# Data Analytics in Healthcare

**Bijan Raahemi, Ph.D., P.Eng, SMIEEE**
Associate Professor
Director of the Knowledge Discovery and Data Mining Lab
University of Ottawa
Ottawa, Canada

Knowledge Discovery and Data Mining is the nontrivial process of extracting implicit, novel, and useful information from large volume of data. It is a multi-disciplinary field of science and technology ranging from machine learning, artificial intelligence to computer algorithm design, information retrieval, and database systems. It spans a wide range of applications in Engineering (e.g. intrusion detection and network security, flow classification), business (e.g. fraud detection, decision support systems, forecasting market trend), medicine and population health (e.g. study of drug implications, disease outbreak), and environmental science (e.g. flood prediction).

In healthcare, major applications of Knowledge Discovery and Data Mining fall into four categories: (a) Clinical Medicine: analyzing complex clinical, laboratory, equipment use, and drug management data for disease diagnosis and decision making; (b) Public Health including early outbreak detection, healthcare and syndromic surveillance; (c) Healthcare Text Mining including mining medical literature, as well as mining clinical data such as patients' clinical records; and (d) Healthcare Policy and Planning including detecting expensive clinical profiles among patients diagnosed with a specific chronic illness, and helping with resource allocation and capacity planning.

In this talk, I will present the results of two recent studies we conducted in our Knowledge Discovery and Data Mining lab pertinent to category (a) and (d), respectively:

**(I) Brain-based Biomarkers for Depression Diagnoses**

We built predictive and descriptive models to diagnose depressed individuals based on the EEG signals recorded of brain activities in three frequency bands (Alpha, Beta, and Theta). The data was analyzed through the use of various analytical models and data mining techniques including neural networks, decision trees, and k-means clustering. This study identifies significant biomarkers in the EEG signals that can help physicians in their decision making. The findings of this study will contribute to a larger clinical trial, aiming to determine whether treatment with two antidepressants is more effective than treatment with only one.

**(II) Predicting High Cost Patients in General Population using Data Mining Techniques**

We applied data mining techniques to a nationally-representative expenditure data from the US population to predict very high-cost patients in the top 5 cost percentiles. A dataset of 100,000 records was pre-processed and modeled by Decision Trees, and Neural Networks. Multiple predictive models were built and their performances were analyzed using various measures including correctness accuracy, and G-mean. We concluded that among a primary set of 66 attributes, the best predictors to estimate the top 5% high-cost population include individual's overall health perception, history of blood cholesterol check, history of physical/sensory/mental limitations, age, and history of colonic prevention measures. The prediction is independent of the patients' visits to doctors or hospitalized. Consequently, the results from this study can be used by policy makers, health planners, and insurers to plan and improve delivery of health services in advance.