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. Observations about data capture methods and storage were noted. 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 and 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:

4.2.1 Visit pattern analysis

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*).

4.2.2 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.

4.2.3   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 [113]

4.2.4   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
Snehana	-	-	-	-	-	+	+	+
Swedana	-	-	-	-	-	+	-	+
Vamana	-	-	-	+	-	-	-	+
Virechana	-	-	-	+	-	-	-	+
Niruha Basti	+	-	-	-	-	-	-	-
Anuvasana Basti	+	-	+	+	+	-	+	+
Siravyedha	+	+	+	-	+	-	-	+
Raktamokshana	-	-	-	-	-	+	+	-
Agnikarma	+	-	+	-	+	+	-	+
Shastrakarma	-	-	-	-	-	-	-	+

Ch - Charaka; Su - Sushrutha; A.H.- Astanga Hrudaya; B.P. - Bhava Prakash; Y.R. - Yoga Ratnakara; H.S. - Haritha Samhitha; B.S - Bhela Samhitha; C.D. - Cakradatta

4.2.5   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 [114] [115]. 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 [116]
- MedDRA: Medical Dictionary for Regulatory Activities [114]
- WHO-ART: World Health Organization Adverse Reactions Terminology [115]
- WHO-DDE: World Health Organization Drug Dictionary Enhanced [115]
- ACD: Ayurvedic Classification of Diseases
- NCI: National Cancer Institute Code list [113]
- LOINC: Logical Observation Identifiers Names and Codes standards [117]
- Any other dictionaries, as recommended by AYUSH

4.2.6   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 (Figure *3-1*). 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.

4.2.7   Patient profile module

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 (Figure *3-4*, Figure *3-5*). 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 medically coded as per section 4.2.5
  - 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 status still ongoing or discontinued (need an algorithm)

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
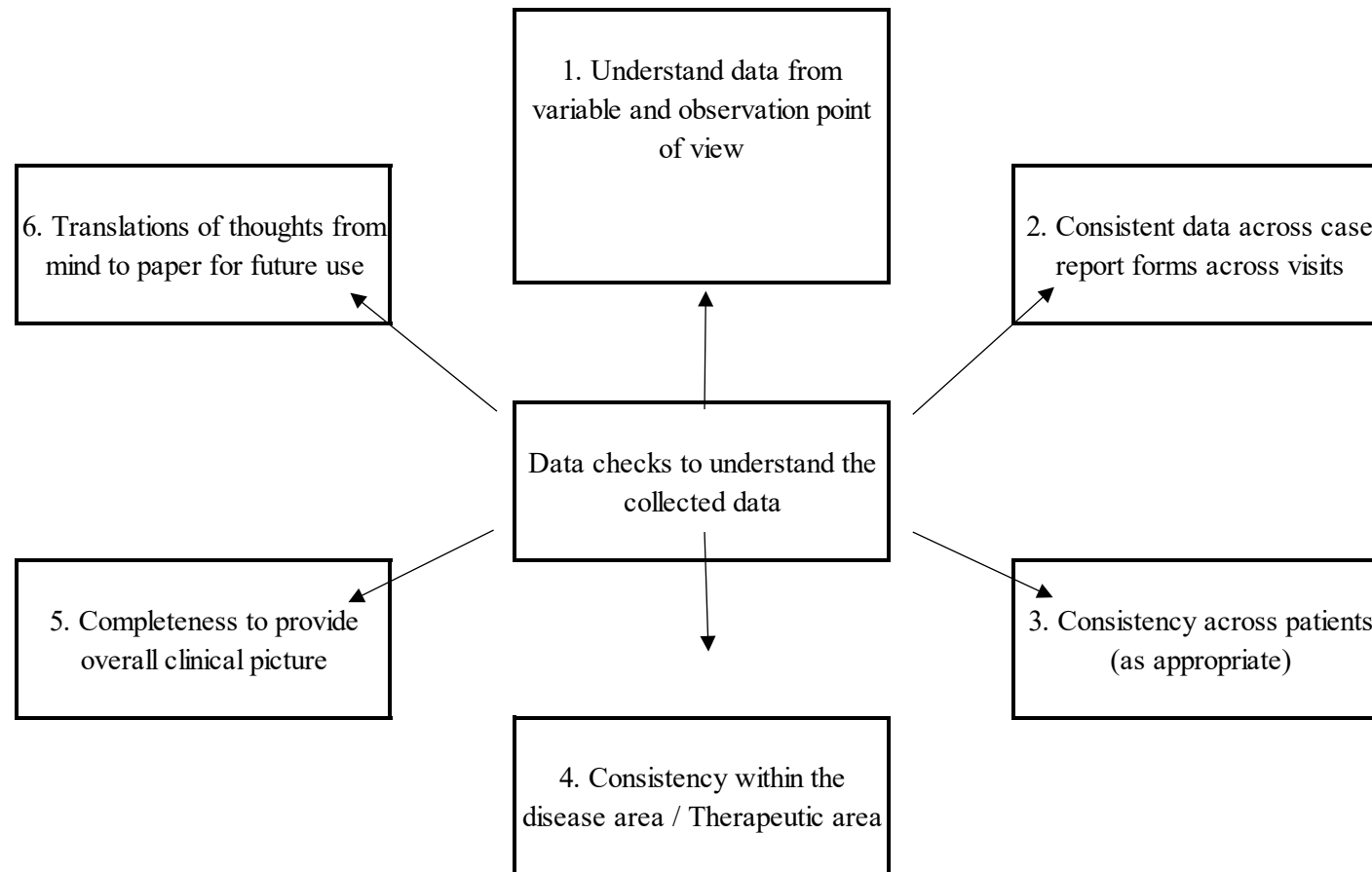
4.2.8 Improvements to the system architecture

