Standardization and Analysis of EHR Data

Raj Kapoor Gupta

NIT, Rourkela

May 06, 2019

Overview

- Introduction
- Literature Survey
    - Motivation
    - Problem Statement
- Methodology
    - Dataset Description
    - Datset Preprocessing
        - Data Aggregation
        - Data Cleaning
    - Preventive Analysis

Overview - Continued

- Preliminary Result
  - Clustering by population
  - Clustering by patient category
  - Clustering by gender
  - Clustering by location of living
  - Clustering by Doctors
  - Clustering by seasons
- Holt-Winters' Seasonal Method
- Conclusion and Future Work

Motivation

The health care industry historically has generated large amounts of data driven by record keeping, compliance regulatory requirements and a range of health care functions[1], including among others clinical decision support , disease surveillance and population health management [2-5]. Big Data in health care is overwhelming not only because of its volume but because of the diversity of the data types and the speed at it which must be managed.[6]
Through the course of this project we plan to study the medical patterns of the localized ecosystem and use the information obtained to take measures to check the spread of diseases.

Problem Statement

- National Institute of Technology, Rourkela, has dealt with almost 157754 medical cases in 2018, and 153136 cases in 2017 previously.We want to study the pattern of occurence of diseases in the institute and the reasons for their occurence. Thus preventive measures can be taken on the basis of these pattern for the wellness of both employees and students. No standardized dataset was however maintained by the institute to carry on this process and so effort was put in to aggregate the data from different sources. Then clusters were made to localise the diseases based on certain factors.

Literature Survey

- Wager (2005) said that in order to understand the relation between information technologies and healthcare, we first need to understand what are the technologies used in healthcare.
- Bhattacherjee and Hikmet (2007) and Castro (2007), granted that information technology has improved healthcare industries, but they also highlighted some of the difficulties related to the use of information technology in healthcare sectors, as they noticed that it is hard to implement information technology in small clinics and organizations, with high costs due to reduced efficiencies of scale.

Literature Survey-Continued

- Big data analytics applications in healthcare take advantage of the explosion in data to extract insights for making better informed decisions [9-11], and as a research category are referred to as, no surprise here, big data analytics in healthcare [12-14].
- Data from healthcare includes clinical data from clinical decision support systems (physicians written notes and prescriptions, medical imaging, laboratory, pharmacy, insurance, and other administrative data); patient data in electronic patient records (EPRs); machine generated/sensor data, such as from monitoring vital signs; social media posts, including Twitter feeds (so-called tweets) [7], blogs [8].
- Via analytics, payers are able to monitor adherence to drug and treatment regimens and detect trends that lead to individual and population wellness benefits [11,15-17].

Methodology

The methodology of the project can be divided into the following subgroups

- Dataset Description
    - The EHR or electronic health record analysis was done of the National Institute of Technology, Rourkela's dispensary. This EHR contained information about the patients visiting the dispensary and was made available for the research by the Automation Cell of the Institute. It was divided into two datasets which contained records from the year 2017 and 2018.The columns of the Dataset were:
    - VDate- Visiting Date
    - ID- Patient ID
    - Category - Patient Category
    - birthyear- Birth Year
    - Gender- Gender of Patient
    - State - State of Patient
    - Hall- Hall of Patient
    - BloodGroup- Blood Group of Patient

Methodology

- YearofAdm - Year of Admission
- chiefcomplain - Chief Complaint
- Diagnostics - Diagnostic Made
- MedicineName - Name of medicine Described
- Referral Nature - Type of Referral
- Referal Reason - Reason given for Referral
- Ref Hospital - Hospital for which Referral Given
- Doctor - Doctor which attended to the patient

Dataset Preprocessing

Data Aggregation

- The automation cell earlier only kept a small subsample of the data in a table but it lacked adequate entries in other columns for data analytics.
- the primary challenge was to aggregate the data from different tables kept in the database and connect them using suitable joins to form the aggregated dataset.
- The ID of the patient was used to find out the category,birth year, gender, state and hall from tables like Student and Employee Details
- This aggregation was stored in the form of procedures so that it can be reused in the future.

