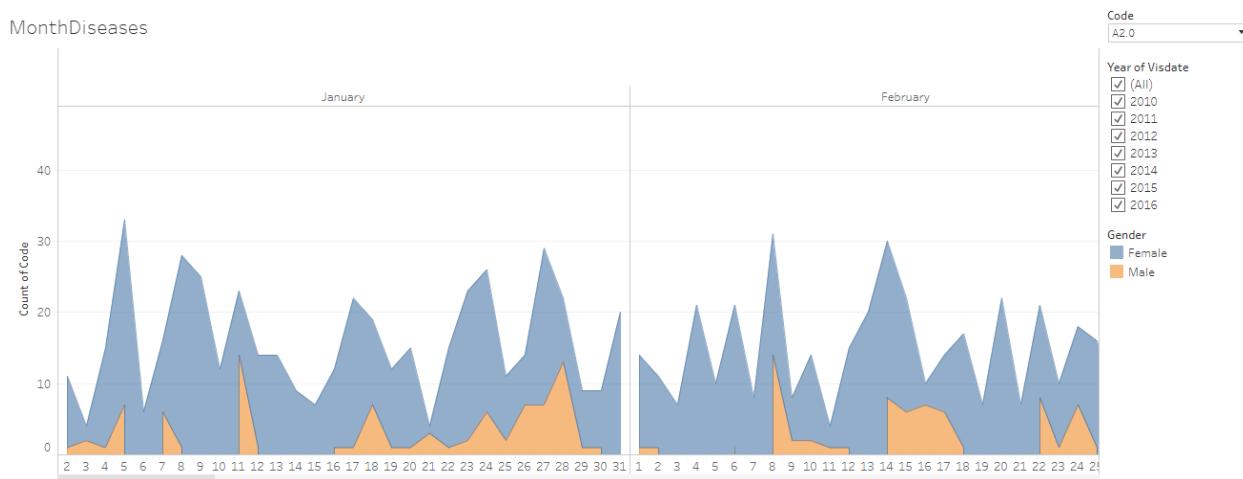
Figure 3-30: Area graph representation of diseases



Area graph representation of diseases: x-axis: Month, y-axis: frequency count of unique patients, disease code: the underlying data for each disease, peach colour: counts for male, blue colour: counts for female.



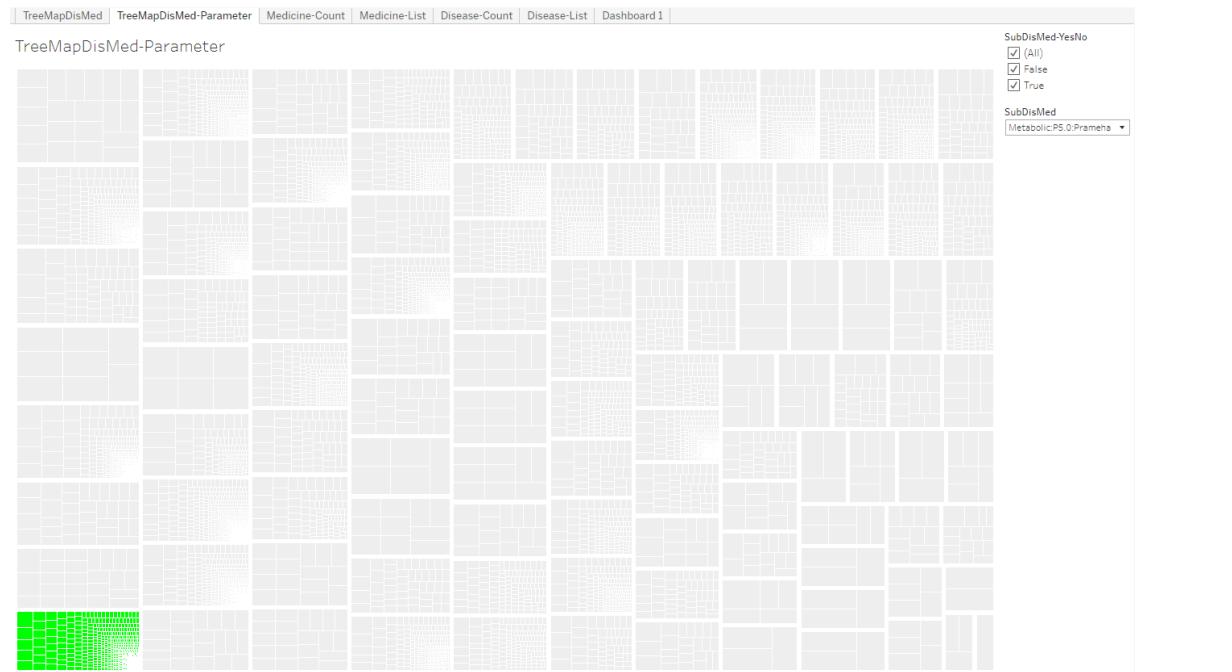
x-axis can be expanded to convert the monthly view to more granular unit (week, day).



The original data displayed in the 1st part of the presentation is opened for daily view for a particular disease code A2.0, Data version: 2011 to 2016

https://public.tableau.com/views/IndividualPatientCalendar/MonthDiseases?:language=en&:display_count=y&:origin=viz_share_link

Figure 3-31: Mosaic plot: Disease and treatment representation example 1: Prameha

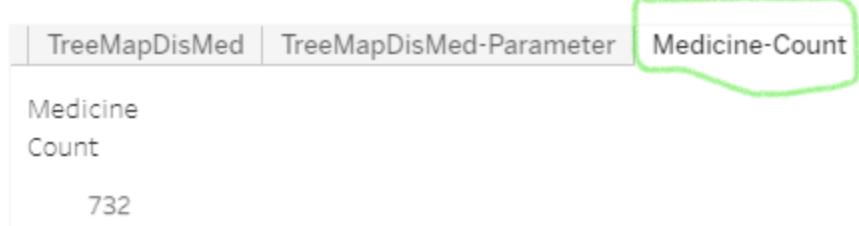


Mosaic plot: Each box is one disease, the selected disease is marked in Green colour, smaller boxes inside each disease display one intervention each, Data version: 2011 to Oct 2017

https://public.tableau.com/views/305_medicine_dur_by_dis/TreeMapDisMed-Parameter?:language=en&:display_count=y&:origin=viz_share_link



The snapshot above is the zoomed version of the Prameha block which highlights one of the many treatments prescribed for the same.



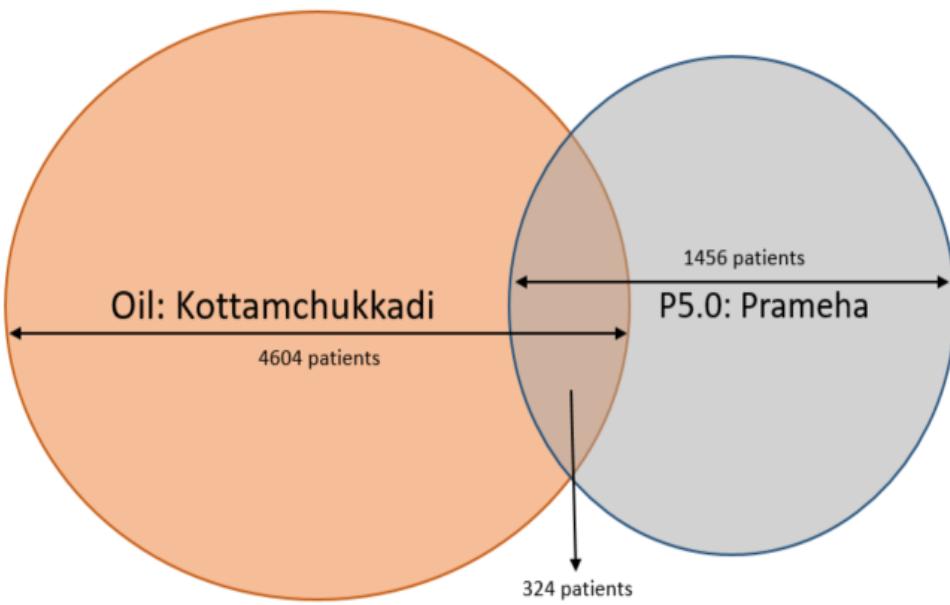
The Medicine count tab provides the information on total number of different medicines prescribed for Prameha. There are 732 distinct interventions.

Medicine-List							SubDisMed
Med02	Code02	totpatmed	totpatdis	n	percdis	percmed	Metabolic:P5.0:Prameha
Oil:Kottamchukkadi	Metabolic:P5.0:Prameha	4604	1456	324	22.25	7.04	Abc
Tablet /Gulika / Vati:Neuron	Metabolic:P5.0:Prameha	2065	1456	247	16.96	11.96	Abc
Oil:Chinchadi	Metabolic:P5.0:Prameha	2947	1456	221	15.18	7.5	Abc
Tablet / Gulika / Vati:Nisa Amalaki	Metabolic:P5.0:Prameha	798	1456	206	14.15	25.81	Abc
Tablet / Gulika / Vati:Rumastal Forte	Metabolic:P5.0:Prameha	1844	1456	199	13.67	10.79	Abc
Tablet / Gulika / Vati:D Nil	Metabolic:P5.0:Prameha	549	1456	155	10.65	28.23	Abc
Tablet / Gulika / Vati:Palsinuron	Metabolic:P5.0:Prameha	1511	1456	154	10.58	10.19	Abc
Tablet / Gulika / Vati:Diabecon DS	Metabolic:P5.0:Prameha	476	1456	145	9.96	30.46	Abc
Liniment:Myostal	Metabolic:P5.0:Prameha	1991	1456	145	9.96	7.28	Abc
Tablet / Gulika / Vati:Cervilon	Metabolic:P5.0:Prameha	1151	1456	143	9.82	12.42	Abc
Kashayam:Rasnasapthakam	Metabolic:P5.0:Prameha	1866	1456	133	9.13	7.13	Abc
Kashayam:Prasaranyadi	Metabolic:P5.0:Prameha	974	1456	124	8.52	12.73	Abc
Kashayam:Rasnaerandadi	Metabolic:P5.0:Prameha	2265	1456	122	8.38	5.39	Abc
Tablet / Gulika / Vati:Nuro - XT	Metabolic:P5.0:Prameha	1026	1456	120	8.24	11.7	Abc
Kashayam:Rasnayograjra Guggulu	Metabolic:P5.0:Prameha	967	1456	117	8.04	12.1	Abc
Kashayam:Asanadi	Metabolic:P5.0:Prameha	418	1456	113	7.76	27.03	Abc
Tablet / Gulika / Vati:Rumalaya Forte	Metabolic:P5.0:Prameha	1175	1456	106	7.28	9.02	Abc
Kashayam Tab:Rasnaerandadi	Metabolic:P5.0:Prameha	1182	1456	100	6.87	8.46	Abc
Kashayam:Sahacharadi	Metabolic:P5.0:Prameha	1603	1456	92	6.32	5.74	Abc
Tablet / Gulika / Vati:Trayodasanga Guggulu	Metabolic:P5.0:Prameha	1830	1456	88	6.04	4.81	Abc
Tablet / Gulika / Vati:Lumbaton	Metabolic:P5.0:Prameha	1433	1456	84	5.77	5.86	Abc
Tablet / Gulika / Vati:Ksheerabala	Metabolic:P5.0:Prameha	594	1456	76	5.22	12.79	Abc
Kashayam:Ashtavargam	Metabolic:P5.0:Prameha	548	1456	76	5.22	13.87	Abc
Tablet / Gulika / Vati:Flexy	Metabolic:P5.0:Prameha	562	1456	73	5.01	12.99	Abc
Oil:Narayana	Metabolic:P5.0:Prameha	374	1456	73	5.01	19.52	Abc
Oil:Murivenna	Metabolic:P5.0:Prameha	2329	1456	69	4.74	2.96	Abc
Oil:Sahacharadi	Metabolic:P5.0:Prameha	1100	1456	67	4.6	6.09	Abc
Kashayam:Maharasnayograjra Guggulu	Metabolic:P5.0:Prameha	660	1456	67	4.6	10.15	Abc
Tablet / Gulika / Vati:Chandraprabha	Metabolic:P5.0:Prameha	612	1456	66	4.53	10.78	Abc
Tablet / Gulika / Vati:Trayodasanga Guggulu DS	Metabolic:P5.0:Prameha	1249	1456	65	4.46	5.2	Abc
Tablet / Gulika / Vati:Mehamil	Metabolic:P5.0:Prameha	214	1456	59	4.05	27.57	Abc
Tablet / Gulika / Vati:Glukostat	Metabolic:P5.0:Prameha	175	1456	57	3.91	32.57	Abc
Oil:Pinda	Metabolic:P5.0:Prameha	937	1456	57	3.91	6.08	Abc
Rasayanam:Vasant Kusumakar Plain	Metabolic:P5.0:Prameha	535	1456	55	3.78	10.28	Abc
Tablet / Gulika / Vati:Sallaki M R	Metabolic:P5.0:Prameha	2972	1456	54	3.71	1.82	Abc
Kashayam:Kathaka Kathiradi	Metabolic:P5.0:Prameha	167	1456	53	3.64	31.74	Abc
Tablet / Gulika / Vati:Dia P	Metabolic:P5.0:Prameha	134	1456	51	3.5	38.06	Abc

A detailed list of medicines with the total patients prescribed with them as well as total patients suffering from Prameha are displayed in the snapshot above.