Overall the hospital data is not captured in a standardized format as explained in the above points (Figure *3-2*). 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 [118] [119]. Appropriate drop-down menu lists with predefined inputs to be built into the system, with the help of experts, to ensure good quality data.

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 "second 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 undoubted epidemiological value in this presentation. There are obvious mistakes in documenting the blood groups observed via this tabulation – another secondary use of this tabulation is to build data quality related efficiencies (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 (Figure *3-22,* Figure *3-23,* Figure 3-24, Figure *3-25,* and, Figure *3-26*).

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 *3-34,* Figure *3-35,* 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*).

4.5   Use cases with in-depth illustrations
This section outlines a few use cases in-depth and how can the key stake holders make use of the information generated earlier.

4.5.1   Illustration 1: In-depth review of Visit pattern analysis
Data / observations: In a competitive world, understanding how patients plan their visits to a hospital and their social demographics can help in running an operation which would make the patient as the focus point of efficient operations. Operational efficiency allows businesses to build an edge and it is applicable in any field. Visit pattern analysis (Figure *3-3*) allows insights into the same. This visualization depicts data from the very first of the hospital in 2010 on a calendar. In Feb 2011 there were more than 50 patients visit on a single calendar day. In May of 2011, there were more than 200 patients visit on a single day. Otherwise, the average number of visits was hovering around 30 – 50 patient visits. The underlying data shows that the patient inflow increasing year on year. Out of 2100+ days of data from year 2010 to 2016, there were 158 days on which more than 100 patient visit days were there, which accounts for 7.33% of all the days. Most of these 100+ patient visit days happened on Saturday, showing patients' tendency to visit hospital on a weekend. Figure *3-7*: Country-wise Visualization shows that almost 98% of patients are coming from India. Analysis carried out as per Figure *3-25* for V2.0 (Vaatavyadhi) shows slightly a greater number of patients visiting in June, July months compared to other months, providing data on seasonal variations. The numbers start tapering down across other months. Figure *3-15* shows this data in terms of Indian rutus. Jwara (J1.0) in ACD is "flu" in Figure *3-25*, subset for

PrimaryCode = J1.0, should show seasonal variations, as people getting with flu is seasonal in nature. But the overall number of patients reporting Jwara is low (128 females, 140 males) not allowing for these variations to be detected. If Jwara is not reported by large number of patients in the database then that would mean less patients are taking ayurveda treatment to get rid of Jwara. If patients having certain disease conditions are not represented in same proportion in hospital database, considered as a sample, as epidemiologically seen, then the underlying data will not represent disease condition appropriate, this explains relationship between sample and population.

