**Progress report – June 2019**

**Updated title:** Analysis of hospital based ayurvedic clinical practice to gain real world data knowledge

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

By Vinay Mahajan, Girish Tillu, Ashwini Mathur, Ashwini Godbole

Summary: The following progress has been made so far between Jan 2019 and June 2019

All the work done so far has been written as 5 chapters so far. Refinement and updates are expected as the research work continues.

I was supposed to work on specific diseases to understand the patterns and trends. The following 2 areas were chosen to work on to complement the ongoing activities in TDU.

* Chronic Kidney Disease (CKD)– Vrikka roga, – any disease description containing word “vrikka”
* Cancer diseases – any disease description containing word “arbu”

For both these disease categories following datasets and analysis are performed so far:

1. Patient level data ADSLs have been created for disease groups. These datasets contain patient level information for treatments and diseases at each visit. Additional derived variables are derived to get complete clinical picture.
   1. 01adsl\_vrikka.rds
   2. 01adsl\_cancer.rds
2. These 2 datasets are used to create basic analysis displays in tableau, these displays provide fundamental understanding of patient conditions. 15 interactive analyses are carried out for each of the disease conditions.
   1. 01VrikkaRoga\_SQL\_Dis\_Med\_Ser
   2. 01Cancer\_SQL\_Dis\_Med\_Ser
3. Before and after analysis for disease conditions. The following analysis uses 1st occurrence of any disease as day 1 at an individual patient basis. Vrikka roga and cancers were considered as the primary disease. Using this as reference day “before period” and “after period” are derived. “Before period” provides significant amount of “baseline data”, “after period” provides specific insights into what would happen after the reference disease. Following tableau displays provide insights into the patient counts.
   1. 01VrikkaRoga\_Before\_After
   2. 01Cancer\_Before\_After
4. Disease and treatment view, listing and summary for Vrikka roga is created. There are 3 analysis views created for individual patient data.

* 1stpart of the analysis: Patients are treated as they come to hospital. This visual provides a patient level view of number of diseases reported for the first time and then repeated, similarly treatment prescribed for the first time vs. a repeat of treatment.
  1. When a disease is reported very first time then that is considered “1st time disease reported”, any subsequent repetition is considered as “Repeat”.
  2. When a treatment is prescribed very first time then that is considered “1st time treatment prescribed”, any subsequent repetition is considered as “Repeat”.
  3. These 2 calculations are repeated through the data for each patient.
* 2nd part of the analysis:This is a cumulative view for an individual patient. This provides a summary of what would have happened to a patient till a certain visit number. There are 2 tables created, first with absolute numbers and second with percentages.
* 3rd part of analysis: This is another version of display of diseases and treatments for individual patients “non-overlapping or non-cumulative” version.
  1. Each line is a patient visit. 1st disease, Repeat disease, 1st treatment and Repeat treatment columns are displayed.
  2. Studyday column shows the visit day.

Finally all these analyses are listed on a “dashboard” to provide a comprehensive view in tableau: 080VrikkaRogaDis\_Med\_analysis.

Episodic view analysis: Patients come to hospital as and when there is a need either for the same disease or for different diseases. Following algorithm creates patient categories:

1. Create a variable to identify episodes of a diseaseif a disease is re-appearing after 30 days then consider that asa new episode, this duration should be specific to each disease in reality
2. Use the variable “eps01” for cumulative addition and get number of episodes for each disease for each patient, if a disease is non-episodic then use 9999 as the duration
3. This calculation should help in understanding the disease specificpseudo outcome and amount of data collected
4. Use 180 days duration to separate episodes as related vs. un-related as an additional layer of relationship, save this information in a variable called as “releps01”
5. The duration between episodes as well as between related episodes provides an insight into how close or how far the recurrence of events
6. Use these variables along with the overall classification of a patient to create a medical story

Data structure: Unique combinations of PatientID + StudyDay + DiseaseCode, Similar structure could be created with additional treatment variable as well.