Figure 3-32: Disease and treatment example 2: P5.0: Prameha and Oil: Kottamchukkadi

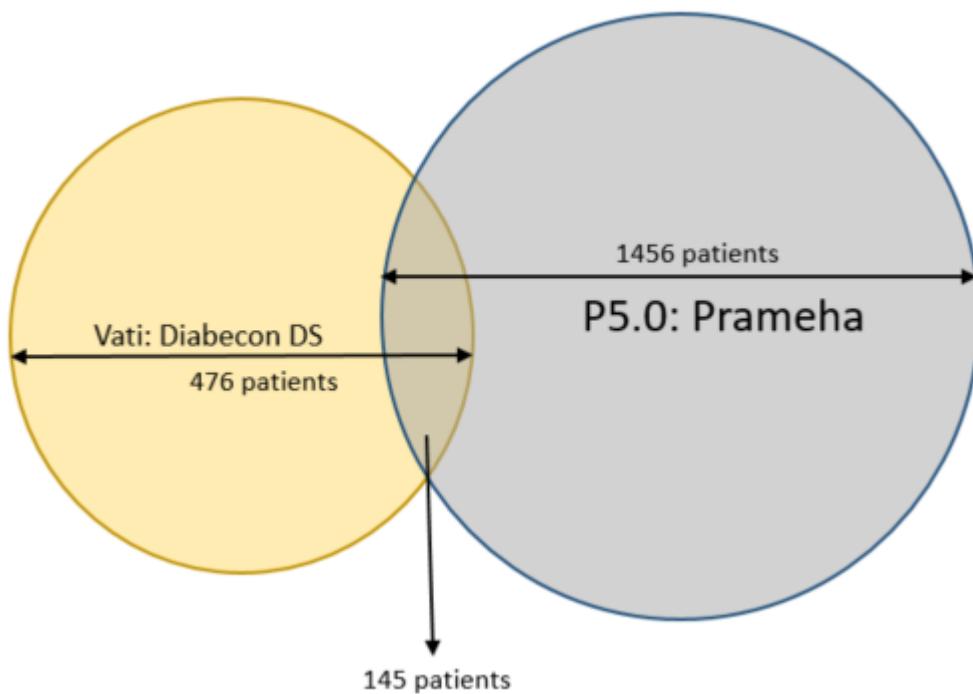
Medicine-List						
Med02	Code02	totpatmed	totpatdis	n	percdis	percmed
Oil:Kottamchukkadi	Metabolic:P5.0:Prameha	4604	1456	324	22.25	7.04



Totpatmed shows that there are 4,604 patients who have been prescribed with Oil: Kottamchukkadi; Totpatdis shows that there are 1,456 patients who are diagnosed with Prameha; this shows that the treatment is not Prameha specific. The n count of 324 shows the patients who had Prameha and were prescribed Oil: Kottamchukkadi. Percdis shows that 22.25% i.e. 324 / 1456 of patients having Prameha are prescribed this particular treatment. Percmed shows that only 7% of the time this medicine has been prescribed for Prameha patients.

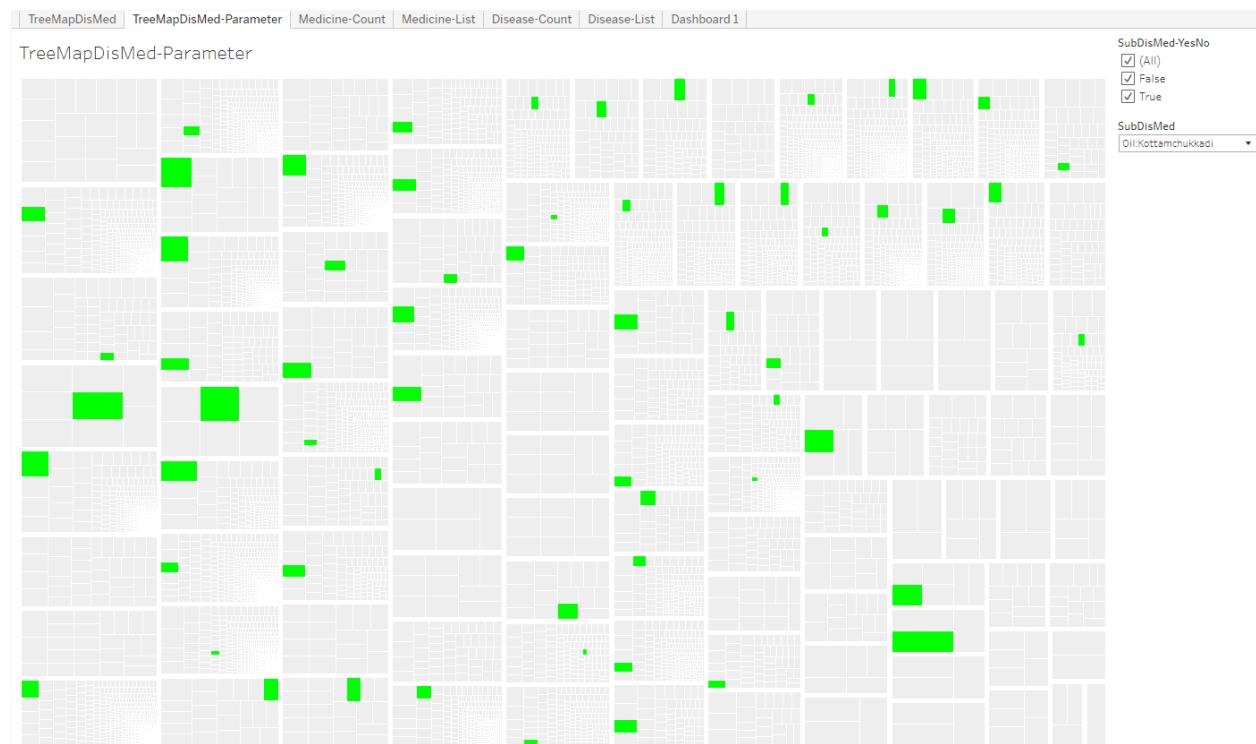
Figure 3-33: Disease and treatment example 3: P5.0: Prameha and Vati: Diabecon DS

Tablet / Gulika / Vati:Diabecon DS	Metabolic:P5.0:Prameha	476	1456	145	9.96	30.46
------------------------------------	------------------------	-----	------	-----	------	-------



Totpatmed shows that there are 476 patients who have been prescribed with Vati: Diabecon DS; Totpatdis shows that there are 1,456 patients who are diagnosed with Prameha; this shows that the treatment could be more prescribed to Prameha patients. The n count of 145 shows the patients who had Prameha and were prescribed Vati: Diabecon DS. Percdis shows that 9.96% i.e. 145 / 1,456 of patients having Prameha are prescribed this particular treatment. Percmed shows that only 30.46% of the time this medicine has been prescribed for Prameha patients.

Figure 3-34: Mosaic plot Disease and treatment representation example 4: Treatment: Oil: Kottamchukkadi



Mosaic plot: Each box is one disease, the selected disease is marked in Green colour, smaller boxes inside each disease display one intervention each, Data version: 2011 to Oct 2017

https://public.tableau.com/shared/48CRZ7B54?:display_count=y&:origin=viz_share_link



The Disease count tab provides the information on total number of different diseases for which the intervention was prescribed. There are 62 distinct diseases.

TreeMapDisMed TreeMapDisMed-Parameter Medicine-Count Medicine-List Disease-Count Disease-List Dashboard 1							
Medicine-List		SubDisMed OilKottamchukkadi ▾					
Med02	Code02	totpatmed	totpatdis	n	percdis	percmed	
OilKottamchukkadi	RMSD.V2.63.Vaatavyaad..	4604	4160	1506	36.2	32.71	Abc
	RMSD.V2.23.Vaatavyaad..	4604	2476	1274	51.45	27.67	Abc
	RMSD.S13.11.Sthaanabh..	4604	942	490	52.02	10.64	Abc
	RMSD.V2.0.Vaatavyaadhi..	4604	2822	396	14.03	8.6	Abc
	RMSD.S14.21.Sthaanabh..	4604	1227	326	26.57	7.08	Abc
	Metabolic:P5.0.Prameha..	4604	1456	324	22.25	7.04	Abc
	RMSD.A2.0.Aamavaata..	4604	790	205	25.95	4.45	Abc
	RMSD.A3.0.Abhigataja S..	4604	703	78	11.1	1.69	Abc
	RMSD.S13.19.Sthaanabh..	4604	120	65	54.17	1.41	Abc
	RMSD.V1.0.Vaatarketa..	4604	506	62	12.25	1.35	Abc
	RMSD.V2.36.Vaatavyaad..	4604	361	51	14.13	1.11	Abc
	RMSD.S14.0.Sthaanabhe..	4604	508	50	9.84	1.09	Abc
	RMSD.V2.9.Vaatavyadhi..	4604	203	44	21.67	0.96	Abc
	Metabolic:M1.0.1.Medoro..	4604	590	42	7.12	0.91	Abc
	Metabolic:M2.0.Madhum..	4604	1395	40	2.87	0.87	Abc
	RMSD.S14.7.Sthaanabhe..	4604	227	35	15.42	0.76	Abc
	RMSD.S14.17.Sthaanabh..	4604	189	26	13.76	0.56	Abc
	RMSD.V2.75.Vaatavyaad..	4604	162	20	12.35	0.43	Abc
	RMSD.V2.30.Vaatavyaad..	4604	78	18	23.08	0.39	Abc
	Metabolic:S16.0.Sthauly..	4604	911	18	1.98	0.39	Abc
	RMSD.S15.42.Sthaanabh..	4604	73	15	20.55	0.33	Abc
	RMSD.S15.36.Sthaanabh..	4604	75	14	18.67	0.3	Abc
	RMSD.S15.31.Sthaanabh..	4604	86	13	15.12	0.28	Abc
	RMSD.V2.77.Vaatavyaad..	4604	55	11	20	0.24	Abc
	Metabolic:M1.0.0.Medoro..	4604	139	11	7.91	0.24	Abc
	Metabolic:P5.4.Prameha ..	4604	140	9	6.43	0.2	Abc
	RMSD.S14.3.Sthaanabhe..	4604	79	8	10.13	0.17	Abc
	RMSD.V2.68.Vaatavyaad..	4604	49	7	14.29	0.15	Abc
	RMSD.S14.24.Sthaanabh..	4604	29	7	24.14	0.15	Abc
	RMSD.S14.18.Sthaanabh..	4604	34	7	20.59	0.15	Abc
	RMSD.V2.47.Vaatavyaad..	4604	24	6	25	0.13	Abc
	RMSD.S10.4.Stambha -Gr..	4604	31	6	19.35	0.13	Abc
	RMSD.A2.3.Aamavaata ..	4604	30	5	16.67	0.11	Abc
	Metabolic:M1.0.2.Medoro..	4604	188	5	2.66	0.11	Abc
	RMSD.V2.42.Vaatavyaad..	4604	52	4	7.69	0.09	Abc
	RMSD.S14.4.Sthaanabhe..	4604	68	4	5.88	0.09	Abc
	RMSD.S14.13.Sthaanabh..	4604	45	3	6.67	0.07	Abc

A detailed list of diseases with the total patients prescribed with the treatment as well as total patients suffering from different diseases are displayed.