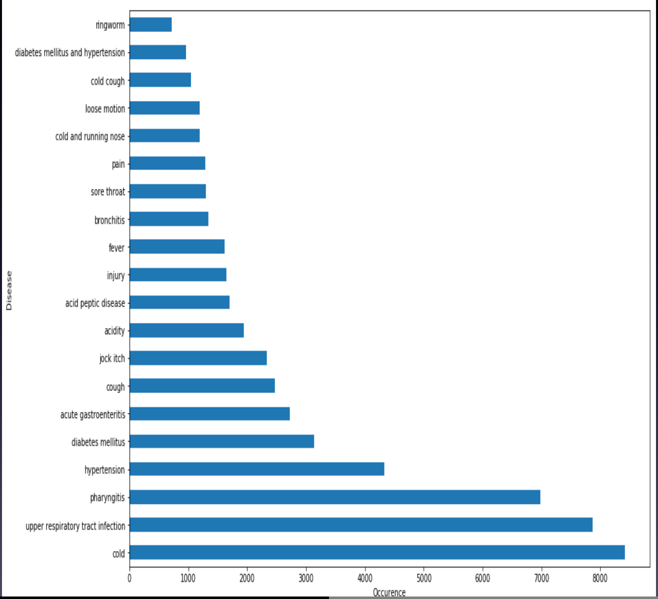
Data Cleaning

- The cleaning of data was a substantial task as it had 157754, 153136 entries for the year 2018 and 2017. Challenges faced were as follows:
- Converting the Visiting Date into a single uniform format.
- State Abbreviations were used in some entries and they had to be converted into full form.
- The chief complaints were written by the doctors in short hand notation across different entries and also sometimes their orders were inconsistent.
- The rows where chief complaints missing and could not be clarified were dropped and not counted for analysis.

Preventive Analysis

- This part of the analysis focuses on methods which observes trends and formulates reasons for the occurence of the trends. It then suggests preventive measures to stop or fix problems that can be causes by these trends.
- Cluster Analysis
- After the pre processing step the data was ready for analysis. It was decided that primarily the focus would be on the spread of diseases in the institute and the reasons for it. This was done by finding the most frequent diseases which occured in the campus.