|  |  |  |
| --- | --- | --- |
| PatientID | StudyDay (date) | DiseaseCode |

1.     Calculate 2 outcomes for each patient

a.      Outcome at a patient level

b.     Outcome at a patient + each disease level

2.     Outcomes would have the following values:

|  |  |  |  |
| --- | --- | --- | --- |
| Outcome at a patient level | Outcome at a patient + each disease level | Category | Response variable |
| Drop out at 1st visit | Drop out at 1st visit for a disease | Drop out | No response |
| Drop out with 2 to 5 visits | Drop out with 2 to 5 visits | Very limited data | No response |
| Visits >= 5 and duration <= 30 days | Visits >= 5 and duration <= 30 days for a disease | Very limited data | Some response |
| Visits >= 5 and duration >= 30 days | Visits >= 5 and duration >= 30 days for a disease | Moderate amount of data | Limited response |
| Duration > 30 days and <= 180 days | Duration > 30 days and <= 180 days –  Should this be considered as a new episode? | Significant amount of data | Good response |
| Duration >= 180 days | Duration >= 180 days | Large amount of data | Good response |

Need diseases getting classified into

* Chronic vs acute some diseases could be classified into both
* Curable vs only maintenance,
* Disease which could be counted as different episodes

Suppose a patient gets treated for diabetes. And (s)he gets treated for diabetes in all subsequent visits, really nothing can be said about the disease getting better or worse, since it is a chronic disease.

On the other hand, it is for some kind of acute pain and then, say over a period of 6 months there is no pain treatment, it means the patient improved. There is a need to come up with such an algorithm. This has to be at a patient and disease level with very specific algorithm for a disease.

Following R program creates the necessary dataset:

* Program: 102\_episodic\_view.R and
* Output file: 102\_episodicdis01.csv

Responder vs. Non-responder classification:

* Relation of each data point for each disease:
* Individual visits create study day
* Individual visits contribute to the events
* Individual events contribute to the related events
* Calculate the number of events for a disease based on 30 day difference between each visit, this should provide 1 to n events.
* Calculate the median duration for each event for each disease.
* Ignore events with only 1 day of duration, as this shortens the median to 1, thus not providing any insights into the data.
* Assumption: a patient comes back to the hospital only if there is some benefit experienced. Hence longer the duration of “related visits” better the response.
  + In case the duration of an event is greater than equal to the median duration then classify that event as “Responder”, else classify that even as “Non-responder”. Events with only 1 day of data are labeled as “Data for only 1 day”.
  + If a disease could be treated within very limited period of time then the labeling will be reversed. Episode duration less than median duration will be labeled as “Responder”, and rest will be “Non-responder”.

Tableau display: 102\_episodic01\_responder\_nonresponder, shows the episodic views designed to understand the patterns

Pattern mining using SPMF program:

Discovering unexpected and useful patterns in databases is a fundamental data mining task. One of the most popular data mining tasks on sequences is sequential pattern mining. It consists of discovering interesting subsequences in a set of sequences, where the interestingness of a subsequence can be measured in terms of various criteria such as its occurrence frequency, length, and profit. Sequential pattern mining has many real-life applications since data is encoded as sequences in many fields such as bioinformatics, e-learning, market basket analysis, text analysis, and webpage click-stream analysis. The patient level data generated for diseases as well as prescribed treatments provide a sizeable data for data mining task.

In the talk "P for Patterns" presented at the World Ayurvedic Conference, we had initiated some data mining activities. From Jan 2019 onward, the research work on the Pattern finding algorithms using an Open Source Java based Library SPMF (Sequential Pattern Mining Library) was initiated. The details for the algorithms could be found at the following link:

<http://www.philippe-fournier-viger.com/spmf/index.php>

[Philippe Fournier-Viger](http://www.philippe-fournier-viger.com/), Ph.D. is a data mining researcher and professor. He is the founder and main author of the SPMF data mining software, open-source software offering more than 150 algorithms for discovering itemsets, association rules, sequential patterns and rules in sequences and transactions. The SPMF software has been cited in more than 640 papers and was visited by more than 700,000 visitors since 2010. He has also written or participated in more than 200 research papers which have received more than 3000 citations. He is one of the two editor-in-chief of the Data Science and Pattern Recognition journal. He edited the book “High-Utility Pattern Mining” (Springer).