Figure 3-35: Cross tabulation of prescribed treatments and disease group by gender Example 1

MedType_DisType

Medicine Name	Distype / Patient Gender						Grand Total
	Metabolic		RMSD		Total		
	F	M	F	M	Total		
Balarishtam					29	10	39
Balarishtam 450 ml	3	2	143	80	223	225	
Balarishtam 450 ml [sna]					16	9	25
Balarishtam 450 ml KPL					13	9	22
Balarishtam 450 ML SNA					9	5	14
Balarishtam 450 ml Vaidya	3	2	158	122	280	282	
Balarishtam [KPL]					2	1	3
Balarishtam [AVP]	11	8	353	256	609	616	
Balarishtam Rajah					1	1	2
Balarishtam(Arya)	1	1	24	12	36	37	
Grand Total	17	13	677	465	1,142	1,154	

Cross tabulation of medicine, disease type and patient gender, Medicine name: source data collected on Case Report Form, Distype: Metabolic and RMSD groups derived in the analysis dataset. Patient Gender: source data. Only 30 patients having metabolic diseases were prescribed the medicine whereas 1,142 patients with RMSD were prescribed. This reflects the ayurvedic principle of who should be prescribed any aristham. Data version: 2011 to Oct 2017

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

Figure 3-36: Cross tabulation of prescribed treatments and disease group by gender Example 2

MedType_DisType

Medicine Name	Distype / Patient Gender						
	Metabolic			RMSD			Grand Total
	F	M	Total	F	M	Total	
Abhrak bhasma (Dooth) 2gm				2	1	3	3
Abhraka Bhasma 1g (Sahasraputi) Dhoot					1	1	1
Abhraka Bhasma 5g [101] (AVP)	3	2	5	1	2	3	8
Abhraka Bhasma AVP	3	2	5	2	1	3	8
Abhraka Bhasma Dooth-100Butas	1	2	3	2	6	8	10
Abhrakam bhasma				2		2	2
Apamarga kshara Bhasma					1	1	1
Godhanti Bhasma 50gm [SDM]	1		1	16	4	20	20
Hajrulayahda Bhasma (Baidyanath)	1	5	6	3	8	11	17
Heeraka Bhasma 100mg [Dooth]				2		2	2
Kanmada Bhasma(10gm) (AVP)	4	4	8	2	2	4	11
Kashis Bhasma (Baidyanath)				1		1	1
Pravala bhasma	1		1	2	2	4	5
Pravala Bhasma Tab Dooth	1		1		1	1	2
Pravala Bhasmam (AVP)	1		1	1		1	2
Pravala Bhasmam Cap - 100				10	2	12	12
Pravala Bhasmam Cap (Kottakkal)	2		2	98	35	133	133
Rajatabhasma Cap (Kottakkal)	1		1				1
Roupya (Rajatha) Bhasma					1	1	1
Roupya (Rajatha) Bhasma 5 gm (Dooth)				2	1	3	3
Sameerapannaga Bhasma (Dooth)				2		2	2
sanka bhasmam				2		2	2
sankha bhasma				1	1	2	2
Sankha Bhasma(AVP)	1			1		1	1
Shilajith Bhasma	1	2	3				3
Shuddha Shilajit Bhasma		1	1				1
Tankan Bhasma 10gm		1	1	1	1	2	3
Tankan Bhasma (Baidyanath)		2	2	3	2	5	7
Tapyadi Loha Bhasma (Dooth)	1		1	3	1	4	4
Trivanga Bhasma (Dooth)	1	1	2				2
Vanga Bhasma 5g (Dhoot)	1	2	3				3
Yashada Bhasma 5gm (Dooth)	3	6	9	10	12	22	28
Grand Total	24	30	54	161	82	243	287

Cross tabulation of medicine, disease type and patient gender, Medicine name: source data collected on Case Report Form, Distype: Metabolic and RMSD groups derived in the analysis dataset. Patient Gender: source data. Only 287 (1.5%) patients have been prescribed bhasma. This reflects the ayurvedic principle of using bhasma based treatment wisely. Data version: 2011 to Oct 2016

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

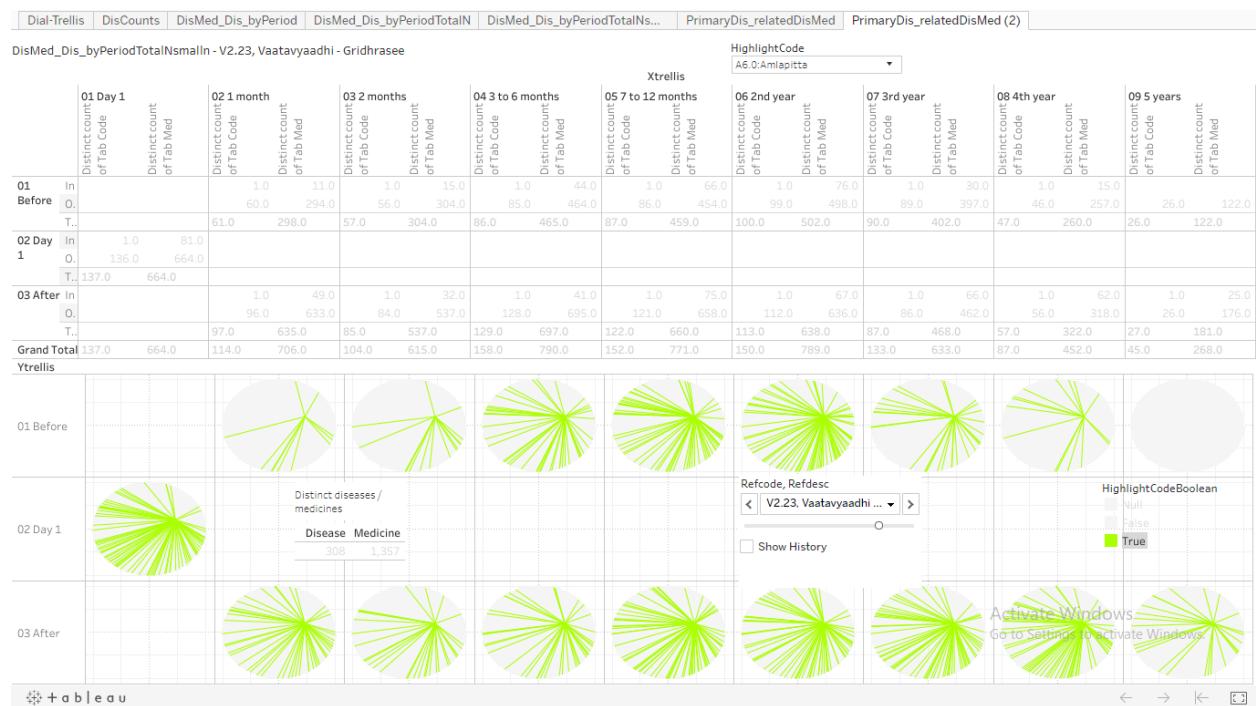
Table 3-2: Summary statistics and t-test for bhasma usage

Overall (N=514)	welch Two sample t-test
Pre_Bhasma	
Mean (SD) 14.8 (14.0)	
Median [Min, Max] 9.00 [2.00, 111]	
Post_Bhasma	
Mean (SD) 10.6 (14.5)	
Median [Min, Max] 6.00 [0, 89.0]	

Patients who were prescribed at least one “bhasma” treatments are summarized, Pre_Bhasma: Duration of treatment before the 1st bhasma treatment, Post_Bhasma: Duration of treatment after the 1st bhasma treatment, Data version: 2011 to Oct 2017

https://github.com/coursephd/PostgreSQL/blob/master/110_rasa_aushadhi_analysis.R

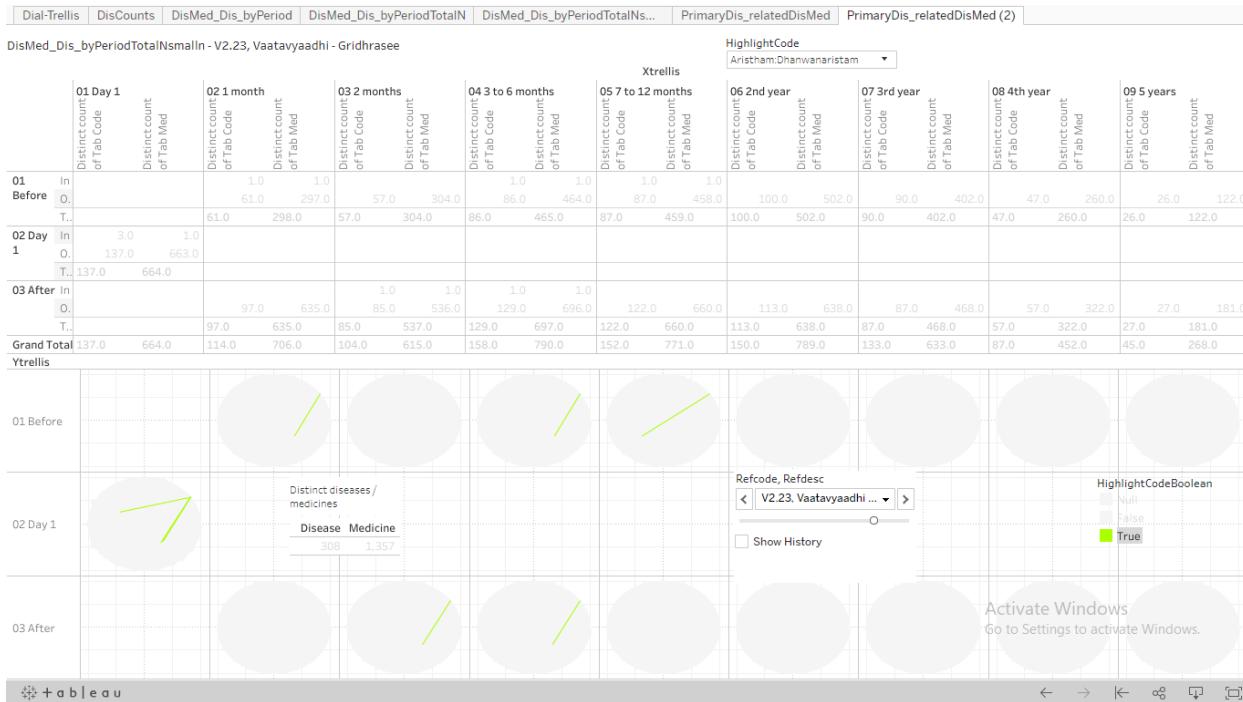
Figure 3-37: Circular view: Co-occurrences of disease – disease Example 1



Example 1: Disease: A6.0: Amavaata and Reference Disease: V2.23: Vaatavyadhi – Gridhrasee, Upper section: Pre and post time windows, count of distinct diseases and count of distinct medicines prescribed at the given time point. Lower section: 1st row represents the co-occurrence of disease – disease and / or disease – treatment before day 1 of the reference disease. Middle row: On day 1 count of distinct diseases and count of distinct medicines prescribed. Last row represents the same co-occurrence data after day 1 of the reference disease. Green bars inside a circle show co-occurrence of chosen disease – disease and / or disease – treatment combination, Data version: 2011 to Oct 2017

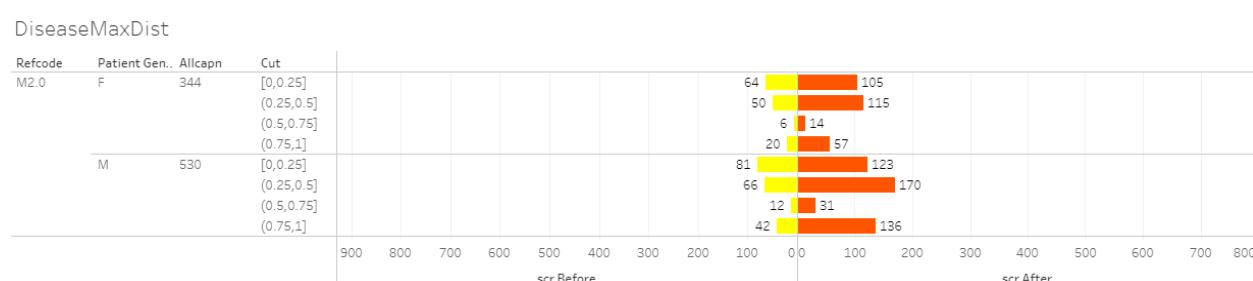
https://public.tableau.com/views/085_dis_count_edges_3rd_byPeriod02try/PrimaryDis_relatedDisMed2?:language=en&:display_count=y&:origin=viz_share_link