Insights: As the sensitive information about how many doctors, how many nurses, how much money was being charged is not used in this analysis, currently there will not be any conclusions drawn w.r.to economic efficiency. As the location of the hospital is almost 4 km inside from the closest main road and the transport options were not regularly available, there were lesser number of patients seen visiting. Visit pattern analysis report can be used on a periodic basis to understand the patient patterns along with any other material hospital management must be using. Even though more than 50+ countries are represented in the database, the number of unique patients and amount of patient data is small to carry out any meaningful analysis. The visit pattern could be repeated by major disease types to get more insights into what type of doctors should be scheduled, what kind of support staff and what type of perishable pharmacy components made available in an optimal manner. Based on this information, the hospital management can have more staff working on weekends to cover patient inflow. Appropriate amount of stock of medicines in pharmacy could also be arranged. If the patient inflow is consistently lower on other days, then should hospital employ lower number of staff on these days, practically this may not be possible, but for operational efficiency this path should be explored. Some of the repairs and upgrades to the hospital infrastructure could be planned for the weekdays avoiding inconvenience to day-to-day functioning.

4.5.2   Illustration 2: In-depth review of summary statistics of number of diseases

Data / observations: Data quality is an inherent requirement for any data related exercise. Some insights based on patterns of missing data, outliers and potential scientific and medical inconsistencies can suggest methods to improve the data quality which would be of essence to hospital administration.

Analysis covering descriptive summary statistics by number of Diseases by age and gender was carried out (Figure *3-11*) and explained in detail in the earlier chapter 3 section 3.3. This table summarizes descriptive statistics for age. The first column categorizes how many diseases were reported per patient going from 1, 2, ..., n diseases. The table shows 23 diseases as the maximum number of diseases reported by a patient. There is a large part of data showing "NA" = unknown number of diseases. Both observations point to potential data issues. Another key finding from the table is to have 12,884 and 14,375 female and male patients reporting only one disease. Other analysis related to duration and hospital visits (Figure *3-14*) show that there is large dropout rate after one visit. The minimum and maximum age listed goes from 1 to 96 years for females, 2 to 108 years for males.

Insights: These observations should be of interest to the hospital management as well as treating doctors. At the beginning of the hospital operations the ACD coding was not implemented and that resulted in missing disease codes, giving rise to the "NA" category. Did the hospital have a patient with 23 different diseases or does this number also point a data issue or duplicate data entries for a patient? Such kind of data points should be checked on an ongoing basis.

How do hospital management as well as treating doctors view nearly 70% drop out rate after reporting first disease? This analysis will help in multiple ways. Are there operational challenges causing patients not to come back? Are the treatment mechanisms too difficult to follow for patients? What type of follow-up methods should be built to keep the patient engagement high? If this data is to be looked at positively then should we conclude that nearly 70% patients got benefitted from the treatment and hence did not need come back to the hospital. Are there truly 96- and 108-years old patients coming to hospital or are these possible data issues? Are these ages 9.6 and 10.8 years, OR 96 and 108 months / days? It is encouraging to see patients of all ages coming to the hospital. This topic should be of interest to the data managing team at hospital. If the training material is not clearly specifying an appropriate way of data entry, then this analysis helps in improving material as well as Hospital management system. Missing disease information could be understood by treating clinicians or anyone doing data entry on their behalf for having complete documentation for future use.

4.5.3   Illustration 3: In-depth review of disease table by gender

Data / observations: If we can understand the disease and underlying conditions, then the insights generated can give better understanding of science behind the disease. Disease categorization is done by the treating doctor at individual patient visit using ACD classification for documentation. One such example of subset of Prameha has been listed in Figure *3-1*. This subset shows disease categories as Prameha, Prameha – Sthula, Prameha – Pidaka, Prameha – Krusha. The numbers for Prameha category are 648 and 849, and for the other categories these frequency counts are in single digits or in 20s. Is this analysis pointing to less than accurate representation of disease?

Mosaic / tile plot analysis for Prameha (Figure *3-31* subset for Metabolic: P5.0: Prameha) shows that there are 732 unique treatments prescribed, for Prameha – Sthula there are 114 unique treatments prescribed, for Prameha – Pidaka there are 48 unique treatments prescribed and, for Prameha – Krusha there are 27 unique treatments prescribed.

