The 35$^{\text{th}}$ Annual Workshop on Mathematical Problems in Industry: Data-driven predictive models for healthcare (Iterex group)

Manuchehr Aminian, Rituparna Basak, Elita Astrid Lobo,
Jenna McDanold, Richard Moore, Ruqi Pei, Geneva Porter,
Kosuke Sugita, Soheil Saghafi

June 17-21, 2019

Executive Summary

Background. Chronic Obstructive Pulmonary Disease (COPD) and sepsis are two conditions which have enormous health impacts in the United States. COPD refers to a collection of diseases resulting in airflow blockage and breathing problems affecting primarily chronic, long-term smokers. Fifteen million people in the US report they have COPD, and as of 2011 was the third leading cause of death, not to mention the economic costs. Sepsis is a similarly pressing health condition. According to the Mayo Clinic, sepsis is "*...a potentially life-threatening condition caused by the body's response to an infection*" and is particularly prevalent in those who are, or have recently been, hospitalized. Sepsis may lead to septic shock, which is a life threatening condition. The Centers for Disease Control report that approximately 1.7 million adults in the US develop sepsis, and accounts for one third of deaths in hospitals. Given these statistics and the failures of traditional approaches to prediction and treatment, it is clear that there are opportunities for improvement.

In modern years, the ongoing development of machine learning tools and the maturation of data analysis as a disciplined field have resulted in improvements in state-of-the-art in a wide variety of fields, including healthcare. These approaches have been particularly in their ability to develop predictive models for complex problems for which first-principles modeling may be out of reach, or infeasible in practice. For these reasons, Iterex has motivated our group at MPI 2019 to develop data-driven approaches to early prediction and triage decisionis related to sepsis and COPD.

In the context of the workshop, we were provided with two data sets related to each condition. With the provided COPD data set, our main objective was to accurately predict exacerbation events, while in the case of sepsis, patient time series were provided and the goal was to predict the onset of sepsis. We discuss our key findings for the two data sets in the paragraphs below.

Analysis of COPD. The data set provided associated with COPD is a collection of approximately 2600 scenarios; each containing a collection of vitals and self-reported symptoms. Each case was marked as resulting in an exacerbation event, and evaluated by one or more doctors, who provided a triage recommendation. We associated numerical values with categorical user inputs, and made decisions relating to either imputing data or discarding features, depending on missingness and appropriateness of those features.

Once the data was cleaned, initial visualizations were done to identify simple patterns in the data. One approach was to create pairwise scatter plots of the data by features using `pairplot` in the `seaborn` Python package. A few basic patterns were identified, such as the (expected) relationship between height, weight, and body mass index (BMI). Visualization also revealed clustering of the data into three age brackets. Principal Component Analysis was additionally used, but revealed no obvious patterns in the projection of the data on the first few the principal components, or in the principal components themselves.

We then attempted prediction of exacerbation events and identification of the most predictive features in the data. Exacerbation events were assigned 0 (nominal) or 1 (exacerbation) to which machine learning tools could be applied. One analysis used a Random Forest (RF) regressor with a typical training/testing validation scheme. We built a simple RF model using an implementation of RF in `scikit-learn` giving simple accuracy of 67.6% using all features; this also provided a ranking of feature importance. Repeating the process using the top six features resulted in an accuracy of 69.6%, effectively allowing us to neglect the remainder of the features for the purpose of classification. We expect in future work these could be further interpreted to build a more comprehensive understanding of the onset of exacerbation events.

Additionally, we used LASSO, an approach to build sparse linear models, to simultaneously predict exacerbation and identify important features. Sparse linear models have a trade-off of providing interpretable models but failing to identify strong nonlinearities in data. With this approach three features were identified in the training set and the model resulted in 84% accuracy (AUC= 0.88) on testing data – though we observed large false positive rates (approx. 45%) which could be further studied in the future.

Analysis of Sepsis. The sepsis data set was a collection of 4006 cases of hospitalized patients, 400 of whom succumb to sepsis, with a collection of forty vitals reported on an hourly basis. In addition to challenges of data missingness (many features are associated with lab tests which are only collected rarely), this data set is distinguished from COPD since the data are time series, which opens up dynamical approaches to prediction. We discussed advanced techniques towards time series analysis with machine learning such as recurrent neural networks and Long Short-Term Memory networks, but due to the many complications with this type of data, opted instead to work with linear models such as logistic regression (LR) and LASSO, or with random forests (RF) with the original features or possibly augmented features. For the purpose of comparison of models accounting for false negatives, we used the area under the curve of the receiver operating characteristic (AUC) as a holistic measure of performance. An AUC value near 0.5 is associated with random guessing of the label, while a value of 1.0 is a perfect separation of the two classes by the model. Given the time constraints of the project, we simply opted to impute features by median or modal value by feature across the entire data for our analyses, despite many biases introduced by such a simple approach.

Initial visualization and study of the relationships between features was performed, but no obvious relationships could be seen. To provide a benchmark for comparison, we implemented a clinical definition of systemic inflammatory response syndrome (a precursor to sepsis). The clinical definition incorporates four vitals and flags a case when two or more of these are outside of nominal ranges. This definition was able to predict the sepsis label across all data at an 87% rate, but had a 65% false positive rate.

Our initial approach was to predict the *current sepsis state* (via the provided `SepsisLabel`), rather than prediction of sepsis in the future, while throwing away time information and patient label (despite the many flaws in doing this). We applied LR and RF regression towards this task on the entire data set and obtained AUC values near 0.5 mainly due to severe imbalance in the sizes of data in each class. Attempts at training on balanced data by simple random sampling of sepsis-negative data (also flawed) resulted in AUC values on testing data in the neighborhood of 0.70.

A follow-up analysis was to create a "time until sepsis" value per-datapoint based on the patient, then ask the regressors to predict the time until sepsis (an arbitrary value of 99 hours was assigned to non-septic patients). This is the simplest approach to implementing time information without further feature engineering or implementation of more complex time series tools. We found the regressors succeeded in predicting these values on testing data to within an hour (as measured by root mean square error) but failed to naively associate this with the current `SepsisLabel` (AUC= 0.72). Much follow-up work could be done here given the raw predictive power we observe.

Our most successful machine learning approach included two extra steps. First, additional features were created based on moving averages of patients' data, to reduce noise and incorporate time information (although interpretation is challenging). Second, rather than predicting the current sepsis label, the models were asked to predict whether a data point comes from someone *who will at any point* become septic in their time history. This approach proved much more successful (AUC= 0.93) but many follow-up analyses need to be done to verify how robust this result is, and to what degree the combination of moving averages and variability in time series for sepsis-positive patients dilutes its predictive power in practice. Nevertheless, we believe this is a promising approach worth further investigation.

Distinct from our machine learning approaches, several group members learned about and began developing a Kalman filter approach to analyzing the time series. Kalman filters are a popular tool to incorporate observable data (such as temperature, lactate levels, etc) with unobservable underlying states (such as pathogen levels and the detailed immune response). This allows one to use a principled mathematical model and couple it with real data in a way not otherwise possible, inferring the underlying state of the immune system as new data arrives in an intelligent way. To do this, our group members took a minimal model from a recent paper by Ramirez-Zuniga *et al*, describing immune response to pathogen in 8 state variables – one of which is lactate level – and implemented the Kalman filter update equations in Matlab. Ultimately, while we could not make this approach work while incorporating data from the provided data set, we feel that this will be an impactful approach to solving this problem to be pursued in the future.