Figure 3-38: Circular view: Co-occurrences of disease – treatment Example 2



Example 2: Treatment: Arishtam:Dhanwanaristham and Reference Disease: V2.23: Vaatavyadhi – Gridhrasee, Upper section: Pre and post time windows, count of distinct diseases and count of distinct medicines prescribed at the given time point. Lower section: 1st row represents the co-occurrence of disease – disease and / or disease – treatment before day 1 of the reference disease. Middle row: On day 1 count of distinct diseases and count of distinct medicines prescribed. Last row represents the same co-occurrence data after day 1 of the reference disease. Green bars inside a circle show co-octene of chosen disease – disease and / or disease – treatment combination, Data version: 2011 to Oct 2017
https://public.tableau.com/shared/94DNM35MQ?:display_count=y&:origin=viz_share_link

Figure 3-39: Pre and Post distance analysis for disease: M2.0: Madhumeha

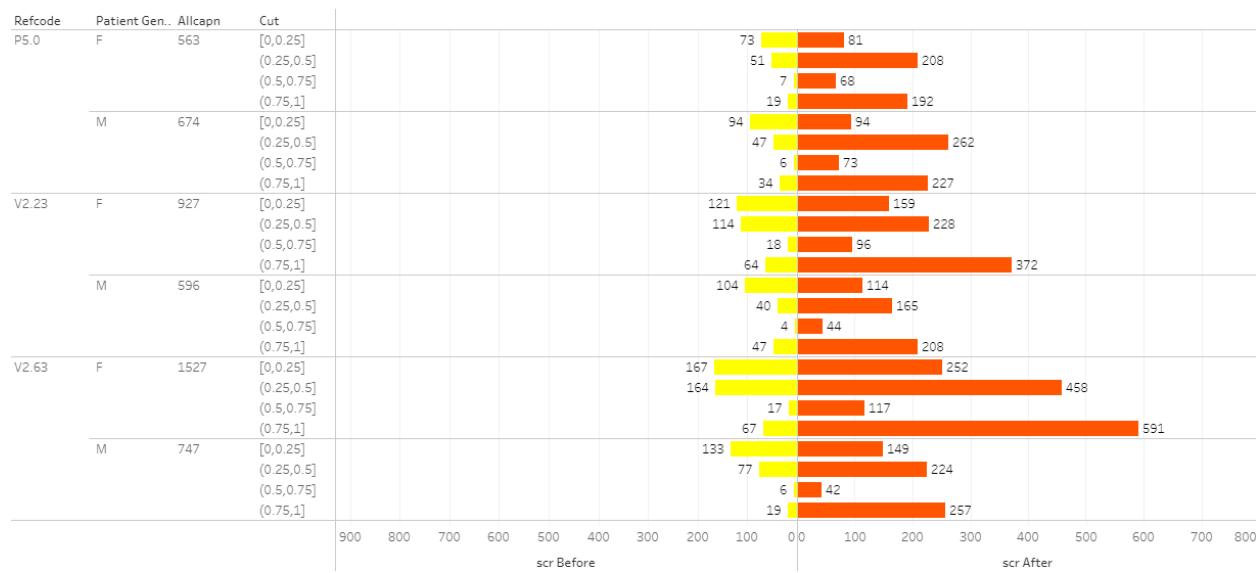


Refcode: reference disease, Patient gender, Allcapn: Total number of patients in each of the categories, Cut: Jaccard distance, scr Before: patients falling in a particular category before day 1 of reference disease, scr After: patients falling in a particular category after day 1 of reference disease, Data version: 2011 to Oct 2017

https://public.tableau.com/views/DistanceMeasuresTimePeriod-086prgm/DiseaseMaxDist?:display_count=y&:origin=viz_share_link

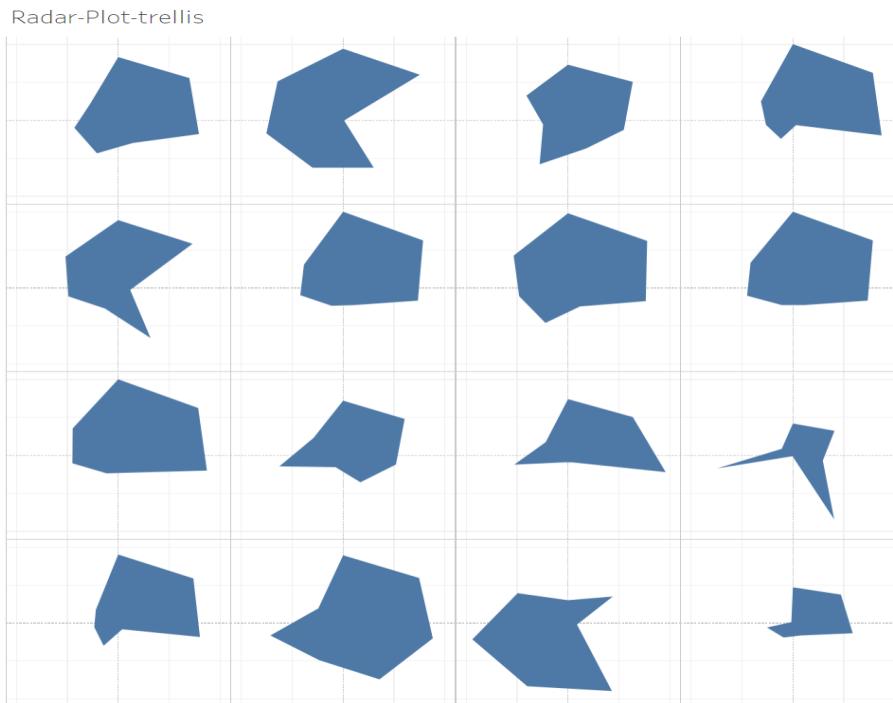
Figure 3-40: Pre and Post distance analysis for medicines given for diseases: P5.0, V2.23, V2.63

DiseaseMaxDist



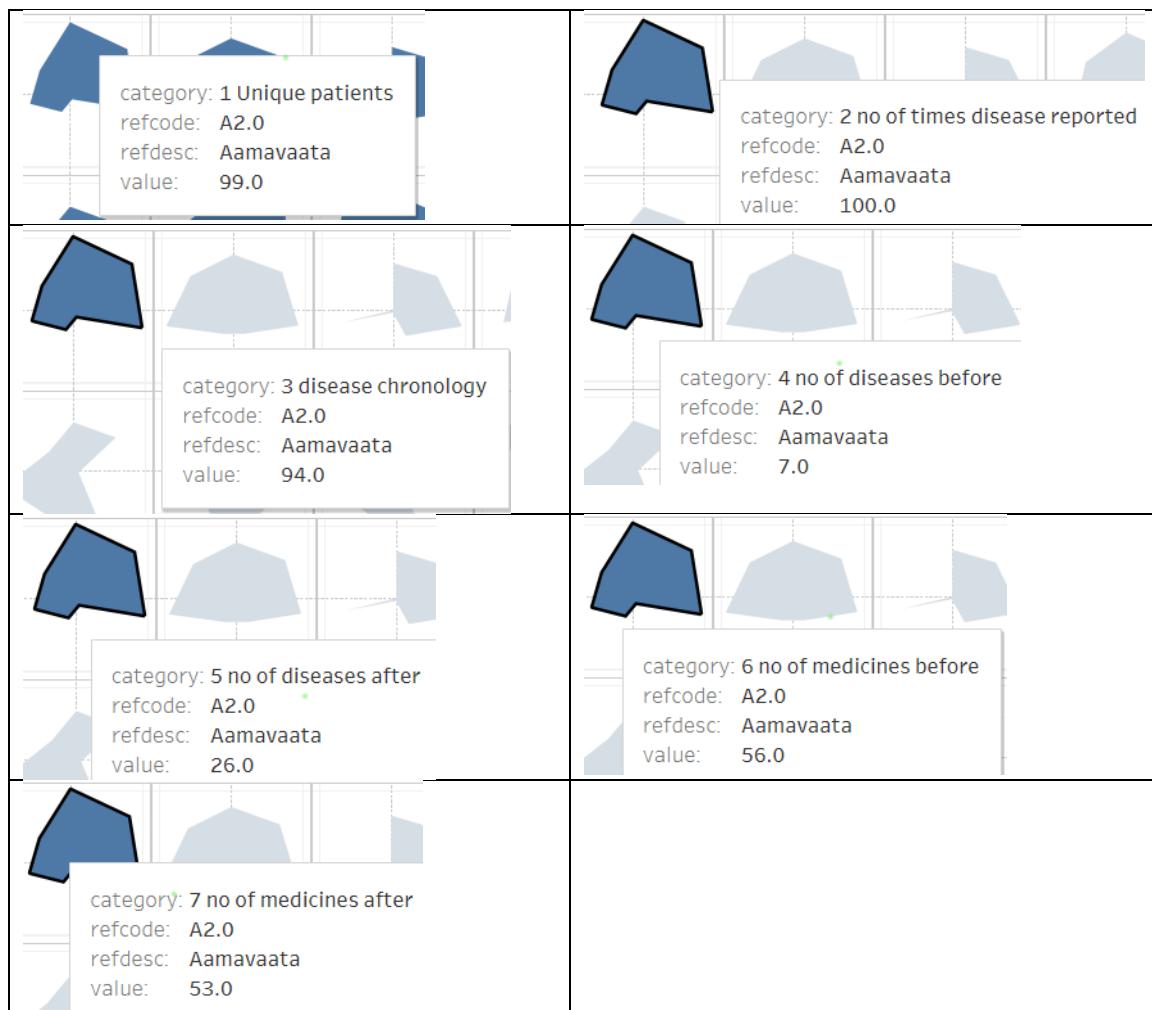
Refcode: reference disease, Patient gender, Allcapn: Total number of patients in each of the categories, Cut: Jaccard distance, scr Before: patients falling in a particular category before day 1 of reference disease, scr After: patients falling in a particular category after day 1 of reference disease. This distance calculation was done on the basis of trajectory of prescribed treatments, Data version: 2011 to Oct 2017.

Figure 3-41: Radar plot



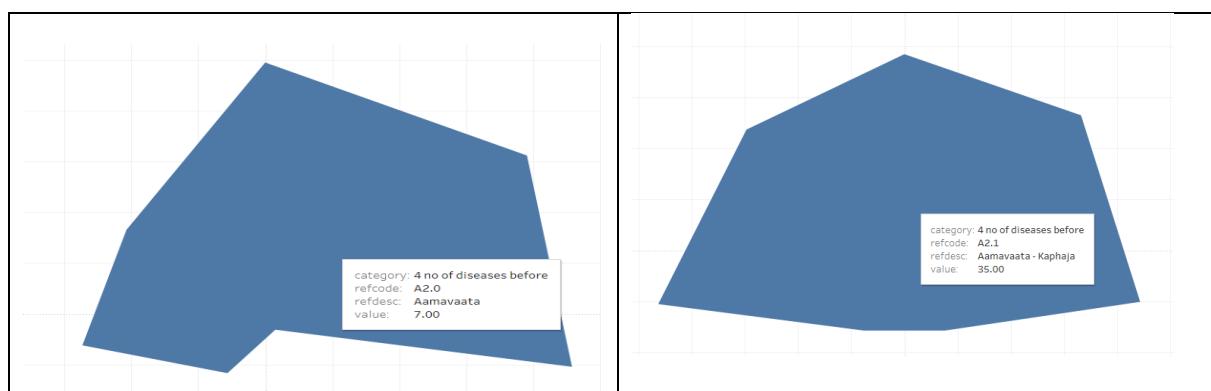
Radar plot showing multiple diseases displayed side by side.