Preliminary Results - By Population



By Patient Category

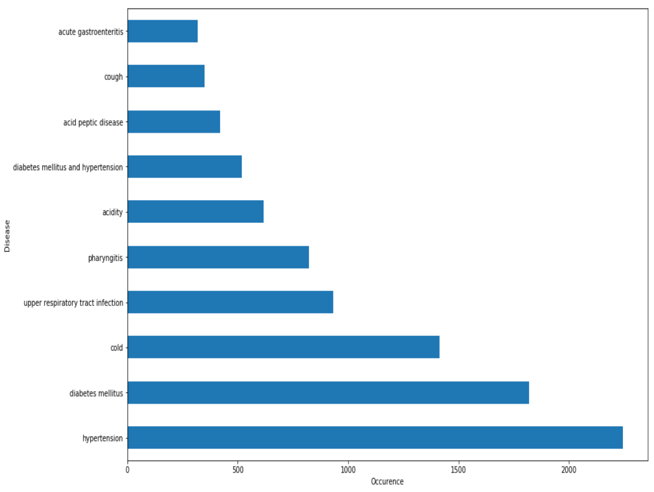


Figure: Bar Graph showing most frequent diseases with respect to Employees

By Patient Category

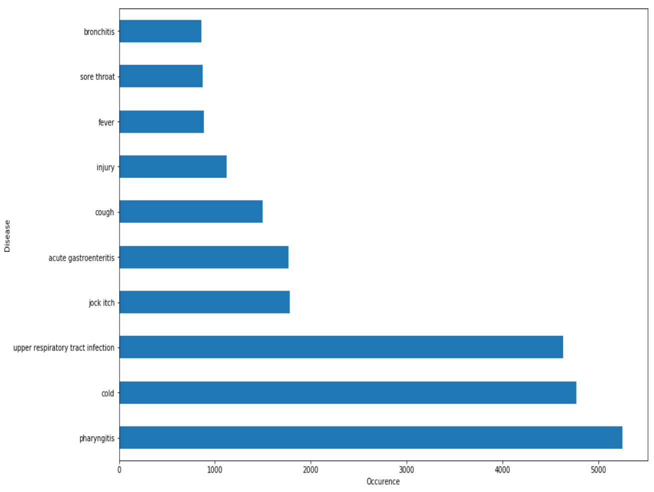


Figure: Bar Graph showing most frequent diseases with respect to students

By Gender

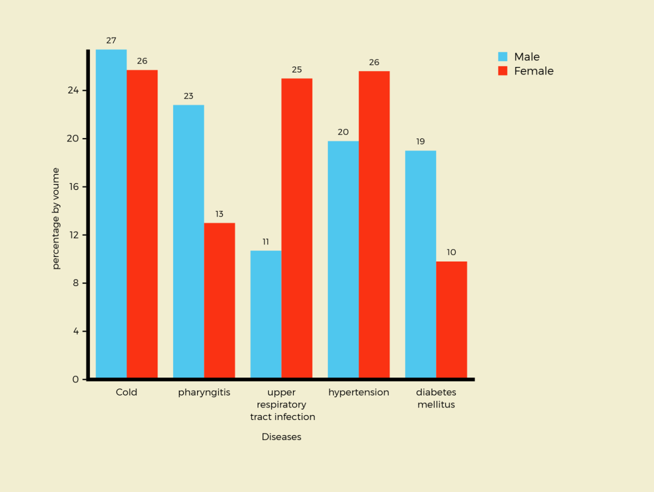


Figure: Bar Graph showing trends with respect to gender

By Location of Living

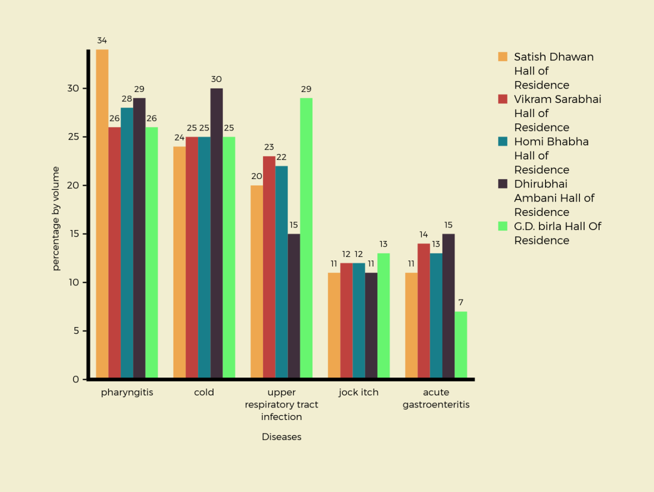


Figure: Bar Graph showing trends with respect to hostel

By Location of Living

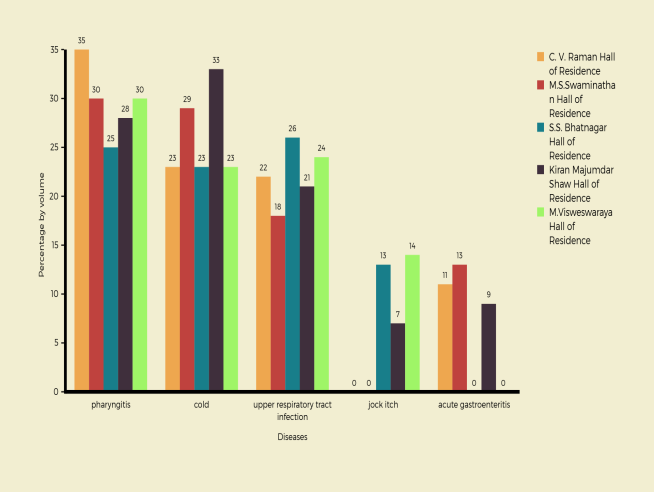
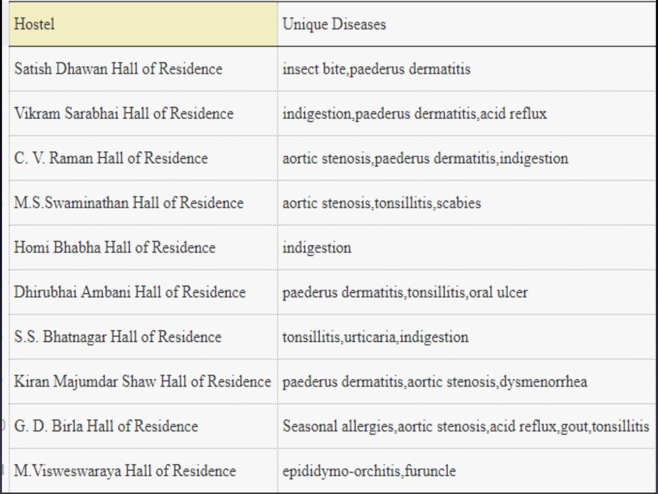