Figure *3-1* has data Aamavata related data represented. Aamavata – Vaataja, Kaphaja and Pittaja categories are tabulated in the analysis. The frequency counts are quite different for each category. 652 females in only Aamavata category followed by 27, 13 and 4 patients in Aamavata – Vaataja, Kaphaja and Pittaja categories. Same is the case for male category.

Insights: Should we assume that reporting just "Prameha" was a lot easier than reporting more details about it in the ACD class? The treating doctor must have treated patients as accurately as possible but thoughts have not be transferred accurately from "mind" to "paper" for future usage. Any analysis carried out using this data for which treatments were prescribed for which variation of the underlying "Prameha" could be a suboptimal approximation. This observation points to a possible shortcoming in usage of ACD and maybe a hospital level training is needed.

Is the current classification accurately expressing the underlying disease, perhaps no. There is a need to have an in-depth discussion with treating doctors to get more insights into this mislabelling. The ayurvedic experts from hospital and if needed experts from other institutions could come together and define guidance documents on how to document a disease.

Due to inaccurate representation of disease data, some of these interpretations will not be fully reliable, even though the big picture view may not be adversely impacted (this sentence should be validated by an ayurvedic vaidya). Can carefully labelling of disease at each visit be done to increase the overall quality of data and possible by products from them. This observation is not only for ayurveda but must be happening in any clinical assessment across the world. These nuances are of great importance to basic researchers. We can find similar examples for other diseases in our existing dataset. If disease classification by Vata, Pitta, Kapha by a treating doctor is not easy to be achieved at each visit to produce "close to 100% real picture of patient" then what could be possible solutions proposed? If there are existing validated tools for treating doctors to use then they should be used.

4.5.4   Illustration 4: In-depth review of individual patient disease journey

Data / observations: If we can understand the patient journey which includes, diagnoses, treatments, outcomes over time, it will allow the treating doctor to adjust and do better for the patient. The insights can as well give more understanding of the disease and the science behind it. This in-depth analysis shows us how to combine three analyses to understand individual patient journey. Figure *3-5*: Patient visit profile – Vertical view, Figure *3-27*: Patient Disease and Treatment administration by Study Day, and Figure *3-29*: Patient Cumulative Disease and Treatment administration by Visit are combined to show how can a treating doctor use these three tools simultaneously. Subset these 3 analyses for patient MR000774. This data is for a 68-year-old female patient who has been coming to the hospital for close to 3 years (960 days). In this period, she has visited hospital for 29 times. There were 9 different diseases reported and 35 prescribed medicines in the hospital database. For a few of these diseases ACD code has not been reported, giving rise to missing data. Madhumeha and Prameha have been reported as first two diseases reported. Is this a correct representation having both Pramhea and Madhumeha being reported?

Diabetes condition is treated by DNil and Nisakathakadi kashayam. Prameha was treated by Nisakathakadi, kathaka kathiradi. Nisakathakadi was prescribed as Choornam / powder and Kashayam, Nisakathakadi treatment was stopped and then Glokustat and Diabecon DS were prescribed. This line-by-line data review provides information about treatment protocol being followed.

The analysis carried out in Figure *3-29* helps in understanding first time disease reported and repeat disease representation may be used to understand the co-morbidities. 26 times diseases were reported and only 9 were unique showing a smaller number of complications, if we ignore "Not coded terms" then there are only 5 unique reported. 95 times treatments were prescribed out of which 35 were unique. There are 2 vaatavyadhis, arthritic conditions, reported after a few visits. Similar details about the treatment can be found in the medicines section.