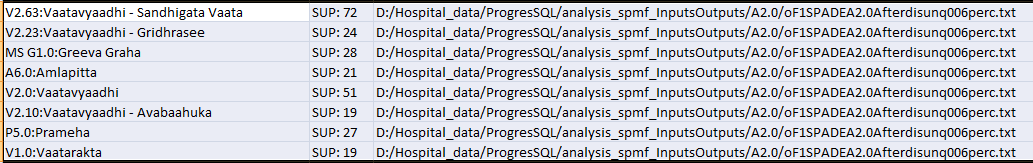
The SPMF utility offers implementations of 171 data mining algorithms as of Feb 2019:

1. association rule mining,
2. itemset mining,
3. sequential pattern
4. sequential rule mining,
5. sequence prediction,
6. periodic pattern mining,
7. episode mining
8. high-utility pattern mining,
9. time-series mining.
10. clustering and classification,

Medicine and disease data are classified before and after for each reference disease (as explained in “15.15 Dataset with each disease considered as a reference disease having day 1”). Some of the data mining algorithms available on the SPMF library are executed on the underlying data.

The following steps are programmed using R program to automate the execution of algorithms:

1. The disease data is split into unique disease trajectories before the onset of reference disease, after disease and across all times. Same is done for the medicine data as well. The data mining utility accepts the data in numerical values hence the data is converted from textual data to numerals.
2. The data for each disease and medicine is saved in text files separated either by blank or by predefined separator as expected by the program.
3. The text file is used as an input by the java program executable file. Appropriate algorithm names, parameters for the algorithm are passed along with the output file name. The text file is passed to individual algorithms to generate the support, confidence and lift.
4. The output file is in not readily available as a human readable file; hence the file will be post processed. The numerals will be formatted back to the appropriate disease and medicine names.
5. For an algorithm there are many parameters settings used hence there are multiple files generated. All the files with different settings after post processing are combined into a single csv file.
6. E.g. for disease “A2.0 -- Aamvata” input file naming “F1SPADEA2.0Afterdisunq.txt” is as follows:
   1. <File separator> [F1 / SPC]
   2. <Algorithm name> [Names available in Java executable file – SPADE in this case]
   3. <Disease Code name> [ACD code name]
   4. <After/Before/All>
   5. <Unique trajectories> [unq]
   6. <Disease/Medicine> [dis/med]
7. The output file for the above algorithm is “oF1SPADEA2.0Afterdisunq\_formatted.csv”, example rows from the output, patients who have reported A2.0 – Aamvata have reported the following diseases in “after” time period.



Presentation at TDU in the following workshop: Real World Evidence and its role in Ayurveda clinical practice and basic research – A data science perspective, 13th and 14th June 2019

2 topics were presented along with Dr. Ashwini Mathur:

1. Software to capture clinical trends from hospital HMIS systems and need for set theory and pattern recognition for interpretation
2. The talk focused on:
   1. Real World Evidence & Epidemiology
   2. Real World Evidence & Clinical practice
   3. Real World Evidence & Basic Research
   4. Real World Evidence & Data requirements

Manuscript written under the following title: “Use of databases and software tools to build empirical evidence of Ayurvedic practice”.

**14. Viz name: DistanceMeasures**

* There are 100s of Distance measures available in mathematics and statistics. These provide the similarity / dis-similarity between objects.
* Diseases experienced by each patient are sorted by date and only 1st instance of a disease is retained. This way a disease trajectory is created for each and every patient for each and every reference disease, before and after the occurrence of the reference disease.
* Cartesian product of patients is created for each reference disease, so that distances can be calculated.
* The similarity measure is calculated for each disease trajectory, e.g. Jaccard distance is used as a distance measure for this display.
* Jaccard distance closer to 0 shows dissimilarities and closer to 1 show similarities.
* The distances are cut into 4 categories 0 to 0.25, 0.25 to 0.5, 0.5 to 0.75 and 0.75 to 1.