Example 1: Multidimensional view a single disease: A6.0: Aamavaata



7 parameters displayed on one plot: (1) Unique patients, (2) Number of times disease reported, (3) Disease chronology, (4) Number of diseases reported before the reference disease, (5) Number of disease reported after the reference disease, (6) Number of medicines prescribed before the reference disease, (7) Number of medicines prescribed after the reference disease, Data version: 2011 to Oct 2017 https://public.tableau.com/shared/GSRS5P4DF?:display_count=y&origin=viz_share_link

Example 2: Multiple disease comparison



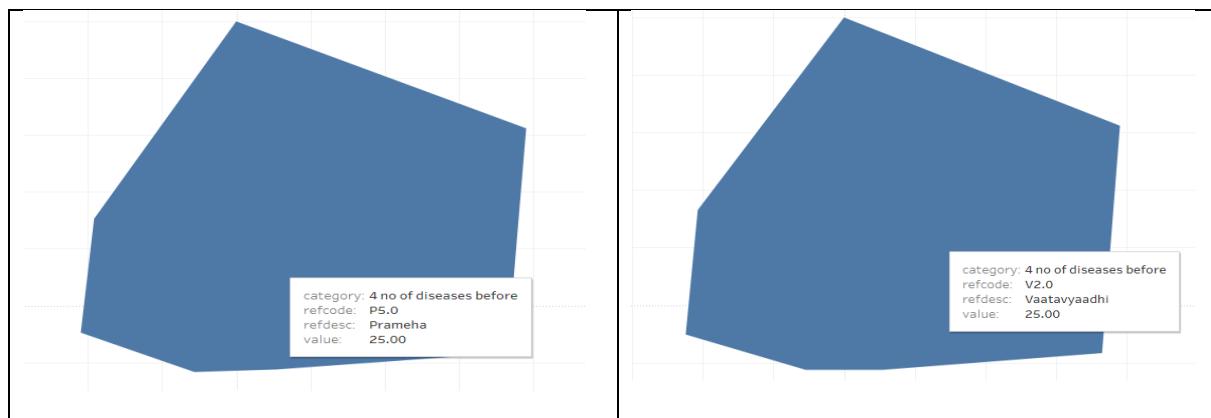
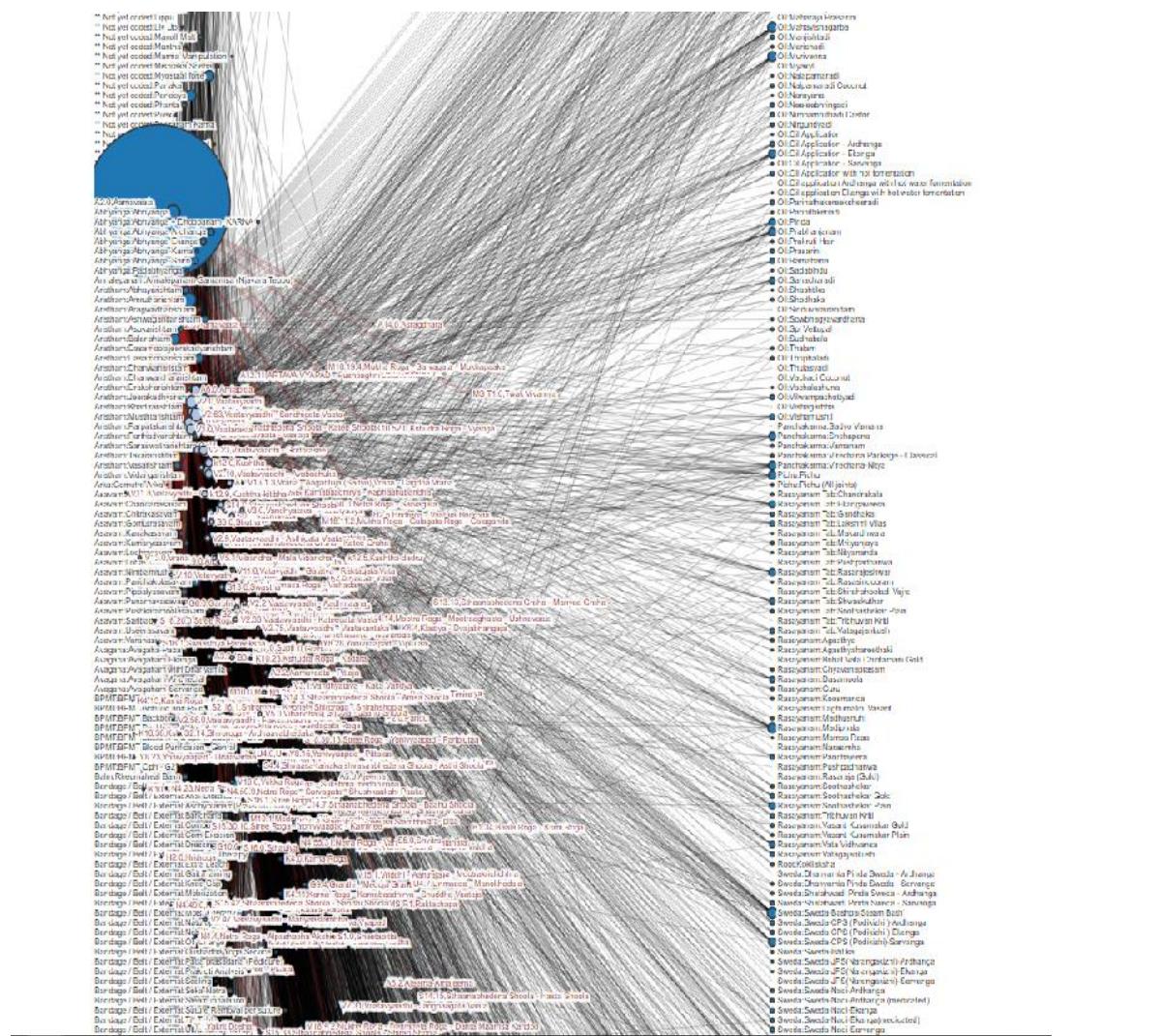


Figure 3-42: Dynamic bubble plot: Example 1: Disease: A6.0: Amavaata



This dynamic bubble plot shows relations between diseases and medicines, the bubble plot size is based on number of unique patients, Data version: 2011 to Oct 2017, Amavaata [https://coursephd.github.io/nodediagram/A2_0/]

4 Discussions

4.1 Converting real life clinical data into analyzable format

The TDU / I-AIM team should be congratulated first before any discussion to create an electronic database right from the inception of the hospital. This foresight has allowed us to have significant amount of data. There are a lot of learnings from this exercise which can be beneficial to many institutes and hospitals.

Conversion of real-life clinical data from an individual data point to a logical dataset was done. Logical relationships were established post inspection of the datasets and the columns. Relational datasets were identified. How data was captured and stored was observed (data formats). Some shortcomings and errors in the data were seen and noted (missing data, inconsistent values, or unresolved duplicates). This exercise of understanding technical architecture from “an end user point of view” will help in running analysis of various types. If data generation for future use is one of the top priorities for the hospital, then there should be a project plan put together, appropriate steps should be taken to plug the existing gaps.

4.2 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.

Table 4-1: Proposed vital sign data structure

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 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 [79]

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, 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.

Figure 4-1: Treatment principles defined in different texts

Treatment	Ch	Su	AH	BP	YR	HS	BS	CD
<i>Snehana</i>	-	-	-	-	-	+	+	+
<i>Swedana</i>	-	-	-	-	-	+	-	+
<i>Vamana</i>	-	-	-	+	-	-	-	+
<i>Virechana</i>	-	-	-	+	-	-	-	+
<i>Niruha Basti</i>	+	-	-	-	-	-	-	-
<i>Anuvasana Basti</i>	+	-	+	+	+	-	+	+
<i>Siravyedha</i>	+	+	+	-	+	-	-	+
<i>Raktamokshana</i>	-	-	-	-	-	+	+	-
<i>Agnikarma</i>	+	-	+	-	+	+	-	+
<i>Shastrakarma</i>	-	-	-	-	-	-	-	+
<i>Ch - Charaka; Su - Sushruta; A.H.- Astanga Hrudaya; B.P. - Bhava Prakash; Y.R. - Yoga Ratnakara; H.S. - Haritha Samhitha; B.S - Bhela Samhitha; C.D. - Cakradatta</i>								

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 used medical coding dictionaries for coding medical terms are MedDRA and WHO DDE [76] [77]. A few examples of medical dictionaries are:

- COSTART: Coding Symbols for Thesaurus of Adverse Reaction Terms
- ICD xx CM: International Classification of Diseases xx Revision Clinical Modification [78]
- MedDRA: Medical Dictionary for Regulatory Activities [76]
- WHO-ART: World Health Organization Adverse Reactions Terminology [77]
- WHO-DDE: World Health Organization Drug Dictionary Enhanced [77]
- ACD: Ayurvedic Classification of Diseases
- NCI: National Cancer Institute Code list [79]
- LOINC: Logical Observation Identifiers Names and Codes standards [80]
- Any other dictionaries, as recommended by AYUSH

Classification and Sub-classification of the Doshas / Diseases: Almost each researcher understands that research results are as good as the data using which the conclusions are drawn. Most scientists do not receive guidance in methods for controlling the quality of research data which is fundamental to clinical research. An exhaustive list of all possible diseases should be created, and a checklist of disease classification and sub-classification should be maintained so that the doctors based on their judgment can classify the dosha appropriately and use the recommended disease classification term. This should help in reducing inconsistency and disparity in reporting. The problem should be split into operational and scientific components. Identify fields which would require coding: Diagnosis codes description, Compliant, Drug description. Operational steps to be taken as follows:

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

Table 4-2: Proposed idea for clinical coding timetable

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 team of 3-4 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.

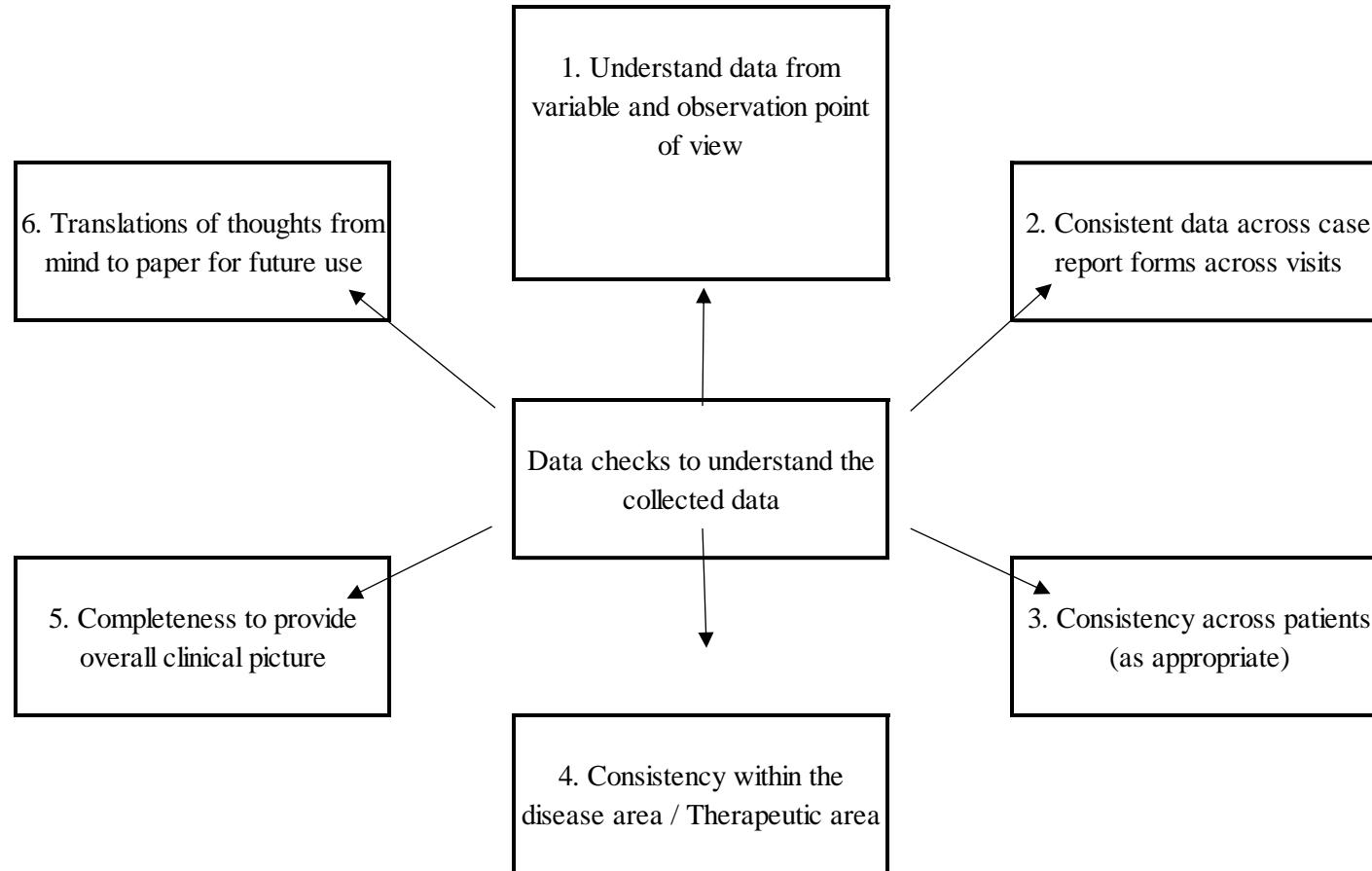
Overall data standardization: Overall the hospital data is not captured in a standardized format as explained in the above points. Create standardized CRF pages for consistent and correct data capture. Already established standards like Clinical Data Interchange Standards Consortium (CDISC) or International Organization for Standardization (ISO) standards can be implemented [81] [82]. 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 good 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 latest ICD version, 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 (Figure 4-2), summarizes the cycle of understanding the hospital data so that a meaningful interpretation can be arrived at. 6 steps provide a way of generating very good quality data: (1) Understand the data from variable and observation point of view, (2) Collect consistent data across case report forms across visits, (3) Maintain consistency across patients, (4) Maintain consistency across disease areas, (5) Strive to maintain completeness to provide overall clinical picture, and (6) These steps should enable translation of thoughts from mind to data for future use.