The analysis carried out in Figure *3-27* provides a tabular and line-by-line listing view at one go, providing one more variation of the same analysis. This report shows reported diseases and medicines as well as updates across timepoints. These three patient journey analyses should be good additions to existing tools to study patient status before a patient visit for any doctor.
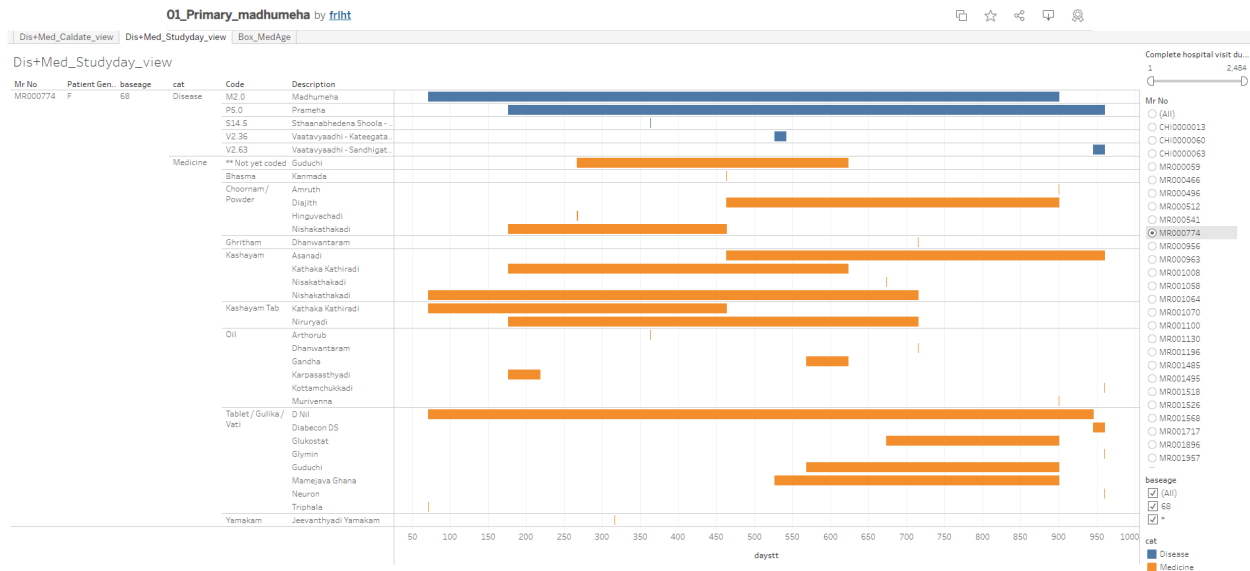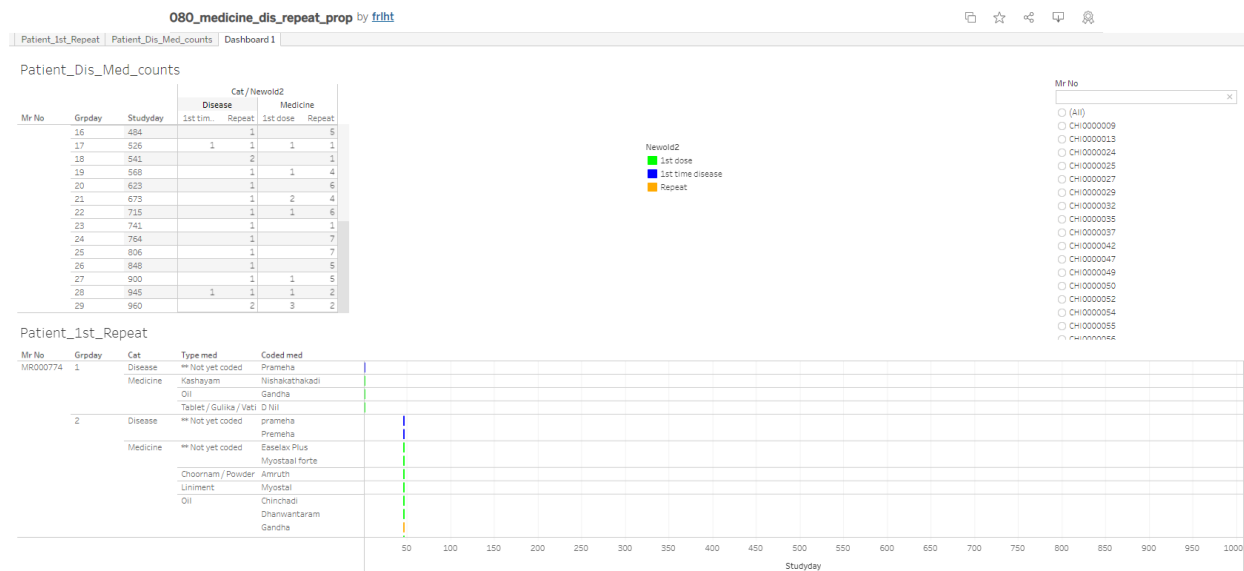

