**Progress report – Dec 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 July 2019 and Dec 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.

Review of Real World Evidence methods and corresponding analysis.

EHR or EMR analysis using variety of methods:

When a patient visits a doctor or a hospital, what comes to anyone’s mind: “what is wrong?” and “what happens next?” The first question suggests the diagnosis of the illness; the other part is about prediction of future medical risk. At least till late 1990s, these questions were answered as well as solved by individual doctors. Year 2000 or so onwards, rapid growth in use of Electronic Medical Records (EMRs) or Electronic Health Records (EHRs) offers promises for healthcare analytics [Hillestad et al., 2005], such as chronic disease management and personalized medicine.

EMRs contain a wealth of healthcare information, including medication, procedure, diagnosis codes and lab test results. EHRs have created large amounts of data in different pockets of the world.

Bare minimum version of any EHR should include the following data points:

* Patient ID
* Patient gender
* Patient age
* Visit date
* Nature of visit (Inpatient / Outpatient)
* Disease diagnosis(s) coded in one of the many standard dictionaries
* Prescribed medicine(s) coded in one of the many standard dictionaries
* Administered procedure(s) coded in one of the many standard dictionaries
* Primary / secondary nature of the reported disease
* Lab tests
* Etc …

The IAIM hospital EHRs contain all of the above-mentioned data points plus a few more for the ayurvedic parameters. In ayurvedic context where there are many practicing doctors writing this valuable only on case sheets, this large and rich data is not readily available for any kind of learnings.

Structure of EHRs fall into following categories:

1. Patient level data: EHRs are populated for each patient at each visit, as and when a patient visits a medical institute.
2. Heterogeneous nature of EHRs: Each institute could have a way of collecting the data on EHR making them heterogeneous. The individual patient data is not readily available for due to various reasons like GDPR, IPR, sensitive information, etc.
3. Longitudinal: For each visit for each patient an EHR is filled, hence they become longitudinal. The type of visit (in-patient, out-patient), regular visits, emergency visit, etc. are time stamped to make the visits contextual.
4. Temporal relationship: Longitudinal nature allows EHRs to build the ordering of medical events, helping in building medical history, diagnosis, prognosis, potential causality, etc.
5. Irregular timing of visits: EMR varies greatly in length – young patients would have just one visit for an acute condition and older patients with chronic conditions may have hundreds of visits.
6. Episodic nature: EMRs are created only when a patient visits, hence each disease condition is represented as an episode. The episode is often dependent on the type of a disease, typically ranging from a day to two weeks for acute diseases of mild and moderate severity, going all the way up to a few years for severe chronic diseases. This dictates the timing of visits, largely random.
7. Sparse and high dimensional: Hundreds and thousands of disease conditions are reported for individual patients using many standard dictionaries like ICD for diagnosis codes, CPT for procedures, LOINC for lab tests, SNOWMED for medicines, etc. As many conditions could be reported only a few times in many visits, this reporting structure makes the data sparse and high dimensional.
8. Structured and unstructured content: Along with the coded terminologies, free text of doctor’s notes section contains precious medical information in many different languages.
9. Progressions of diseases and recovery reported: EHRs are a combination of the course of reported disease, the evolving and the intervening processes

Based on this type of data, we can get many insights into following questions:

1. How to predict the next disease?
2. How to predict the next prescribed medicine?
3. How to predict the next visit?
4. How to predict the course of treatment?
5. How to represent the clinical diagnostic and prognostic algorithms?
6. How to predict a “risky” patient?

This gives rise to different kinds of problem solving:

1. Prediction algorithms based on neural networks
2. Natural language process problem [NLP] from the unstructured medical data
3. Medical Concept Embedding from visit view of EHRs
4. Representation of high dimensional data into low dimensional data

Following algorithms will be studied based on the IAIM EHR data:

1. word2vec algorithm
   1. Continuous Bag of words (CBOW)
   2. Skipgram
2. Glove algorithm
3. Wang2vec algorithm
4. Med2vec algorithm
5. MCE algorithm
6. If possible DeepCare, Deepr [a few programs are not available publicly]
7. Any other relevant approaches possible to be run on the EHR data

These steps need to be executed for many algorithms.

Raw data available

Convert to the data structure as needed by the classification, embedding, prediction program

Split the data into

(1) Training set, (2) test set, (3) validation set

Execute the necessary algorithms and complete the training model

Calculate the model performance parameters and report them

CoMorbidity analysis using “CoMorbidity” package in R based on EHR data:

The term comorbidity refers to the coexistence or presence of multiple diseases or disorders in relation to a primary disease or disorder in a patient. A variety of plots and heatmaps are provided to display the results of the comorbidity analysis. The tasks that can be performed with the comoRbidity package are the following:

1. Age and sex analysis of the population diagnosed with the disease of interest.
2. Clinical comorbidity analysis, based on diagnosis data, in a specific sex and age interval.
3. Analysis of the comorbidity temporal directionality.
4. Visualization of the results in a clear, easily interpretable manner.

The comoRbidity package also expedites the integration of comorbidity results with other R packages, and allows the development of complex bioinformatics workflows.

A paper was submitted to the Current science in August 2019, but the editorial team has provided no response yet.

A presentation to the AYUSH secretary was done in Nov 2019: the link the presentation is provided here: <https://github.com/coursephd/PostgreSQL/tree/master/thesis/Pune-Nov2019>