Figure: Bar Graph showing trends with respect to hostel

By Location of Living



Hostel	Unique Diseases
Satish Dhawan Hall of Residence	insect bite,paederus dermatitis
Vikram Sarabhai Hall of Residence	indigestion,paederus dermatitis,acid reflux
C. V. Raman Hall of Residence	aortic stenosis,paederus dermatitis,indigestion
M.S.Swaminathan Hall of Residence	aortic stenosis,tonsillitis,scabies
Homi Bhabha Hall of Residence	indigestion
Dhirubhai Ambani Hall of Residence	paederus dermatitis,tonsillitis,oral ulcer
S.S. Bhatnagar Hall of Residence	tonsillitis,urticaria,indigestion
Kiran Majumdar Shaw Hall of Residence	paederus dermatitis,aortic stenosis,dysmenorrhea
G. D. Birla Hall of Residence	Seasonal allergies,aortic stenosis,acid reflux,gout,tonsillitis
M.Visweswaraya Hall of Residence	epididymo-orchitis,furuncle

Figure: Name of Unique Diseases in Hostels

By Doctors

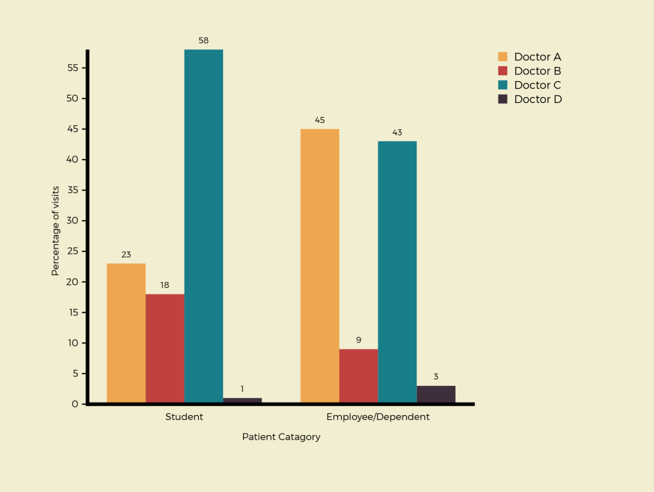


Figure: Bar Graph showing trends with respect to doctors

By Season

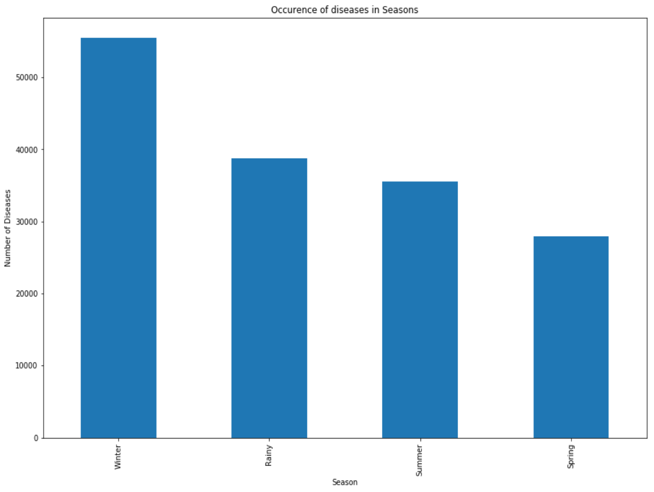


Figure: Bar Graph showing trends with respect to season

Holt-Winters’ Season Method

There are two variations to this method that differ in the nature of the seasonal component.

- Additive Method (preferred when the seasonal variations are roughly constant through the series)
- Multiplicative Method (preferred when the seasonal variations are changing proportional to the level of the series.)

Holt-Winters multiplicative method

The Holt-Winters seasonal method comprises the forecast equation and three smoothing equations one for the level

$$l_t$$

, one for the trend

$$b_t$$

, and one for the seasonal component

$$s_t$$

, with corresponding smoothing parameters α,β and γ. We use m to denote the frequency of the seasonality, i.e., the number of seasons in a year. For example, for quarterly data m=4 , and for monthly data m =12.

Holt-Winters' Multiplicative Method

The component form for the multiplicative method is:

$$Y_{t+\frac{h}{t}} = (l_t + hb_t)S_{t+h-m(k+l)}$$

$$\text{l}_t = \alpha\frac{Y_t}{S_{t-m}} + (1-\alpha)(l_{t-l} + b_{t-1})$$

$$\text{b}_t = \beta^*(l_t - l_{t-1}) + (1-\beta^*)b_{t-1}$$

$$\text{S}_t = \gamma\frac{Y_t}{l_{t-1}+b_{t-1}} + (1-\gamma)S_{t-m}$$

where,

k is the integer part of $(h-1)/m$,

which ensures that the estimates of the seasonal indices used for forecasting come from the final year of the sample.

