Data Driven Medicine : Using Post-Operative Vital Signs as Markers for Pancreatic Cancer Surgical Complications

Sneha Iyer
Department of Mechanical and Aerospace Engineering
sgiyer@princeton.edu

Preeti Iyer
Department of Computer Science
psiyer@princeton.edu

Abstract

This study aims to formulate a more intelligent and proactive way to detect post operative complications in patients undergoing pancreaticduodenectomy surgeries ("Whipple" surgeries). Using vital signs in the days following surgery and patient medical history, we aimed to gain a better understanding for factors that effect post operative complications and create a predictive model to help health systems have a comprehensive way to monitor patients and provide appropriate medical attention those who are at high risk for certain complications. The study used both unsupervised and supervised machine learning to analyze if meaningful clustering could be the done with patient profiles and see if certain vital signs at key post operative times could give more insight into the eventual onset of a post operative complications. Several interesting trends were found and at large heart rate was the best predictor.

1 Introduction

A pancreaticduodenectomy ("Whipple surgery") is a high-risk surgical procedure performed for tumours of the pancreatic head and other periampullary structures; while mortality has drastically decreased over the years for this high risk procedure, still the rate of post-operative complications (where complication was defined as a predetermined list of medical issues by the Department of Pancreatic Surgery at TJ Hospital, these included conditions like Sepsis, Chyle Leaks, Bleeding, Pneumonia, Urinary Track Infections,etc. which happened at some point usually 4 days or later out of surgery) remains high and there is a need to better understand the nature of common and unusual post-operative complications to hopefully inform predictive analytics which can provide more proactive care during post operative recovery. We hope to get specific insights into ways different vital signs and patient history data can be used to more intelligently monitor patients. More broadly we hope to truly find a way to make post operative complications more manageable for hospitals by translating our findings into tangible deliverable for a hospital's post operative team. We dug more into relevant work in the field to understand what was needed to make this study transnational, but also found that much of the analysis which is most medically relevant in this field is not in fact machine learning focused (more simple analysis actually might be more transnational as the applications of some machine learning analytics remains tricky in a clinical setting like this), so we struck a balance to make sure our core study was still motivated by the techniques focused in this class.

1.1 Related Work

Data-Driven Medicine is a new and burgeoning field with so much potential to drive more insightful, strategic medical decisions. Hospitals, labs, and medical companies alike are starting to utilize the

powers of analytics, computation, and pattern prediction to better understand latent structure with medical phenomenons, especially given there is already such massive amounts of raw medical data recorded on a continuous basis for all patients.

Some facilities such as Duke Medical Center have already adopted many predictive medicine and machine learning implementations to drive proactive care and increased efficiency. Their implementations in ways such as a predictive algorithm for live-time changing risk of cardiac failure in Emergency Medicine and procedure outcome estimations allow for additional insight that is impossible for doctors and humans to manually detect. [1] Freenome is a company based in Palo Alto that uses AI and intelligent analytics to predict cancer by running pattern recognition on blood work. [2] We actually got the chance to visit Freenome and talk to their CEO, through Princeton's TigerTrek program. It was fascinating and inspiring to learn about the cross-functional work they are tackling.

In particular to Vital Signs (VS) and their involvement in every medical decision and monitoring period, many research groups have started to explore the trends and hidden predictive capabilities that may lie in these biomarkers. For example, the University of Virginia is conducting an ongoing study using the variance of heart rates to predict neonatal mortality. [3] It is generally understood that there are linear trends with VS intensity and complication outcome but the specifics of which ones are particularly important and relevant for specific cases has not been properly understood.

As Thomas Jefferson Hospital has relayed in our meetings with them, a translational application of this and ongoing studies is the redefinition and hyper definition of what exactly is important in medical settings. A fever, generally understood in the medical community to be 99.5 and 100.9 Â°F, is an indication that something larger in the body is happening. However, this definition of a fever has been extrapolated from other cases and applied to medicine as a whole. For temperature and all other VS, there can very well be better thresholds and indicative numbers that are better suited to be the value at which other care and monitoring must be administered. By performing an analysis specific to pancreaticduodenectomies, we can unearth precision and correctly-focused medical parameters that can drive better post operative care.

