CLININCAL DATA SIMULATION AND PATIENT CLASSIFICATION

COMPUTATIONAL AND SYSTEMS BIOLOGY 185, WINTER 2020
GROUP PROJECT PAPER

Deepthi Gangiredla **Angela Gu** **Shaili Mathur** **Saahil Patel** **Aditya Pimplaskar** **Kaushal Rao**

March 17, 2020

ABSTRACT

Calculating postoperative mortality risk score is an important element to consider before surgical procedures, and this can help physicians evaluate the feasibility of even performing the procedure on the patient. Here, we aimed to evaluate the performance of statistical models in predicting patient mortality risk. We simulate electronic health record data using a graphical model, and use a random forest model to classify patient outcomes and predict mortality risks in patients. We found regression on EHR data was more proficient at computing a continuous survival probability compared to computing a binary classification of mortality.

Keywords Data Simulation · Clinical Data · Random Forest

1 Introduction

Pre-operative prediction of mortality risk is an important assessment in determining a course of treatment in critical care settings. Currently, pre-operative mortality risk is predicted using the American Society of Anesthesiologists (ASA) Physical Status Classification. This is a useful metric, but is generated subjectively by a clinician. Previous work in the field (Hill, et al) has shown that machine learning on EHR data can be used to predict pre-operative mortality risks in patients. Hill et al identify the most efficient statistical model to generate an alternative alternative to ASA scores, and found that the random forest classifier model was the most proficient predictor of mortality scores.

Here, we study the MIMIC-III dataset, a publicly available dataset which has de-identified patient records associated with over 40,000 patients who stayed in critical care units of the Beth Israel Deaconess Medical Center between 2001 and 2012. We identify features in clinical data that are commonly available in ICU patients, and simulate data with these features.

We simulate four datasets with differing levels of variance in the dataset. We use a hierarchical graphical model to simulate our data. We incorporate empirical mortality rates in patients across age and sex in our model and define relationships between general health, organ system status, and clinical measurements with realistic values. We fix whether a patient survives based on their assigned mortality rate, and infer their outcome and mortality rate from only clinincally measurable features. We implement principal components analysis, a random forest classifier, and a random forest regression model, and evaluate their effectiveness.

2 Methods

2.1 Feature Selection and optimal values

Features for our model were selected based on two main factors: preoperative features that were tested as predictors of mortality for the models in Hill et al. (creatinine, urea, platelet count, INR/prothrombin time, arterial blood pressure, heart rate, white blood cell count, oxygen saturation, age and sex) and features that were not used in Hill et al. but are general biomarkers for disease (LDL, HDL, red blood cell count, and urine volume). Since the MIMIC dataset did not have ASA statuses, Apache II mortality scores and the Coefficient of Hospital Mortality (CHM) were used as replacements in our features. To provide a reference for our simulated data, the range of optimal values for each lab-collected feature was found as well.

2.2 Data Simulation

We use a graphical models to generate data on health measurements of patients in critical care units. In this model, each value is drawn from a distribution whose parameters depend on the parent node and on the level of variance that we want set in the data, and on empirical age and mortality rate distributions for the top level nodes (in blue in the Figure below). We generate four sets of data, with no variance, low variance, medium variance, and high variance.

We categorize health measurements into four groups: kidney function, immune response, cholesterol panel, general blood measurements, and hospital mortality predictions. The graph we construct is shown below. The blue nodes are high level choices, and the orange node is an abstract health score for a particular organ system. The age of the patient is drawn from the age distribution from Atramont et al, and the age along with the sex is used to predict the mortality rate (also from Atramont et al). Each node is drawn from a distribution with parameters depending on the value of the parent node. The survival probability (health) is drawn from a a Gaussian distribution with mean defined by average empirical mortality for the patient's age and sex, and variance (σ^2) which is fixed by the variance level of the data. The values for organ system level nodes (in orange) are drawn from a Gaussian distributions with the survival probability as the mean, and the same fixed standard deviation used previously to generate the survival probability. The values of σ^2 are 0.0, 0.05, 0.1, and 0.2 for the no variance, low variance, medium variance, and high variance respectively. These values are then used to generate the measured values for the 14 clinical features (in red).

