**Clinical Deterioration Prediction Model: Inferential Statistics**

For this project, I used publicly available Electronic Health Records (EHRs) datasets. The MIT Media Lab for Computational Physiology has developed MIMIC-IIIv1.4 dataset based on 46,520 patients who stayed in critical care units of the Beth Israel Deaconess Medical Center of Boston between 2001 and 2012. MIMIC-IIIv1.4 dataset is freely available to researchers across the world. A formal request should be made directly to [www.mimic.physionet.org](http://www.mimic.physionet.org/), to get access to the data. There is a required course on human research ‘Data or Specimens Only Research’ prior to data access request. Copy of my human subject certification can be found here -[www.citiprogram.org/verify/?kb6607b78-5821-4de5-8cad-daf929f7fbbf-33486907](http://www.citiprogram.org/verify/?kb6607b78-5821-4de5-8cad-daf929f7fbbf-33486907)

The MIMICIII dataset has 26 relational tables including patient’s hospital admission, callout information when patient was ready for discharge, caregiver information, electronic charted events including vital signs and any additional information relevant to patient care, patient demographic data, list of services the patient was admitted or transferred under, ICU stay types, diagnoses types, laboratory measurments, microbiology tests and sensitivity, prescription data and billing information.

The final dataset used for the inferential statistics project includes unique ICU admission of 46,234 patients’ demographic (age), vital (blood pressure, heart rate, body temperature, and Glasgow Comma Scale), underlying conditions (HIV, metastatic cancer, and hematologic malignancy), admission type (scheduled surgical, medical, or unscheduled surgical), renal (urinary output, and Blood Urea Nitrogen), and others (serum bicarbonate level, sodium level, potassium level, and bilirubin level) data. This dataset is build based on the commonly used mortality prediction tool, Simplified Acute Physiology Score II (SAPSII).

This report is a brief summary of inferential statistics results from a frequentist, bootstrapping, and Bayesian calculation. The code can be found [here](https://github.com/abebual/Clinical-Deterioration-Prediction-Model--Inferential-Statistics).

Using the frequentist theory I performed a hypothesis test of SAPSII total score, a score based on a total sum of scores of all 17 independent variables mentioned above, between ICU patients who survived the first 24 hours of their stay at ICU and those who were deceased during the first 24 hours of their ICU stay. The result from *ttest\_ind* (*scipy.stats*) shows there is a statistically significance difference in SAPSII total score between ICU patients who have survived and deceased within the first 24 hours of ICU stay (T=-113, p-value of 0.0). That is it gives us the probabilistic confidence to reject the null hypothesis - that claims there is no difference in SAPSII total score between survived and deceased.

Using the same final dataset developed for clinical deterioration prediction model, I made inferences about the ICU patients using bootstrapping (ie. Uses computer power to essentially re-run the sample draw again and again and again to see what actually happens.) Using bootstrap inference, the 95% confidence interval and a bootstrap hypothesis test for difference of means results were consistent to the frequentist theory. Performing the bootstrap (permutation samples and compute the p-value) replicates to observe a random difference between ICU patients survived and deceased during 24 hours of ICU observation.

Finally, using the same dataset, I employed Bayesian probabilistic model. Here the probabilities are framed not so much in terms of "how many times would I expect this event to occur if the experiment were to be rerun many times" but rather in terms of "what is my belief in the likelihood of this event occurring?" In a Bayesian probabilistic programming context, we can build models for systems and then let the data tell us how likely certain values for our model parameters are. This can be a very useful way to incorporate prior knowledge and deal with limited data.

We may suspect from the above that there is some sort of exponential-like distribution at play here. The probability of deceased at ICU may possibly be Poisson distribution. The gamma distribution may be applicable and we could test this for the distribution of SAPSII total score that weren't deceased. In this exercise, we have postulated a distribution to describe the individual SAPSII score for deceased ICU patients. This distribution has two required parameters, which we do not know, but we used PyMC3 to perform Bayesian inference to find our level of "belief" in a range of values for them. We then used the average parameter values to create one simulated data set of the same size as the original, but the distribution of our posteriors for these parameters will allow us to perform simulations of any sample size we desire and for a range of scenarios of different α and 𝛽. The result shows a plausible form for the distribution of SAPSII total scores of deceased and potential values for that distribution's parameters so we can already draw random variates from that distribution to perform simulations. The plot between simulated and observed data distribution shows consistent overlap.