A subset of Figure *3-5* for patient = MR000774

Patient_dis_med_Cumulative

Mr No	Cumday2	Disease 1st time.	Disease Repe.	Medicine 1st dose	Medicine Repe.
MR000774	Till visit 1	1		3	
	Till visit 2	3		12	2
	Till visit 3	4		13	5
	Till visit 4	5		15	8
	Till visit 5	5	1	18	9
	Till visit 6	6	1	18	14
	Till visit 7	6	2	18	19
	Till visit 8	6	3	18	22
	Till visit 9	6	4	20	22
	Till visit 10	6	5	20	26
	Till visit 11	6	6	21	26
	Till visit 12	7	7	22	26
	Till visit 13	7	8	23	30
	Till visit 14	7	9	23	35
	Till visit 15	7	10	25	39
	Till visit 16	7	11	25	44
	Till visit 17	8	12	26	45
	Till visit 18	8	14	26	46
	Till visit 19	8	15	27	50
	Till visit 20	8	16	27	56
	Till visit 21	8	17	29	60
	Till visit 22	8	18	30	66
	Till visit 23	8	19	30	67
	Till visit 24	8	20	30	74
	Till visit 25	8	21	30	81
	Till visit 26	8	22	30	86
	Till visit 27	8	23	31	91
	Till visit 28	9	24	32	93
	Till visit 29	9	26	35	95

%Patient_dis_med_Cumulative

Mr No	Cumday2	Disease 1st time disease	Disease Repeat	Medicine 1st dose	Medicine Repeat
MR000774	Till visit 1	100.00%		100.00%	
	Till visit 2	100.00%		85.71%	14.29%
	Till visit 3	100.00%		72.22%	27.78%
	Till visit 4	100.00%		65.22%	34.78%
	Till visit 5	83.33%	16.67%	66.67%	33.33%
	Till visit 6	85.71%	14.29%	56.25%	43.75%
	Till visit 7	75.00%	25.00%	48.65%	51.35%
	Till visit 8	66.67%	33.33%	45.00%	55.00%
	Till visit 9	60.00%	40.00%	47.62%	52.38%
	Till visit 10	54.55%	45.45%	43.48%	56.52%
	Till visit 11	50.00%	50.00%	44.68%	55.32%
	Till visit 12	50.00%	50.00%	45.83%	54.17%
	Till visit 13	46.67%	53.33%	43.40%	56.60%
	Till visit 14	43.75%	56.25%	39.66%	60.34%
	Till visit 15	41.18%	58.82%	39.06%	60.94%
	Till visit 16	38.89%	61.11%	36.23%	63.77%
	Till visit 17	40.00%	60.00%	36.62%	63.38%
	Till visit 18	36.36%	63.64%	36.11%	63.89%
	Till visit 19	34.78%	65.22%	35.06%	64.94%
	Till visit 20	33.33%	66.67%	32.53%	67.47%
	Till visit 21	32.00%	68.00%	32.58%	67.42%
	Till visit 22	30.77%	69.23%	31.25%	68.75%
	Till visit 23	29.63%	70.37%	30.93%	69.07%
	Till visit 24	28.57%	71.43%	28.85%	71.15%
	Till visit 25	27.59%	72.41%	27.03%	72.97%
	Till visit 26	26.67%	73.33%	25.86%	74.14%
	Till visit 27	25.81%	74.19%	25.41%	74.59%
	Till visit 28	27.27%	72.73%	25.60%	74.40%
	Till visit 29	25.71%	74.29%	26.92%	73.08%

A subset of Figure *3-29* for patient = MR000774

A subset of Figure *3-27* for patient = MR000774

Insights: The spelling was different in different prescriptions which does not change the underlying treatment but creates a perception about extra treatments prescribed to tackle the same disease. If this can be controlled by the coded medicine dictionary and type of formulation, then this will help in data analysis. The arthritic conditions reported by the patient, could be just age-related ailments reported by the patient. This elderly female patient does not have many co-morbidities reported which is a good sign from health point of view.