| **Sheet name** | **Description** |
| --- | --- |
| DiseaseDist | -Most of the disease trajectories have distance score between 0 and 0.25 for before and after the reference disease. - This could be interpreted as dissimilar. Not many patients have similar diseases before and after the reference disease. - The underlying patient population could be considered as a heterogeneous population. - This should also be seen in distance measures calculated for prescribed medicines. |
| DiseaseMaxDist | - Maximum distance measure from Jaccard is considered for calculations. - This way most similar patients (as per disease trajectory) are analysed. - Similarity scores between 0.75 and 1 increase in "After period" for almost all the diseases when compared to the "Before period". - After an onset of a reference disease similar diseases are experienced and could help in building causal relationship between diseases. |
| DistIndPatientFreq | This view shows frequency count of distinct patients for each reference disease by similarity distance categories for Before and After periods |

**15. Viz name: DistanceMeasures-Medicines:**

* This analysis is same as "DistanceMeasures" analysis. In place of disease trajectories, prescribed treatment trajectories.
* The similarity scores are worse than that for the disease trajectories.
* Most of the prescribed treatments are dis-similar for both the periods.

**16. Viz name: HumanBody02:**

* This visual shows the disease trajectories on a human skeleton.
* Before and after periods are shown as 2 skeletons next to each other.
* Diseases which could be approximately assigned to a body part are displayed on the body, otherwise are displayed on the side of the body.
* Following Frequency counts are displayed:
  + Total number of patients
  + Total number of patients by gender
  + Number of patients experiencing the disease trajectory is displayed
  + Similarity score
  + Total number of diseases listed in the disease trajectory

**17. Viz name: Primary\_disease\_and\_all\_other\_diseases (On Older data)**

Dashboard 2 in this visual is explained below.

| **Sheet name** | **Description** |
| --- | --- |
| Dashboard 2 | This dashboard has 4 sections 1. Summary statistics for age for various diseases. 2. Boxplot of age for various diseases. 3. Number of other diseases contracted at any time while having the underlying disease. 4. Bubble plot using frequency counts for diseases.  The dashboard is controlled by a "PrimaryCode" of a reference disease. The corresponding data is displayed on the page. - The bubble plot displays the number of distinct patients having the primary disease. - Other bubbles 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 diseases experienced.  This display provides a comprehensive view of the disease clsuters. Some diseases could be experienced a lot more than some of the other diseases. Some diseases could be experienced differently by different genders, at different age groups. |

**18. Viz name: 01Cancer\_SQL\_Dis\_Med\_Ser**

Basic analysis carried out to understand the Cancer patients

| **Sheet name** | **Description** |
| --- | --- |
| Cancer\_patients | Frequency table by gender. There is more number of females. Many of these patients may have come for palliative treatment, but there is no evidence to claim this. - Additional classification into different types of cancers is not clearly done in the database. |
| Visit\_Duration | Total duration of hospital visits is calculated as the maximum date of hospital visit - minimum date of hospital visit + 1 in day for each patient. Median duration for females is more than twice for males. |
| Patient\_Visit\_View | This is a listing of individual patient by disease categories. The x-axis displays study day going from day 1 to last visit for each patient. Each bar represents a single study day. The IP visits are marked in Blue and OP visits are marked in Orange colour. For each visit, what kind disease type has been reported is displayed. The tooltip provides additional information related to Total duration of visits to hospital, description of disease, medicine name, first (minimum) day on which Cancer was reported. Many patients with cancer have come only once. |
| DisType\_Diseases | Frequency table for individual diseases by disease category and gender. - Arbuda Maamsaja has been reported by 81 females vs. 6 males, is this breast cancer? |
| MedType\_DisType | Frequency table for medicine categories by gender. Frequency counts for each prescribed medicine is reported. - Cruel Plus, Heeraka Bhasma, Kanchanara are few of the most frequently prescribed medicine |
| Medicine\_DisCode | Frequency table for disease categories, Medicine by gender and individual disease. - as explained earlier, the medicine name variable needs to be modified to get an accurate picture of prescribed medicines |
| DiseaseByStudyDay | Frequency counts of patients reporting a disease on a particular study day under each disease category. - Most of the diseases have the highest frequency reported on day 1 and a steep drop in frequency counts is observed. |
| DiseaseByStudyDay - by season | Analysis similar to DiseaseByStudyDay, the frequency counts are reported by rutus |
| MedByStudyDay | Frequency counts of prescribed medicines on a particular study day under each disease category. |
| SeasonDisease | Frequency table for individual diseases by disease category, rutu and gender. - This analysis should provide insights into seasonal variations of diseases |
| SeasonMedicine | Frequency table for prescribed medicines by rutu and gender. - This analysis should provide insights into seasonal variations of prescriptions of treatments |
| Box\_AgeMed | Box plot representation of age for each prescribed medicine by gender. - This analysis should provide insights into age groupings |
| Slopegraph\_disPatients | Line chart for each disease by calendar year by gender. - Count of distinct patients is plotted on y-axis, the calendar years are displayed on x-axis. - The x-axis can be expanded to an individual month or week or a day to understand the number of patients at a specific time point. - This provides an easy comparison on similar or dissimilar reporting of a specific disease across gender. |
| Slopegraph\_disVisit | Line chart similar to Slopegraph\_disPatients. This visual shows number of distinct visits to the hospital. |
| MedicineByDay | This visual is for prescribed medicine. The medicines are classified into different kinds, Aristham, Asavams, Bhasmas, Arkas, Dhara, Drops, etc. - Frequency counts by each day is plotted by gender. |