2 Data

The data comes from the Thomas Jefferson Hospital (TJH) Department of Surgical Research, Pancreatic Cancer Division. Sneha helped the origins of building and the initial use of this database as a part of a summer internship 3 years ago. The Department of Surgical Research at TJH is pursuing exploratory research to find ways that past outcomes and systematic complications can be detected through means that are already accessible (vital signs, lab results, standard post-operative patient care). The data is an aggregation of patients over a 5 year period who underwent a Whipple surgery. As part of summer work, the data was initially cleaned alongside trained medical researchers at TJH to take into account appropriate medical metadata. An initial proof of concept analysis was performed in 2017 showing that data-driven approaches can be used to predict post-operative complications for this data set. TJH and the doctors in the Cancer department gave full permission to use this dataset and we've established an ongoing partnership to progress this project past the extent of the course. The data used had all private medical descriptors permanently erased before it was given to us as well as randomized all patients as an extra precaution all though not necessary.

From previous work from the TJH Department of Surgery and supplemented by our own initial exploratory analysis, it was determined that the best time frame of data to use as predictive data would be the 4 days following surgery. Since 85% of complications occur after the 4th day of surgery, we want to the longest amount of data which will still be useful for predicting future complications; using more data was considered but it was finalized that this study would be limited to using Vital Sign data from the 96 hours following surgery.

While the data proved to be a rich source for medical analysis it was also limiting in that so much additional data cleaning was required to get it into a machine learning ready format, and also due to medical regulations and constraints on patient confidentially / sending data remotely we were not able to get detailed demographics on each patient beyond some prior medical history.

2.0.1 Data Processing and Feature Engineering

In summary the data set contained usable data for 707 patients (once unusable data was dropped) with 100 features, once the features not relevant to our direct study (such as date of stay, detailed surgery classifications, etc) were dropped and the resultant data set was 707 patients with 53 features each. The following characteristics were available for each patient: *Prior Medical History data of prevalence of: Cardiac History, Smoking, and Chronic Obstructive Pulmonary Disease; 96 hours of vital sign data recording during 8 hour intervals including- Heart rate, temperature, and blood pressure (systolic and diastolic)*

A good amount of pre-processing had to be done before we were able to use the data. Due to typos by nurses and other medical staff some of the encoding for certain complications were messed up and we had to intelligently deal with the incorrectly formatted data and typos; we also had to convert all string answers that often offered short descriptions into binary encoding indicating the prevalence of such a complication and other extracted information. In accordance with the literature we also set thresholds for severe complications by following the Clavien Grade - a scale of severity for medical issues specific to Whipple surgery patients - a threshold of a Clavien Grade of 3 was determined. There also were still some patients with unrecorded data for key periods of time int this time interval - once we dropped patients who had the majority of their data not recorded correctly (determined by looking at if temperature right after surgery was recorded which proved to be a proxy for what we were screening for), there was a remaining small amount of missing data ($n = 38$). *Imputation based on the mean* was performed for this small amount of data; it was determined that this was the best choice since.

From exploratory analysis, it was also determined that our predictive classes of complication rates, infectious complication rates, and hospital re admissions were imbalanced (with mostly lower proportions of patients in the positive classes for each), so Synthetic Minority Over-Sampling Techniques were preformed (using the imblearn library) to balance the classes in training data.

Lastly, the squares of all data were taken and added to the training sets in the hopes that the additional feature engineering would improve performance and increase the robustness of the model. However, not much performance gain was seen, so this was ultimately not analyized further.

2.0.2 Initial Exploratory Analysis

The data set contained a population with a 57% complication rate and a 38% infectious complication rate. 19% of patients had a complication of Clavien Grade 3 or higher (severe). The 90 day mortality rate was 3.9% .

2.1 Methodology

In order to do a holistic and comprehensive analysis of the usefulness of the first 96 hours of post operative vital signs data in informing proactive medical treatment related to future complications and other medical diagnostics, our approach was structured as follows:

1. Unsupervised Learning with Goal of better understanding structure and possibility of clustering behind distributions of patient profiles
For this we used Gaussian Mixture Models - a probabilistic model for representing normally distributed sub populations within an overall population utilizing the expectation maximization algorithm. We used Guassian Mixture Models as opposed to the K-Means clustering algorithm because it's more robust and applicable to deviating data sets of continous data. We looked further into the scores assigned and partitioned the data set based on the predictions and expected clustering probabilities to understand the latent structure and confounding features.