Figure 4-2: Data understanding from an observation – patient – disease to a clinical picture



Improvements to the system architecture: Usually any team in any organization 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 Electronic Medical Record (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 capitalization. Create guidelines and instruct accurately when to use caps and otherwise. In case all caps are used, ensure appropriate warning message or suggestion is generated to alert the end user.
- Abbreviations: Prepare a catalog of all allowed abbreviations along with their meanings. If possible, create inbuilt checks in the building forms to avoid incorrect abbreviations.
- Workflows: Evaluate and monitor the workflows to check the effectiveness on an ongoing basis. 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 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.
- Regular training: 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 can 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. 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

Variable classification analysis: Review of the database suggested that the case report form completion was carried out by different doctors differently giving rise to differences in the way the data was captured. This lack of documentation should be addressed. Hospital management,

treating doctor, researchers, and insurance companies could be the key stakeholders benefitting from these improvements.

4.3 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 (Figure 3-6, Figure 3-7). 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 (Figure 3-8). 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 (Figure 3-11). 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 undoubtedly 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 (Figure 3-9). 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 (Figure 3-12). The low rate of reporting of some of the diseases may explain the natural variations or may reveal inconsistent labelling of the diseases (Figure 3-15). 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 (Figure 3-13).

4.4 Studying diagnostics and interventions

The ACD and ICD mapping exercise shows that the current hospital data demonstrates all the types of diseases being catered to at the hospital. Large spectrum of diseases getting treated at the hospital. This provides insights into the health seeking behaviour of patients. Additionally, this data can be used by insurance companies and policy makers to strengthen ongoing efforts. If the ICD Code can be included in the data collection, then this correlation analysis can be done easily. The ICD code getting populated using the medical expertise will be more reliable than the current analysis (Figure 3-17, Figure 3-18, Figure 3-19, Figure 3-20).

The Prakriti data was not available for all patients across all visits. This points to shortcomings in the data collection methods (Figure 3-21). This analysis shows natural variations in the diseases getting treated. It should be studied by treating physicians to take a deep dive into the data. The process of data collection needs to be looked at and improved. Would it be possible to create an online tool to generate prakriti information? Can this data be collected before a patient goes in for doctor's consultation? Can this data field be made a mandatory field so that there is no missing data generated? If prakriti derivation is a complex process Vikriti or Dosha current dominance must be looked at. Tag the treatment or formulations as -kara and -hara e.g., Pippali is Kaphahara and Pittakara.

Co-morbidity analysis can be used to understand the disease clustering. Which disease(s) cause(s) the other disease(s) to manifest, which disease(s) could be precursor to subsequent disease(s). This analysis can be used to validate the existing hypothesis. This analysis could be further enriched for predictive abilities – turning this into a possible disease preventive tool. Some examples like prameha (causing many diseases for both the genders), pandu roga (mainly reported by females with many disease, and relatively low numbers reported by males), sandhigata vata (reported by more females), etc. have shown that meaning of shlokas can be shown in the modern data format. This type of exercises can be carried out with help of ayurvedic experts tree (Figure 3-22, Figure 3-23, Figure 3-24, Figure 3-25, Figure 3-26).. Explain the above-mentioned examples with pictures

Treatment and disease analysis at individual patient level: This analysis at individual patient level shows life journey of each patient. This may help in understanding the severity of the disease, co-morbidities and the number of medications prescribed to treat the condition. This can also provide an overview of the practicing physician's style of treatment and may be help draw parallels in treating medical conditions. If a new disease is reported and new treatments are added, then it is understood that this is a part of treatment regimen. But if a treatment is reduced then would it be considered as a part of treatment regimen or would it be considered as a removal due to side effect? This cannot be ascertained without an ayurvedic clinician's opinion. If a new disease is reported and if no new treatment is added, then also it raises some questions? Is it as per treatment protocol or are the existing treatments sufficient to cover this new imbalance? More detailed discussions with ayurvedic clinicians will help in understanding this analysis. This may give rise to better ways of collecting the data (Figure 3-27, Figure 3-28, Figure 3-29).

Area graph representation of diseases provides information about 800+ diseases in very short space. Disease patterns are interesting due to following reasons: Diseases vary seasonally, diseases are experienced differently by gender, and it gets shown easily by looking at the distributions. This view is very useful for both operational excellence as well as clinical judgment. The interactive nature of visualization allows for real time subset of diseases. One of the 4 diseases displayed has very few patients compared to other 3 diseases showing different nature of diseases. Another interpretation could be that the disease shown with very low frequency may not be treated by very regularly by Ayurvedic treatments (Figure 3-30).

Treatment and disease analysis at summary level: Mosaic plot and cross tabulation analysis: the Patient Report form or Case Report form captures diseases reported on a particular visit along with treatments and services prescribed to a patient. Due to the nature of the CRF page, multiple diseases and treatments are captured on the same visit. This creates many-to-many relationships which makes it difficult to identify the disease treatment relationship. Even though this challenge exists, the data at a summary level provides good view on treatment and disease relationship. The cross tabulations of balaristham and bhasma provide additional evidence of how these analyses can be used to validate facts and / or generate new concepts. Traditionally bhasmas of any kind are prescribed in very limited quantity and same is reflected in observed data. Naming convention and spelling correctness need to be considered while capturing data in future. The clinical utility of this analysis was shown (Figure 3-31, Figure 3-35, Figure 3-34, Figure 3-36). The t-test at 5% significance level shows statistically significant difference between duration of treatment before bhasma treatment and duration of treatment after bhasma treatment. The study was not powered to detect any specific difference in treatment duration, so the p-value and significance should be interpreted cautiously. How should one interpret ~15 days vs. ~11 days of pre and post bhasma treatment, would it be considered clinically meaningful? These discussions with experts will provide more ideas about these plain numeric observations.

Interpretation from additional Disease – treatment analysis with pre and post visit window approach are as follows: in circular data representation many green lines means that there is a greater chance of diseases reported by patients, there is a greater chance of a medicine prescribed for a disease. If there are very few lines then the combination is clinically not meaningful or if it is meaningful then it is a very rare combination which needs to be studied further (Figure 3-37, Figure 3-38). On a single page there are multiple dimensions of the disease – disease and / or disease – treatment combinations are shown.

Distance score-based analysis reveals the following observations: More number of patients with Jaccard distance closer to 1 was seen for the Post reference day 1 period. This could be pointing to similar biological activity caused by a particular disease. This could be a very important finding from this analysis (Figure 3-39). In the medicinal display the similarity scores are lower as compared to that for the disease trajectories. Which implies that most of the prescribed treatments are dis-similar for both the periods. It is observed that around 50% of treatments could form the base of treatment regimen and could be same for the patients. The remaining part of the treatment regimen is driven by individual patient characteristics. The before and after medicine trajectories would show such underlying data. E.g., for M2.0, there are very few patients having

distance above 0.5 for both genders (Figure 3-40). This analysis should be executed using other mathematical distances to understand the consistency of results. If the disease classification and treatment tagging in the underlying data is improved then we should be able to see much better results, with lesser confounding effect.

Radar plot for multiple diseases is shown next to each other. This is showing massive amounts of information immediately. Differing shapes provide differences reported in the data and an easy way of identifying differences. If there is additional data made available in a structured format, then these parameters could also be added on the radar plot. This radar + trellis combination provides a more powerful tool to visualize large amounts of data on a single page (Figure 3-41).

5 Conclusion

The introduction of this thesis outlines the need to undertake such a study, by providing 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 asks 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 can 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 many 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. While studying both the structure and the content of the hospital database, it was observed, that standardization of database along with effective curation is needed. This type of data could provide a gold mine of information which when summarized could lead us to a lot of supportive evidence. 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. Public health domain can be benefitted by disease surveillance and population health. It also provided actionable inputs to hospital management, practicing doctors and for research publications.

Diagnostics and interventions section show multivariate relations between diseases and interventions. Comorbidities as well as combinations of interventions showed the complex clinical decision-making process. The disease and treatment comorbidity analysis was performed and presented using a variety of plots and heatmaps. This analysis showed how individual observations can be transformed into meaningful stories at hospital.

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

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 figure below (Figure 5-1) explains the different contexts that play a vital role in how the RWD shapes up and supports in providing key outcomes.

Figure 5-1: 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				Longitudinal data Co-morbidities and Cost effectiveness
Methodology context	Low adherence	Quality of life outcome				
	Confounding and Population homogeneity	Economic outcome	Stakeholders: Regulatory Authorities, Policy Makers, Government and Payers			
	Un-blinded treatment and Treatment switch	Primary / Secondary data	Retrospective / Prospective study	Big data, large sample size	Social media	Individual practice Hospital EHRs Real life data, clarity of treatment impact and AEs
	Operational challenges	Comparable data	Patient level data access	GDPR and Anonymization	Incomplete data	More data available on drug and life style interaction
				Data context		

Due to the above-mentioned outcomes, the following contributions will be possible:

1. Contribution to Public health data creation based on large data at our disposal which is not marred by artificial boundaries imposed on patient disease conditions, treatments prescribed
2. Generate Real World Evidence (RWE) to supplement classical understanding which is rooted within the community
3. Make recommendations to the practitioners for standardized way of data collection, analysis and reporting which will support future RWE studies
4. Understand the hidden wealth of data for Transdisciplinary expansion of thoughts
 - a. Shlokas to diagrams – generating supplementary material to classical texts, thus creating numerical evidence for the shlokas (shastra tatva, vyavahar)
 - b. Sustainable treatment solutions for non-communicable diseases readily available
 - c. Thought provoking work to generate new needs, unconventional use of the data
 - d. Expand the use of modern IT solutions like IT infrastructure, electronic health records, cloud, etc. within Ayurvedic area where appropriate – Ayur IT solutions.
 - e. Take advantage of freely available cutting-edge software(s) to create new approaches
 - f. Introduce statistical programming (Ayurdata analyst) as a tool to Ayurvedic area
 - g. Generate viable financial model for making data available for insurance

A lot of work carried is out by health authorities, pharmaceutical companies, and nonprofit organizations. This section provides references to ongoing work. Some of these resources could be used to become a world class data generator:

Clinical Data Interchange Standards Consortium (CDISC) is a global nonprofit charitable organization with administrative offices in Austin, Texas, with many people contributing across the world. CDISC brings experts together to create and advance data standards. This allows for accessibility, interoperability, and reusability of data for competent research that has greater impact on global health [81]. Many of the leading health authorities use the standards developed by the CDISC teams in various parts of drug applications.

TransCelerate BioPharma Inc. is a nonprofit organization with a mission to collaborate across the global biopharmaceutical research and development community to identify, prioritize, design, and facilitate the implementation of solutions designed to deliver high quality new medicines. They have many open-source solutions which could be used to improve delivery model [83].

Based on the 21st century act, 2016, the US FDA has created a regulatory framework for the Real-world data (RWD) and real-world evidence (RWE) which are playing increasing role in the health care decisions [84], [85]. The European Medicines Agency (EMA) has established a center to provide timely and reliable evidence from real world healthcare databases on the use, safety, and effectiveness of medicines for human use, including vaccines, across the European Union (EU). This capability is called the Data Analysis and Real-World Interrogation Network (DARWIN EU®) [86].