**19. Viz name: 01VrikkaRoga\_SQL\_Dis\_Med\_Ser**

Basic analysis carried out to understand the Vrikka Roga patients

| **Sheet name** | **Description** |
| --- | --- |
| -- | The explanation for each of the visual displays is similar to the earlier display for Cancer patients, the data presented here is for Vrikka Roga patients. |

**20. Viz name: 01VrikkaRoga\_Before\_After**

Analysis carried out for the before and after periods of reporting Vrikka Roga

Vrikka Roga is considered as a reference disease.

* Day 1 is calculated as the reference day 1 for individual patient for Vrikka Roga.
* Other diseases for the same patient are positioned either before or after compared to this reference disease (Vrikka Roga).
* Duration w.r.to this reference day is calculated before and after day 1. This calculation provides the background view as well as future view.
* This referencing allows for more informative background disease as well as background medicine information.
* The duration is split into the following time points:

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

* 1 sheet for each reference disease (Vrikka Roga) is created.
* Frequency count of diseases and prescribed medicines is displayed.
* Prior counts are displayed in red colour and After counts are displayed in Green colour.

This view should provide good insights into the causal relationships.

**21. Viz name: 01Cancer\_Before\_After**

Analysis carried out for the before and after periods of reporting Cancer

The explanation for each of the visual displays is similar to the earlier display for Vrikka Roga patients, the data presented here is for Cancer patients.

**22. Viz name: 080VrikkaRogaDis\_Med\_analysis**

Analysis of number of diseases and number of prescribed treatments for Vrikka Roga patients

This shows individual patient data for disease and treatment for Vrikka Roga patients. It provides diseases and treatments per patient as either  
"disease reported 1st time" or "repeat" (data displayed in Sheet Disease),  
"treatment reported 1st time" or "repeat" (data displayed in Sheet Medicine).

* It is reported by studyday (or visit) when a disease and medicine is reported in the data.

The data is analysed in 2 forms (1) what would have happened at every visit and (2) cumulative form (till a particular visit)  
Interpretation  
(1) When a new disease is reported, usually a new treatment(s) is (are) reported  
(2) If there is only a new treatment added then it could indicate, the earlier treatment may not have worked, or it explains the nature of treatment regimen.

| **Sheet name** | **Description** |
| --- | --- |
| Disease | "disease reported 1st time" or "repeat" |
| Medicine | "treatment reported 1st time" or "repeat" |
| Patient\_1st\_Repeat | Individual patient listing of disease and medicine classified as 1st or "repeat" by study day |
| Patient\_Dis\_Med\_counts | Frequency count of disease and medicine classified as 1st or "repeat" by study day (i.e. each visit) |
| Patient\_dis\_med\_Cumulative | Frequency count of disease and medicine classified as 1st or "repeat" by cumulative study day (i.e. till xx visit) |
| %Patient\_dis\_med\_Cumulative | % of the cumulative classification |
| Dashboard 1 | All the above data is displayed on 1 page, the patient subset applied on this page is applied to all other pages for easy navigation and review |