2. Supervised Learning with Goal of predicting complications
We used a variety of different classifiers and compared the respective performances of each. We also did distinct analyses for predicting the prevalence of (1) a complication at large, (2) an infectious complication (subset of the complicating set), (3) Readmission (having to be admitted into the hospital again at a later date because of a complication); not all of these turned out to be significant in our analysis but all were initially looked at. The following classification methods were used in

some capacity, and hyptertuned with either the built in cross validation to the classifier or GridSearch Cross Validation techniques.
1. Random Forest Classifier (RFC)
2. Logistic Regression Classifier using 5 fold cross validation (LR)
3. Support Vector Machine Classifier (SVM)

While we can extract meaning from the relative performances and feature importances of certain classifiers, it would be medically useful to understand which values of vital signs for significant time periods are important - thus we decided to build and train and decision tree with a subset of the most important features determined from previous steps of the analysis (to limit the size of the tree and make node traversals interpretable). Thus a forth classifier was used for this extension on our classification analysis:
4. Decision Tree Classifier - used for threshold analytics

2.2 Evaluation

For the Gaussian Mixture Models, We chose the amount of clusters based on the Akaike information criterion (AIC) and Bayesian information criterion (BIC) by iterating through 2-10 clustering options and finding the lowest AIC/BIC score in balance with how many clusters could be useful and interesting to look into. Once we applied the learning mixture model for our purposes, we dug into the distributions of the time series post operative vital signs at critical time intervals and distinguishing profiles between our sub populations. This is further explained in the results section.

For an initial evaluation of the classification methods we used Accuracy, F_1 score (a metric which is a function of both precision and recall sought in this case because we wanted to evaluate on a balance between precision and recall), and most importantly the % of False Negatives. In our case % of False Negatives was critical since incorrectly classifying a patient as likely to have no complications when they actually had one has detrimental medical applications (and is not practical for translational purposes as we saw in our literature reviews) so we want to pick a classification that skews in the opposite direction - we can be conservative and overclassify patients as being likely to have complications and potentially offer them additional medical attention that might not be needed with just some financial and labor repercussions for the hospital. From our performance results, shown in Table 1 below and Figure 1, we can notice that the logistic regression classifier gave us a relatively high accuracy along with the lowest rates of False negatives. It was also noted that infectious complications performed much better for classification which makes sense, since medically these are tied closer to the presence of a fever or abnormal heart rates (as seen in our background) and this informed the focus of our remaining study to be on infectious complications and used the Logistic Regression Classifier to closely dig out feature importances (seen in the results section).

Complications, At Large				
Classifier	Accuracy	F_1	% FP	% FN
RFC	.64	.67	34.6	39.3
LR	.58	.58	35.9	46.2
SVM	.43	.08		
Infectious Complications				
Classifier	Accuracy	F_1	% FP	% FN
RFC	.69	.53	48.1	24.4
LR	.70	.54	47.2	23.8
SVM	.62	.42		

Table 1. Classifier Performance

3 Results

3.1 Diverging Patient Clusters

Using Gaussian Mixture Models 4 patient sub-populations were found and 2 were significant in their interpretation. There was visibly a cluster (Cluster 3 in below figures) which had the highest infectious and at large complication rates, highest average length of stays, and the most pre-surgery

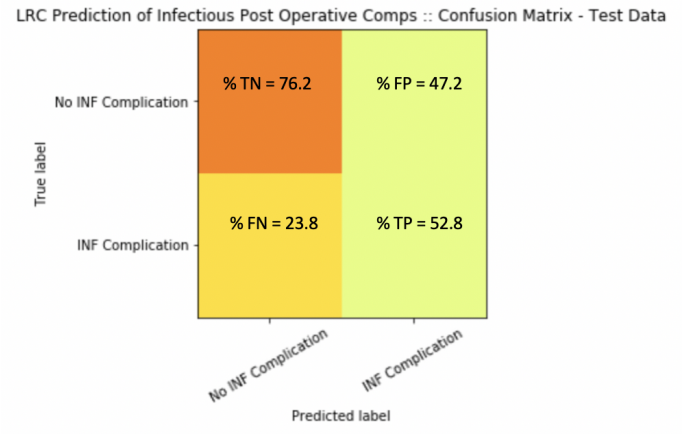


Figure 1: Confusion Matrix for LRC of Infectious Complications

health flags (i.e the highest portion of smokers, and prior cardiac history) - we can refer to this population as more unhealthy patients who at large had some post operative surgery and stayed the longest in the hospital. There was another cluster which also had high rates of infectious and at large complication rates, but also had the highest average prevalence of a sever complication (by their Clavien Grade Scores) and very low indication of any of the mentioned pre-surgery health flags these patients by far had the highest readmission rates - we can refer to this population as healthier patients who had less overall issues on aggregate, but had the most severe issues and we re-admitted the most. (See Figure 3)