Holt-Winters’ Multiplicative Method

- The level equation shows a weighted average between the seasonally adjusted observation $(Y_t - s_{t-m})$ and the non-seasonal forecast $(l_{t-1} + b_{t-1})$ for time t.
- The trend equation is identical to Holts’ linear method. The seasonal equation shows a weighted average between the current seasonal index, $(y_{t-1} - l_{t-1} - b_{t-1})$, and the seasonal index of the same season last year (i.e., m time periods ago).

Holt-Winters' Season Method

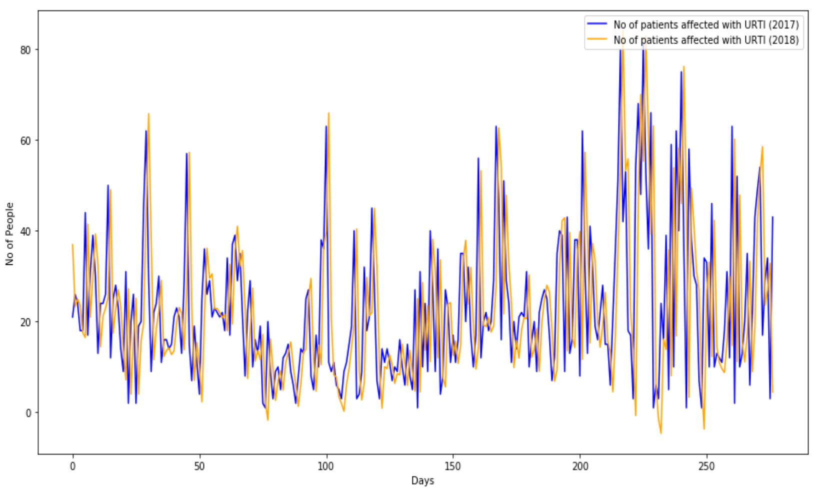


Figure: Graph showing trends with respect to No. of patients affected by URTI

Conclusion and Future Scope

- The trends of occurance of diseases in our institute has been captured in our work with respect to different parameters. Holt-Winters' method has been used to predict the trend of diseases's occurance and set of the features that might depend on it. Thus our work will benefit in both fields of disease prevention and hospital resourse management.
- The columns regarding the medicine have to be converted into a generic formate (for example like calpol belongs to paracetamol) so that the Pharmacy can have adequate stock of medicines when required. This remains a challenging task since there are lakhs of medicines and to convert them into a generic name web scraping techniques may need to be used to scrap the medicine name automatically from different websites.

References

- Raghupathi W: *Data Mining in Health Care.* In Healthcare Informatics:Improving Efficiency and Productivity. Edited by Kudyba S. Taylor Francis;2010
- Burghard C: *Big Data and Analytics Key to Accountable Care Success.* IDC Health Insights; 2012.
- Dembosky A: *Data Prescription for Better Healthcare.* Financial Times, December 12, 2012, p. 19; 2012.
- Feldman B, Martin EM, Skotnes T: *Big Data in Healthcare Hype and Hope.* October 2012. Dr. Bonnie 360; 2012.
- Fernandes L, OâŁ™Connor M, Weaver V: *Big data, bigger outcomes.* J AHIMA 2012
- Frost Sullivan: *Drowning in Big Data? Reducing Information Technology Complexities and Costs for Healthcare Organizations.*
- Bian J, Topaloglu U, Yu F, Yu F: *Towards Large-scale Twitter Mining for Drugrelated Adverse Events.* Maui, Hawaii: SHB; 2012.
- Raghupathi W, Raghupathi V: *An Overview of Health Analytics. Working paper*; 2013.

References

- Ikanow: *Data Analytics for Healthcare: Creating Understanding from Big Data.*
- jStart: *How Big Data Analytics Reduced Medicaid Re-admissions.* A jStart Case Study; 2012.
- Knowledgent: *Big Data and Healthcare Payers*; 2013.
- Explorys: *Unlocking the Power of Big Data to Improve Healthcare for Everyone.*
- IBM: *IBM big data platform for healthcare.* Solutions Brief; 2012.
- Intel: *Leveraging Big Data and Analytics in Healthcare and Life Sciences: Enabling Personalized Medicine for High-Quality Care, Better Outcomes*; 2012.
- IBM: *Data Driven Healthcare Organizations Use Big Data Analytics for Big Gains*; 2013.
- Savage N: *Digging for drug facts.* Commun ACM 2012, 55(10):11âŁ“13.
- Zenger B: *Can Big Data Solve HealthcareâŁ™s Big Problems?* HealthByte, February 2012; 2012.

The End