**23. Viz name: 102\_episodic01\_responder\_nonresponder**

Based on study visits, creation of the disease episodes, related / un-related diseases / Only 1 day visit

**23a. Viz name: 102\_episodic01\_responder\_nonresponder\_vrikka\_roga**

Based on study visits, creation of the disease episodes, related / un-related diseases / Only 1 day visit for Vrikka Roga patients

Episodic view analysis:

* Patients come to hospital as and when there is a need either for the same disease or for different diseases. There is no fixed protocol as well as fixed visit schedule which they would need to follow. A lot of patients (more than 50%) do not visit more than 1 visit. Within a month, the overall patient proportion falls to 30%. Due to this underlying reason the response variable is not properly captured.
* The following algorithm attempts to create an artificial response rate. Following steps create patient categories:

1. Create a variable to identify episodes of a disease if a disease is re-appearing after 30 days then consider that asa new episode, this duration should be specific to each disease in reality
2. Use the variable “eps01” for cumulative addition and get number of episodes for each disease for each patient, if a disease is non-episodic then use 9999 as the duration
3. This calculation should help in understanding the disease specific pseudo outcome and amount of data collected
4. Use 180 days duration to separate episodes as related vs. un-related as an additional layer of relationship, save this information in a variable called as “releps01”
5. The duration between episodes as well as between related episodes provides an insight into how close or how far the recurrence of events
6. Use these variables along with the overall classification of a patient to create a medical story

| **Sheet name** | **Description** |
| --- | --- |
| Resp/NonResp/List | Individual patient listing for each disease (diseases are listed as an individual page), the response is displayed as 1 Day Visit, Responder, Non-responder. \* The subset created on this page is applied to all other pages |
| Resp/NonResp/RelatedEvent | This is similar to the earlier display, an additional "related episode" variable is displayed to understand the recurrence of the related events |
| Respond/NonRespond/Boxplot | Boxplot of responder vs. non-responder by individual episode, due to the current definition, the responder category would have more number days compared to the non-responder category |
| SummaryStatsEpisodeDur | Descriptive statistics for the same analysis are displayed in "Respond/NonRespond/Boxplot" display |
| BoxplotDiffBetEpisodes | This boxplot provides information about the number of days between episodes, the minimum duration between the 2 episodes is 30 days due to the algorithm used. Far apart the duration between the 2 episodes, the possibility of them being clinically independent from each other. |
| SummaryStatsDiffBetEpisodes | Descriptive statistics for the same analysis are displayed in "BoxplotDiffBetEpisodes" display |

**24. Viz name: 105\_trt\_dis\_unq\_mult,**

Identification of prescribed treatments to a specific disease or multiple diseases

**24a. Viz name: 105\_trt\_dis\_unq\_mult\_vrikka\_roga**

Identification of prescribed treatments to a specific disease or multiple diseases for Vrikka Roga patients

* Treatments and diseases are reported per visit on a case report form. Multiple diseases and multiple treatments could get reported in a visit. Due to the database set-up, the 1 to 1 mapping of the disease and treatment may not be possible. In ayurvedic treatment mechanism, the same treatment could be administered for multiple diseases and vice versa.
* Patients are classified as having reported only a single disease and having reported multiple diseases.  
  This analysis carried out for a specific single disease provides a view.  
  E.g. the same analysis is carried out for the Vrikka Roga patients. The single disease defined in that analysis provides information about the prescribed medicines to patients only reporting Vrikka Roga. The most frequently occurring treatments should provide a good view of the medical options for Vrikka Roga. If the patients continue coming for treatment for considerable amount of time (which is not the case for majority of patients) then this approach would have the ability to provide "complete treatment protocol".

**25. Viz name: 107\_prim\_sec\_diag01**

Disease-disease relationship, consider 1 disease as a primary disease and then calculate the duration between 2 reported events, distant events may mean clinically irrelevant events, closer could mean related to each other

**25a. Viz name: 107\_prim\_sec\_diag01vrikka\_roga**

Disease-disease relationship as explained above carried out considering Vrikka Roga as a primary disease

| **Sheet name** | **Description** |
| --- | --- |
| Dashboard 2 | This dashboard has 4 sections 1. Summary statistics for age for various diseases. 2. Boxplot of age for various |