From Figure 2 it can be seen that these clusters (cluster 2 and 3) had a very diverging heart rate from the other clusters, with very high rates compared to other clusters. This shows that heart rate is significant in being a diverging predictor of possible complications and re-admittance, whereas another vital sign which still seemed important during the supervised study (see below) appeared more uniformly changing between all clusters. It was also interesting how for almost all of the vital signs - an aggregate time series views offered a look at certain spikes and drops in readings between all clusters.

3.2 Heart Rate as a predictor

The most important timed predictors of infectious and overall complications are shown in Figure 3, drawn from the feature importances of the built classifiers. It's interesting because there seems to be a cyclical pattern for the most important features for predicting infectious complications - at each 24 hour period after surgery the heart rate is extremely important for prediction. Also while the most important features for complications at large were all in the fourth day of post operative monitoring, heart rate offered an earlier predictive marker for infectious complications. In general, as discussed, it was seen that infectious complications were better able to be predicted.

Heart rate was a strong indicator for infectious complications, and we saw that at 72 hours after surgery both heart rate and temperature were important in the classification. The predictive probabilities at all temperatures and heart rates at this time stamp were plotted and it was apparent that there was a positive trend between heart rate and the predictive probability of an infectious complication and temperature and the predictive probability of an infectious complication. This is expected because in general higher absolute heart rates and temperatures indicate that some medical condition might be more likely, but we expected there to be some confounding effect from patients on the other end of the spectrum – those with extremely low temperatures and heart rates that also got infectious