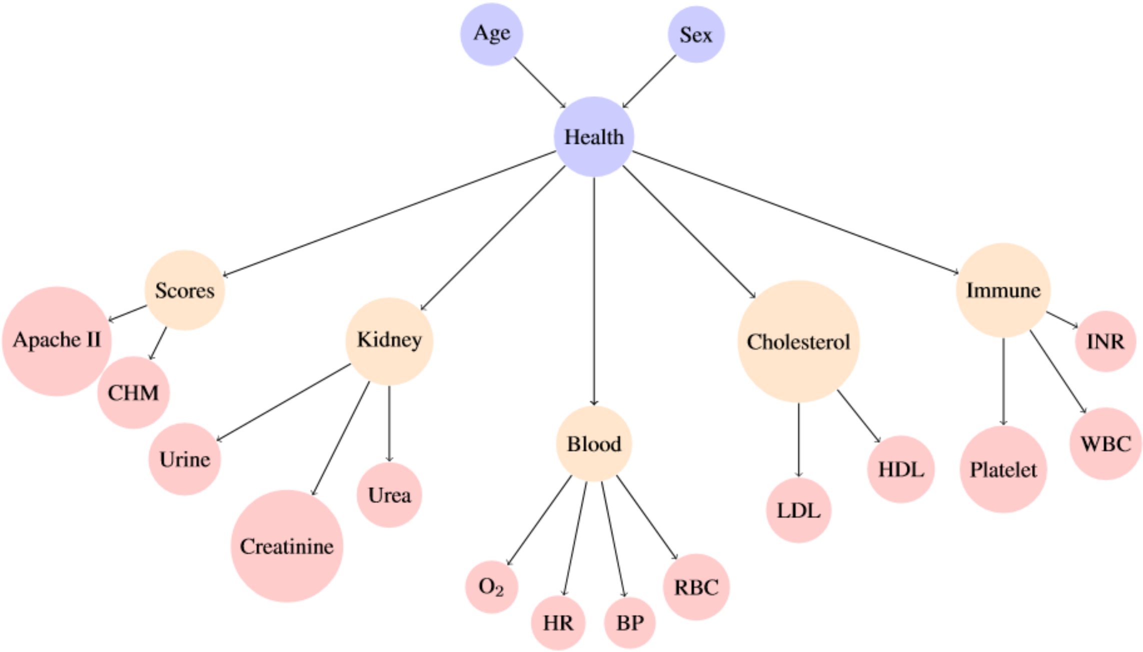


Figure 1: The graphical model which we use to simulate our data. The values for each node in an individual is drawn from a distribution which depends on the parent node. The top level is the most general set of features, age and sex, which determine the patient's survival probability. The orange nodes are abstract health scores for organ systems, and they determine the realistic measurements for the clinically measurable features (in red).

In order to generate these values, we assumed that patients with high values for the parent node of a particular feature would be likely to be close to the optimal value, and patients that are have lower scores must have measurements that are higher or lower than optimal. We used a piece-wise modified Hill function to model this mapping between overall health and realistic measurements:

$$v(H) = \begin{cases} v_{opt} + (v_h - v)\left(\frac{1-H}{H}\right)^{\frac{1}{n}} & \text{if } v \geq V_{opt} \\ v_{opt} - (v_{opt} - v_l)\left(\frac{1-H}{H}\right)^{\frac{1}{n}} & \text{if } v < v_{opt} \end{cases}$$

We find values for each measure that are considered dangerous (either too high or too low), and assume that they correspond to a health value of 0.5 on 1. For each feature we find values that are considered optimal, dangerously high, and dangerously low, v_{opt}, v_h, and v_l respectively. The parameter n controls the sharpness of the curve, and we set it to $\ln 9/\ln 2$, so that health scores of 0.9 correspond to values halfway between optimal and dangerous values. The values are generated by draning from a Gaussian distribution with mean determined by the value that the health score mapped to. In Figure 1 we show this curve for an example feature, urine volume, along with the data corresponding to different levels of variance.