These resources provide a lot of material to enhance overall understanding and allows researchers to be compliant with the regulatory requirements.

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. There is a need to have a profession of a “Statistical programmer” or a “clinical programmer” or an “Ayurdata expert”. This role can contribute to database development, data collection, data cleaning aspects, creating analysis ready datasets, and to finally 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. Theoretical ideas can be converted into practical interactive visualizations and interactive analyses using multiple technologies. These will help convert individual data observations into summaries then into stories thus enabling knowledge generation.

Ayurdata expert can contribute to creating documents for medical journalism, medical education, medical marketing of healthcare products, publications, research documents, and regulatory documents by collaborating with other experts (Table 5-1). We believe that this is a pioneering effort within ayurvedic data analysis area.

Table 5-1: Different types of documents

Medical Reporting	Medical Teaching	Medical advertising of products	Publication
<ul style="list-style-type: none"> • Newspaper & magazine articles • Mostly for public • Written in simple, non-technical language 	<p>For doctors</p> <ul style="list-style-type: none"> • Textbooks, • Continued Medical Education programs, • Slide decks, • Online learning material <p>For Patients learning material</p>	<ul style="list-style-type: none"> • Promotional information for healthcare professionals • Product profiles • Brochures • Sales force training • Online learning material 	<ul style="list-style-type: none"> • Abstracts • Journal articles, case reports, review articles • Posters & presentations for scientific conferences

Research Documents	Regulatory documents #
--------------------	------------------------

<ul style="list-style-type: none"> • Research proposals • Clinical trial protocols • Investigators' Brochure • Informed Consent Documents • Study reports 	<ul style="list-style-type: none"> • Package Inserts • Patient Information Leaflets • Clinical study reports • Web synopses • Subject narratives <p>Common Technical Document (CTD) modules such as</p> <ul style="list-style-type: none"> • nonclinical and clinical overviews & summaries • expert reports • PK, Safety, Efficacy summaries 	Aggregate safety reports such as <ul style="list-style-type: none"> • Periodic Safety Update Reports • Periodic Adverse Drug Experience Reports • Annual safety reports • Policy papers
--	---	---

#: Presently, some of these documents are not applicable within Ayurvedic area

A by-product of this work: Work carried out for this thesis is typically done by a team of people, which I did by myself. In a mid to large sized pharmaceutical organization, this work is carried out by (1) Clinicians and statisticians designing clinical protocol, (2) Database development team creates database and data flow components, (3) Data management team reviews and cleans the data on an ongoing basis, (4) Statistical programming team and statisticians create the necessary Tables Listings and Figures for the analysis, (5) Finally a writing team generates Clinical Study Report / Publication, and last but not the least (6) IT team to handle various systems so that the data and information flow is managed appropriately.

As I got to handle many components myself during designing concepts, writing many programs, creating many interactive analyses, and finally writing thesis, I feel that I have learnt a lot more than I ever imagined.

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

Large spectrum of diseases getting treated at the hospital. Variability in the data: there are many treatments assigned to each disease, KVP classification of the disease is not documented properly due to various reasons – but the unique treatments are documented. e.g., for a disease there could be 1000+ treatments prescribed across timeframes, this is a misrepresentation of data yielding inaccurate scientific results. This can be sorted by improving the process of documentation.

Ongoing and future work:

- Updates to the existing analyses using most recent data
- Creation of similar analyses using data from multiple institutes
- SPMF analysis [Sequential Pattern Mining Library, based on computer science approach, using more than 100 algorithms]
- Episodic analysis, Survival analysis
- Many analyses carried out using Natural Language Processing approaches
- Additional visualizations using hospital data which was not used so far in this thesis

6 Appendix

6.1 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

6.2 Details of analysis dataset

Table 6-1: 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	<p>Derivation of Indian seasons based on the date variable for each visit:</p> <p># Add Indian rutus as new variables</p> <p># https://www.drikpanchang.com/seasons/season-tropical-timings.html?geoname-id=1277333&year=2010</p> <ul style="list-style-type: none"> • 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	<p>Source variable, no derivation needed:</p> <ul style="list-style-type: none"> • C - Cancelled • U - Condn. Unnecessary • Y -Conducted • N - Not Conducted • P - Partially Conducted
presc_type		Source variable, no derivation needed
medicine_name	Medicine name	<p>Source variable, no derivation needed</p> <p>Prescribed medicine names follow a certain predefined naming convention. Medicine name + Quantity + Producer's name are the details recorded for each prescribed medicine.</p>
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

		<ul style="list-style-type: none"> • Kashayam • Asavam • Aristham • Bhasma • Abhyanga • Cream • Rasayanam • Tablet / Gulika / Vati • ...
cat_id		Identification of categories
distype	Disease type	<p>Disease type as OTHER, RMSD, Metabolic</p> <ol style="list-style-type: none"> 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	<p>If a patient has reported any Metabolic disease at least once then that patient is given value Metabolic = 1, else Metabolic =0</p> <p>Metabolic disease group has 10 diseases (Refer 2.4.1.6.1)</p>
RMSD	RMSD	<p>If a patient has reported any RMSD disease at least once then that patient is given value RMSD = 1, else RMSD =0</p> <p>RMSD disease group has 97 diseases (Refer 2.4.1.6.1)</p>
combine	Metabolic RMSD	<ol style="list-style-type: none"> 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,

	Both	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.

Table 6-2: Metabolic and RMSD disease code and de-code

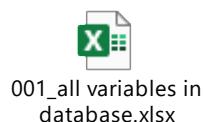
Code	Description	Distype
M10.0	Medoroga	Metabolic
M10.1	Medoroga - Sthula medho roga	Metabolic
M10.2	Medoroga - Sukshma medho roga	Metabolic
M2.0	Madhumeha	Metabolic
P5.0	Prameha	Metabolic
P5.1	Prameha - Krusha	Metabolic
P5.2	Prameha - Pidaka	Metabolic
P5.3	Prameha - Sthula	Metabolic
P5.4	Prameha - Upadrava	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	Sthaanabheda Graha	RMSD
S13.1	Sthaanabheda Graha - Anga Graha	RMSD
S13.11	Sthaanabheda Graha - Katee Graha	RMSD
S13.13	Sthaanabheda Graha - Manyaa Graha	RMSD
S13.14	Sthaanabheda Graha - Marma Graha	RMSD
S13.17	Sthaanabheda Graha - Paada Graha	RMSD
S13.18	Sthaanabheda Graha - Paarshva Graha	RMSD
S13.19	Sthaanabheda Graha - Prishtha Graha	RMSD
S13.20	Sthaanabheda Graha - Shiro Graha	RMSD
S13.22	Sthaanabheda Graha - Uro Graha	RMSD
S13.23	Sthaanabheda Graha - Vaak Graha	RMSD
S13.3	Sthaanabheda Graha - Gala Graha	RMSD
S13.5	Sthaanabheda Graha - Hanu Graha	RMSD
S13.6	Sthaanabheda Graha - Hrid Graha	RMSD
S13.7	Sthaanabheda Graha - Jaanugraha	RMSD
S13.8	Sthaanabheda Graha - Janghaa Graha	RMSD

S14.0	Sthaanabheden Shoola	RMSD
S14.11	Sthaanabheden Shoola - Guda Shoola	RMSD
S14.13	Sthaanabheden Shoola - Gulpha Shoola	RMSD
S14.14	Sthaanabheden Shoola - Hanu Shoola	RMSD
S14.15	Sthaanabheden Shoola - Hasta Shoola	RMSD
S14.16	Sthaanabheden Shoola - Hrid Shoola	RMSD
S14.17	Sthaanabheden Shoola - Jaanu Shoola	RMSD
S14.18	Sthaanabheden Shoola - Janghaa Shoola	RMSD
S14.19	Sthaanabheden Shoola - Kantha Shoola	RMSD
S14.21	Sthaanabheden Shoola - Katee Shoola	RMSD
S14.23	Sthaanabheden Shoola - Kukshi Shoola	RMSD
S14.24	Sthaanabheden Shoola - Manyaa Shoola	RMSD
S14.3	Sthaanabheden Shoola - Amsa Shoola	RMSD
S14.4	Sthaanabheden Shoola - Anga Shoola	RMSD
S14.5	Sthaanabheden Shoola - Anguli Shoola	RMSD
S14.6	Sthaanabheden Shoola - Asthi Shoola	RMSD
S14.7	Sthaanabheden Shoola - Baahu Shoola	RMSD
S15.28	Sthaanabheden Shoola - Nakha Shoola	RMSD
S15.31	Sthaanabheden Shoola - Paada Shoola	RMSD
S15.32	Sthaanabheden Shoola - Paarshni Shoola	RMSD
S15.34	Sthaanabheden Shoola - Parva Shoola	RMSD
S15.36	Sthaanabheden Shoola - Prishtha Shoola	RMSD
S15.41	Sthaanabheden Shoola - Sakthi Shoola	RMSD
S15.42	Sthaanabheden Shoola - Sandhi Shoola	RMSD
S15.43	Sthaanabheden Shoola - Skandha Shoola	RMSD
S15.44	Sthaanabheden Shoola - Snaayu Shoola	RMSD
S15.45	Sthaanabheden Shoola - Sphik Shoola	RMSD
S15.46	Sthaanabheden Shoola - Stanaanta Shoola	RMSD
S15.47	Sthaanabheden Shoola - Trika Shoola	RMSD
S15.48	Sthaanabheden 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 Vaatajaa	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 - Vaataadhistha Vaatarakta	RMSD
V2.0	Vaatavyaadhi	RMSD
V2.12	Vaatavyaadhi - Stabdhaagattra	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

6.3 All variables in the source database

Table 6-3: All variables in the source database



7 References

- [1] M. Masoud Y. Jaradat A. Manasrah Ahmad I. Jannoud, "Sensors of Smart Devices in the Internet of Everything (IoE) Era: Big Opportunities and Massive Doubts," *Journal of Sensors*, p. 26, 2019.
- [2] W. R. Bolislis F. Myriam T. C. Kühler, "Use of Real-world Data for New Drug Applications and Line Extensions," *Clinical Therapeutics*, vol. 42, no. 5, pp. 926-938, 2020.
- [3] AYUSH Minsitry Government of India, "AYUSH website," [Online]. Available: <http://dashboard.ayush.gov.in/>.
- [4] R. H. Singh, "Exploring issues in the development of Ayurvedic research methodology," *Journal of Ayurveda and integrative medicine*, pp. 91-95, 2010.
- [5] I. Krakau, "The importance of practice-based evidence," *Scandinavian Journal of Primary Health Care*, pp. 130-131, 2000.
- [6] B. Patwardhan, "Envisioning AYUSH: Historic Opportunity for Innovation and Revitalization," *Journal of Ayurveda and integrative medicine*, pp. 67-70, 2014.
- [7] G. Tillu TDU hospital staff, "Exploration of the TDU database," 2016 - 2018. [Online].
- [8] Pharmaforum, "Pharmaforum, bringing healthcare together," [Online]. Available: https://pharmaphorum.com/views-and-analysis/a_history_of_the_pharmaceutical_industry/. [Accessed 2020].
- [9] J. Greene S. H. Podolsky, "Reform, regulation, and pharmaceuticals--the Kefauver-Harris Amendments at 50," *The New England journal of medicine*, vol. 367, no. 16, p. 1481–1483, 2012.
- [10] V. Renjith, "Blinding in Randomized Controlled Trials: What researchers need to know?," *Manipal Journal of Nursing and Health Sciences*, vol. 3, no. 1, pp. 45-50, 2017.
- [11] D. S. Jones S. H. Podolsky, "The history and fate of the gold standard," *The Lancet: The Art of Medicine*, vol. 385, no. 9977, pp. 1502-1502, 2015.
- [12] National Library of Medicine USA, "ClinTrials.gov," [Online]. Available: <https://clinicaltrials.gov/>.
- [13] T. Boerma, "Public health information needs in districts," *BMC Health Serv Res*, vol. S12, 2013.
- [14] R. S. Evans., "Electronic Health Records: Then, Now, and in the Future," *Yearbook of medical informatics*, vol. Suppl 1, p. S48–S61, 2016.
- [15] V. R. Suvarna, "Real world evidence (RWE) - Are we (RWE) ready?," *Perspectives in clinical research*, vol. 9, no. 2, pp. 61-63, 2018.
- [16] S. Kamath G. Guyatt, "Importance of evidence-based medicine on research and practice," *Indian journal of anaesthesia*, vol. 60, no. 9, pp. 622-625, 2016.
- [17] G. Nahler, "good clinical research practice (GCRP)," in *Dictionary of Pharmaceutical Medicine*, Vienna, Springer, 2009.