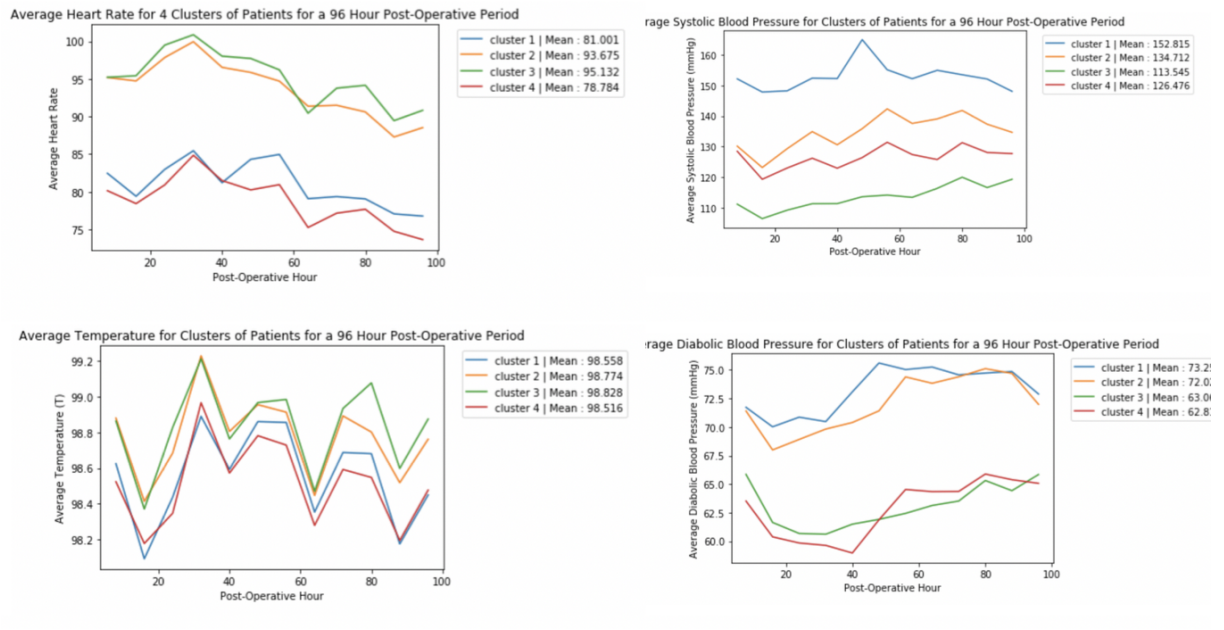


Figure 2: Average Vital Signs for 96 hr post operative period, by cluster

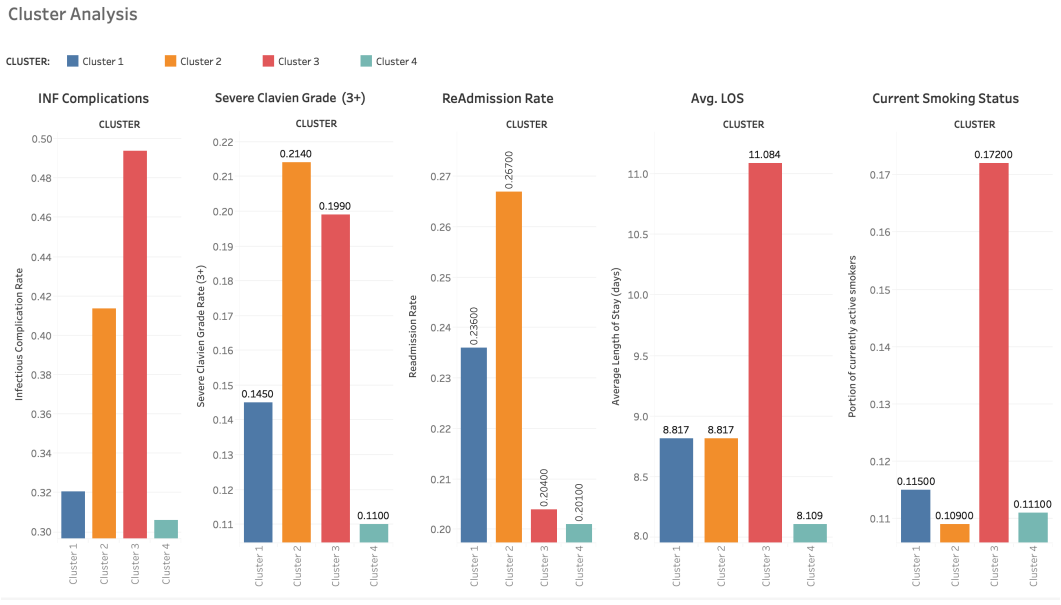


Figure 3: Descriptors Separated by Patient Cluster

complications. However its possible this subset was too small for the trained model to pick up and give importance in the overall classification.

Since we are particularly interested in the results with respect to false negatives and positives, we stratified each of the false classifications for infectious complication prediction at 2 important heart rate time frames (at 24 and 48 hours) to better understand the missclassifications. As can be seen in Figure 5, False negatives on average occurred at much lower rates rates and false positives at higher.

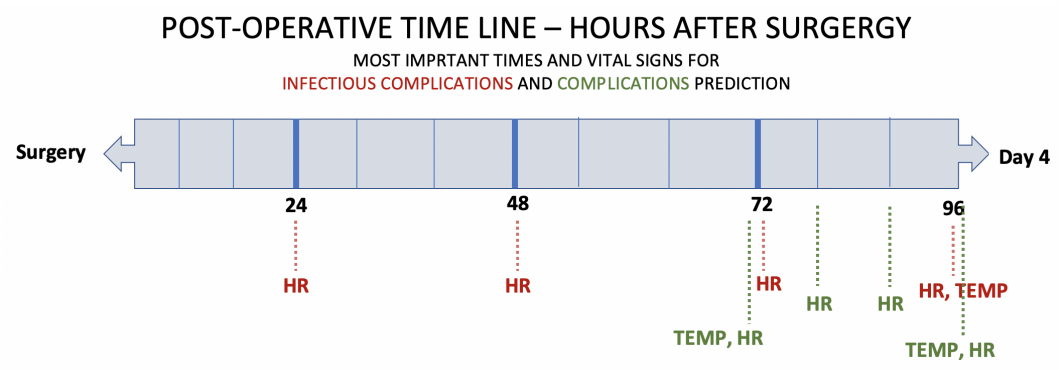


Figure 4: Most important time stamped Vital Signs for Prediction

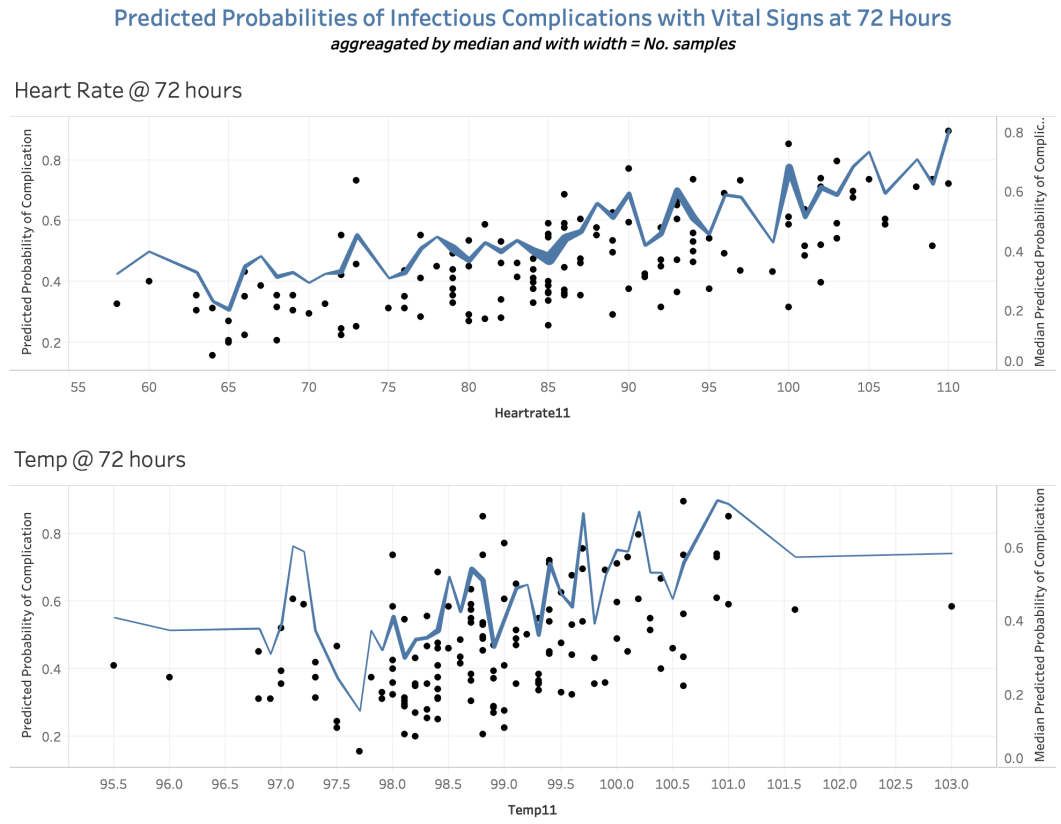


Figure 5: Predictive Probability of Infectious Complication at 72 hours

This makes sense with the way our classifier formulated its model - leaning towards predicting infectious complications for higher heart rates and temperatures.

4 Discussion and Conclusion

Our results section went into detail about the findings but at large we noticed some very interesting things and were able to find both significant sub-populatations of patient profiles and important timed predictors of post operative infectious complications.