- [18] J. Concato, "Randomized, Controlled Trials, Observational Studies, and the Hierarchy of Research Designs," *New England Journal of Medicine*, vol. 342, pp. 1887-1892, 2000.
- [19] J. Avorn, "In Defense of Pharmacoepidemiology — Embracing the Yin and Yang of Drug Research," *New England Journal of Medicine*, 2007.
- [20] A. Anglemyer H. T. Horvath L. Bero, "Healthcare outcomes assessed with observational study designs compared with those assessed in randomized trials," *Cochrane Database of Systematic Reviews*, 2014.
- [21] S. V. Ramagopalan A. Simpson C. Sammon, "Can real-world data really replace randomised clinical trials?," *BMC Med*, vol. 18, no. 13, 2020.
- [22] S. Rimpilainen, "A Review of Electronic Health Records Systems Around the World," 2015.
- [23] "World Health Organization," [Online]. Available: <https://www.who.int/campaigns/world-health-day/2018/campaign-essentials/en/#:~:text=The%20theme%20of%20World%20Health,is%20%E2%80%9CHealth%20or%20All%E2%80%9D..>
- [24] C. Lele, "Development of statistics as a discipline for clinical research: Past, present and future. Perspectives in clinical research," *Perspectives in Clinical Research*, vol. 8, no. 1, pp. 41-44, 2017.
- [25] S. Kaihara, "Information explosion," World Health, 1989.
- [26] J. Klerkx K. Verbert E. Duval, "Enhancing Learning with Visualization Techniques," Katholieke Universiteit Leuven, Belgium.
- [27] H. R. Nagel, "Scientific Visualization versus Information Visualization," Norwegian University of Science and Technology Department of Computer and Information Science.
- [28] B. Chaudhry J. Wang S. Wu et al., "Systematic review: impact of health information technology on quality, efficiency, and costs of medical care," *Ann Intern Med*, vol. 144, no. 10, pp. 742-752, 2006.
- [29] C. Safran M. Bloomrosen W. E. Hammond et al., "Toward a National Framework for the Secondary Use of Health Data: An American Medical Informatics Association White Paper," *Journal of the American Medical Informatics Association*, vol. 14, no. 1, pp. 1-9, 2007.
- [30] D. Shankar, "Health sector reforms for 21(st) century healthcare," *Journal of Ayurveda and integrative medicine*, pp. 4-9, 2015.
- [31] S. Madanian D. T. Parry D. Airehrour et al., "mHealth and big-data integration: promises for healthcare system in India," *BMJ Health Care Inform*, p. 26, 2019.
- [32] D. Shankar B. Patwardhan, "AYUSH for New India: Vision and strategy," *Journal of Ayurveda and integrative medicine*, vol. 8, pp. 137-139, 2017.
- [33] AYUSH ministry Government of India, "AYUSH Namaste portal," [Online]. Available: <http://www.namstp.ayush.gov.in/#/index>.
- [34] AYUSH ministry Government of India, "AYUSH Research Portal," [Online]. Available: <http://ayushportal.nic.in/>.

- [35] AYUSH ministry Government of India, "AYUSH Hospital Management System," [Online]. Available: <http://ehr.ayush.gov.in/ayush/>.
- [36] AYUSH ministry Government of India, "Ayush Grid," 2nd Oct 2020. [Online]. Available: <https://pib.gov.in/PressReleasePage.aspx?PRID=1660936#:~:text=Ayush%20Grid%2C%20the%20emerging%20IT,the%20National%20Digital%20Health%20Mission&text=The%20Ayush%20Grid%20project%20was,backbone%20for%20the%20entire%20sector..>
- [37] AYUSH ministry Government of India, "AYUSH Dashboard," [Online]. Available: <http://dashboard.ayush.gov.in/>.
- [38] A.B. Vaidya, "Reverse pharmacological correlates of ayurvedic drug actions," *Indian J Pharmacol*, vol. 38, p. 311–315, 2006.
- [39] B. Patwardhan, "Bridging Ayurveda with evidence-based scientific approaches in medicine," *EPMA Journal*, vol. 5, 2014.
- [40] B. Patwardhan, "Envisioning AYUSH: Historic Opportunity for Innovation and Revitalization," *Journal of Ayurveda and integrative medicine*, vol. 5, no. 2, pp. 67-70, 2014.
- [41] S. P. Kulkarni, "HURDLES IN RESEARCH IN AYURVEDA AND THEIR POSSIBLE SOLUTIONS," *International Ayurvedic Medical Journal*, vol. 3, no. 1, 2015.
- [42] S. M. Baghel, "Need of new research methodology for Ayurveda," *Ayu*, vol. 32, no. 1, pp. 3-4, 2011.
- [43] D. Shankar, "Directions for revitalization of Ayurveda in the 21st century," *Journal of Ayurveda and integrative medicine*, vol. 9, no. 4, pp. 245-247, 2018.
- [44] H. Walach T. Falkenberg V. Fønnebø et al., "Circular instead of hierarchical: methodological principles for the evaluation of complex interventions," *BMC Med Res Methodol*, vol. 29, no. 6, 2006.
- [45] R. L. Byyny, Spring 2012. [Online]. Available: <https://alphaomegalpha.org/pharos/PDFs/2012-2-Editorial.pdf>.
- [46] A. Patrizio, "IDC: Expect 175 zettabytes of data worldwide by 2025," 3rd Dec 2018. [Online]. Available: <https://www.networkworld.com/article/3325397/idc-expect-175-zettabytes-of-data-worldwide-by-2025.html>.
- [47] T. Cosgrove, "Consult QD: Dealing with Healthcare's Data Explosion by," [Online]. Available: <https://www.google.co.in/amp/s/consultqd.clevelandclinic.org/dealing-healthcares-data-explosion/amp/>.
- [48] K. Adane M. Gizachew S. Kendie, "The role of medical data in efficient patient care delivery: a review," *Risk management and healthcare policy*, vol. 12, pp. 67-73, 2019.
- [49] S. Kandel J. Heer C. Plaisant et al., "Research directions in data wrangling: Visualizations and transformations for usable and credible data," *Information Visualization*, 2011.
- [50] "The Importance of Data Preparation for Business Analytics," 16th July 2019. [Online]. Available: <https://www.ironsidegroup.com/2019/07/16/data-preparation-business-analytics/>.

- [51] "eTutorials.org," [Online]. Available: <http://etutorials.org/Misc/data+quality/Part+I+Understanding+Data+Accuracy/>.
- [52] D. Tobin, "Data Transformation: Explained," 1st June 2020. [Online]. Available: <https://www.xplenty.com/blog/data-transformation-explained/>.
- [53] Committee on Strategies for Responsible Sharing of Clinical Trial Data, "Board on Health Sciences Policy Institute of Medicine. Sharing Clinical Trial Data: Maximizing Benefits, Minimizing Risk," in *The Clinical Trial Life Cycle and When to Share Data*, Washington (DC), National Academies Press (US), 20th Apr 2015.
- [54] A. Kapoor, "Quality Medical Research and Publications in India: Time to Introspect," *International journal of applied & basic medical research*, vol. 9, no. 2, pp. 67-68, 2019.
- [55] J. Tauberer, "Analyzable Data in Open Formats (Principles 5 and 7)," in *Open Government Data: The Book*, 2014.
- [56] K. Tarsi T. Tuff, "Introduction to Population Demographics," *Nature Education Knowledge*, vol. 3, no. 11, p. 3, 2012.
- [57] E. P. Balogh B. T. Miller J. R. Ball, "Committee on Diagnostic Error in Health Care, Board on Health Care Services, Institute of Medicine, The National Academies of Sciences, Engineering, and Medicine," in *Improving Diagnosis in Health Care - The Diagnostic Process*, Washington DC, National Academies Press (US), 29th Dec 2015.
- [58] "PostgreSQL: The World's Most Advanced Open Source Relational Database," The PostgreSQL Global Development Group, [Online]. Available: <https://www.postgresql.org/>.
- [59] The R Core Team, "The Comprehensive R Archive Network," [Online]. Available: <https://cran.r-project.org/>.
- [60] "Python," Python Software Foundation, [Online]. Available: <https://www.python.org/>.
- [61] "Java," Oracle, [Online]. Available: <https://www.java.com/en/>.
- [62] B. Mike, "D3 Data-Driven-Documents," [Online]. Available: <https://d3js.org/>.
- [63] "Tableau," [Online]. Available: <https://www.tableau.com/>.
- [64] M. G. Larson, "Descriptive Statistics and Graphical Displays," *Circulation*, vol. 114, no. 1, pp. 76-81, 2006.
- [65] D. Murray, *Tableau Your Data! Fast and Easy Visual Analysis with Tableau Software*, 1st ed., Wiley Publishing, 2013.
- [66] R. McGill J. W. Tukey W. A. Larsen, "Variations of box plots," *The American Statistician*, vol. 32, no. 1, pp. 12-16, 1978.
- [67] E. Tufte, *The Visual Display of Quantitative Information*, 2nd ed., Cheshire, Connecticut: Graphics Press, 2001.

- [68] D. M. Morales-Silva C. S. McPherson G. Pineda-Villavicencio et al., "Using radar plots for performance benchmarking at patient and hospital levels using an Australian orthopaedics dataset," *Health Informatics Journal*, pp. 2119-2137, September 2020.
- [69] B. McPherson, "The code to generate dynamic bubble plot from Github," [Online]. Available: <https://gist.github.com/brucemcpherson/4684498>.
- [70] C. DeMartini, "datablick," [Online]. Available: <https://www.datablick.com/blog/2018/3/14/layering-data-for-custom-tableau-visualizations>.
- [71] B. McPherson, "Desktop libeartion," [Online]. Available: <https://ramblings.mcpher.com/>.
- [72] H. Hofmann, "Mosaic Plots and Their Variants," in *Handbook of Data Visualization*, Berlin, Heidelberg, Springer Handbooks Comp.Statistics, 2008.
- [73] R. Wicklin, "SAS Blogs," 23 May 2018. [Online]. Available: <https://blogs.sas.com/content/iml/2018/05/23/butterfly-plot.html>.
- [74] T. Siggaard R. Reguant I. F. Jørgensen et al., "Disease trajectory browser for exploring temporal, population-wide disease progression patterns in 7.2 million Danish patients," *Nat Commun*, vol. 11, p. 4952, 2020.
- [75] C. Seung-Seok C. Sung-Hyuk C. Charles et al., "A Survey of Binary Similarity and Distance Measures," *SYSTEMICS, CYBERNETICS AND INFORMATICS*, vol. 8, no. 1, 2010.
- [76] "MedDRA: Medical Dictionary for Regulatory Activities," [Online]. Available: <https://www.meddra.org/>.
- [77] "WHO Drug Global," [Online]. Available: <https://www.who-umc.org/whodrug/whodrug-portfolio/whodrug-global/>.
- [78] "International Classification of Diseases - ICD," [Online]. Available: <https://icd.who.int/browse11/l-m/en>.
- [79] National Cancer Institute. [Online]. Available: <https://evs.nci.nih.gov/ftp1/CDISC/SDTM/SDTM%20Terminology.pdf>.
- [80] "Logical Observation Identifiers Names and Codes - LOINC," [Online]. Available: <https://loinc.org/>.
- [81] "Clinical Data Interchange Standards Consortium: CDISC," [Online]. Available: <https://www.cdisc.org/>.
- [82] "International Organization for Standardization: ISO," [Online]. Available: <https://www.iso.org/home.html>.
- [83] "TransCelerate BioPharma Inc.," [Online]. Available: <https://www.transceleratebiopharmainc.com/>.
- [84] "U.S. Food & Drug administration," [Online]. Available: <https://www.fda.gov/science-research/science-and-research-special-topics/real-world-evidence>.
- [85] "U.S. Food & Drug administration: Framework Real World Evidence Program," [Online]. Available: <https://www.fda.gov/media/120060/download>.

[86] "European Medicines Agency," [Online]. Available: <https://www.ema.europa.eu/en/about-us/how-we-work/big-data/data-analysis-real-world-interrogation-network-darwin-eu>.