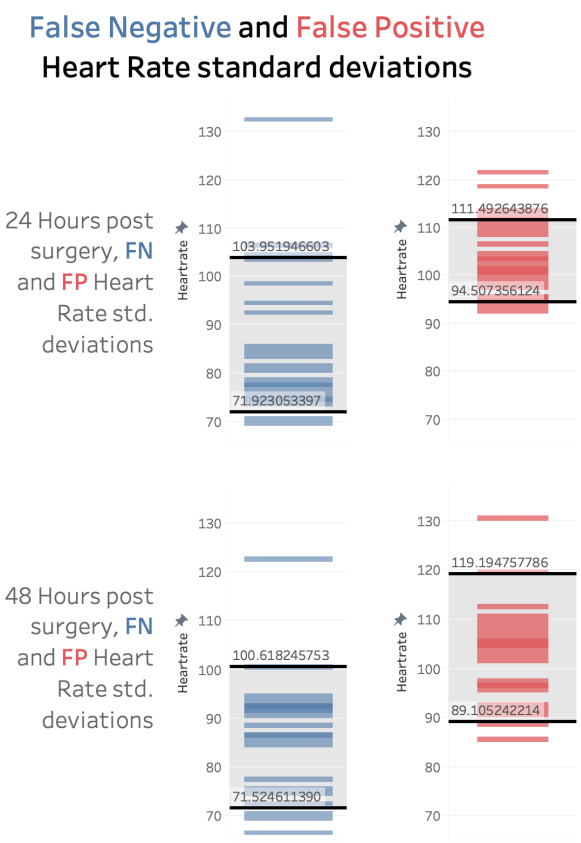


Figure 6: +-1 Std Deviation of HR means at 24 and 48 Hours

At large it seems that those healthier patients who develop complications tend to have more severe complications and in turn be readmitted down the line whereas those who had pre-surgery health-flags on average had higher complications but those that were on average less severe and these patients on average stayed in the hospital the longest. Heart rate also emerged as the most robust descriptor (in both supervised and unsupervised learning) to diverge between patients with eventual complications and those without, while temperature seemed to follow more uniform trajectories between all patients.

The most predictive features of infectious complications were heart rates at 24 hours increments after surgery and it offered earlier prediction ability than complications at large for which features in the final hours of our time period of relevance best predicted onset complications. To better understand the intricacies of the variation of the important features, a decision tree was built with a subset of the most important features shown in Figure 4. To implement checks for high risk patients in a clinical setting it will be useful to have thresholds which can ultimately be developed into something like checklist administered by medical staff where patients' vital sign readings are checked at critical times and informed to decided what action to take based on set values of importance; we can potentially use the decision tree node splits as an informative basis for this and in Figure 7 (See Appendix) it can be seen that the built decision tree does offer node traversals that can be further iterated for a clinical setting. Most of the traversals that led to infectious complication positive classifications were heart rate decisions at the key times we analyzed.

Our next steps are further iterating on the decision tree thresholds for the key time intervals we determined and further looking into the interesting 24 cyclic pattern of heart rate as an important predictor for infectious complications, by looking at a larger sample size and hopefully getting more patient demographic information like age to pair with the current data. We look forward to working with TJH in the upcoming months to make this work translatable into proactive data driven medicine!

please note that the figures got misaligned last minute and we were not able to fix it in time, however everything is labeled by figure number so please use that to guide reading. We apologize for the inconvenience.

5 Acknowledgments

We'd like to acknowledge Professor Barbara Engelhardt and Jonathan Lu for their assistance in talking through concepts and ideas and the learned course material, during the course of this project. We also would like to thank TJH for their collaboration with this study.

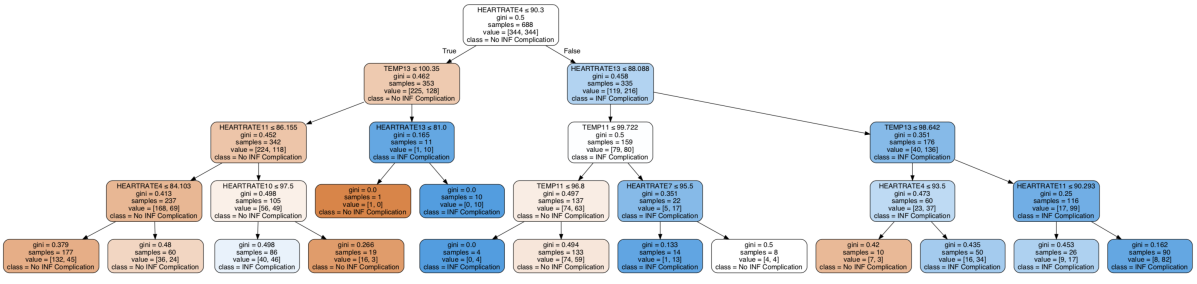
6 Bibliography

[1] Duke Clinical Research Institute. âCenter for Predictive Medicine.â DCRI, dcri.org/our-work/analytics-and-data-science/center-predictive-medicine/.

[2] âUsing AI Genomics for Early Cancer Detection Treatment.â Freenome, www.freenome.com/.

[3] This scientific study was given to us by TJH Department of Surgery. It is currently not availible to the public.

7 Appendix : Supplementary Figures



tree.png

Figure 7: Decision Tree Built with most important features for Infectious Complication Prediction