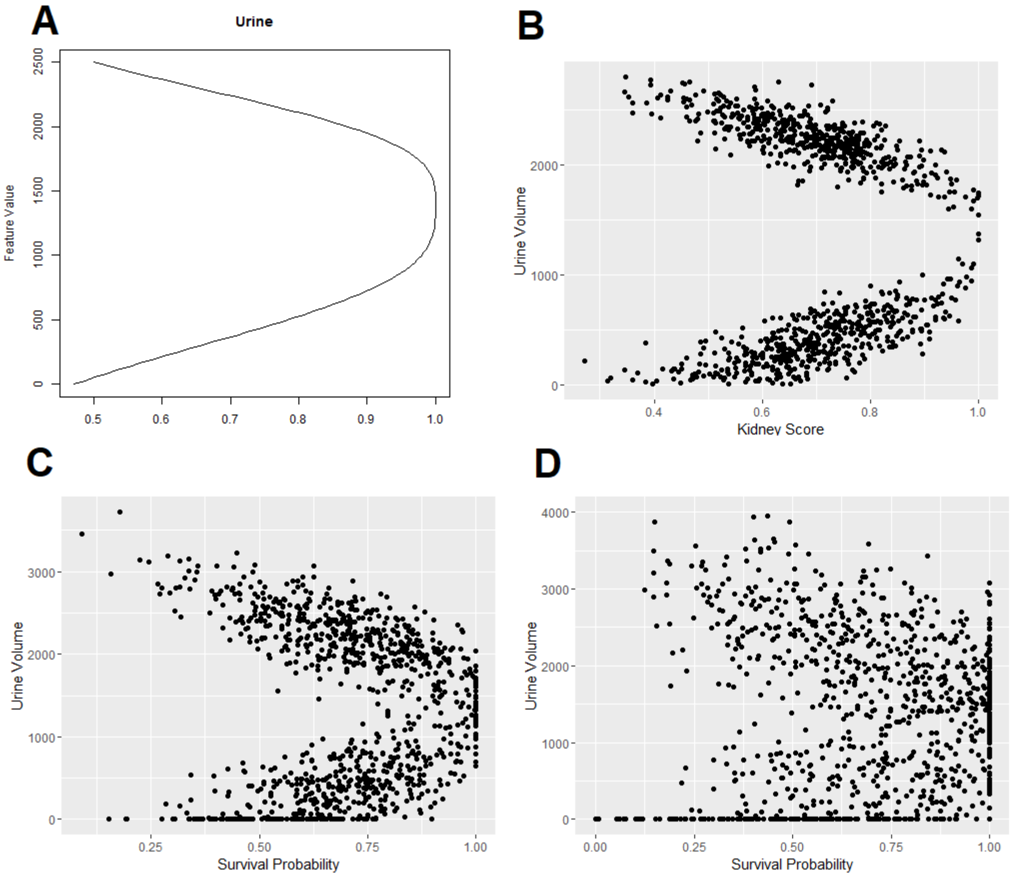


Figure 2: A graph of the modified Hill function (Panel A) for urine volume, and sample data generated from the graphical model across different levels of variance. Panel B corresponds to the low variance dataset, Panel C to medium variance, and Panel D to high variance.

2.3 Principal Components Analysis (PCA)

PCA is a method of linear dimensionality reduction, which generates principal components, which are coordinates in a coordinate space created by orthogonally projecting the data so that the coordinates explain as much of the variance as possible. Principal component 1 (PC1) explains the most variance, PC2 the second most, and so on. We used this method to reduce the dimensionality of our datasets with included clinically measurable features along with demographic features such as age and sex. We used the `prcomp` function in `R` to do implement PCA.

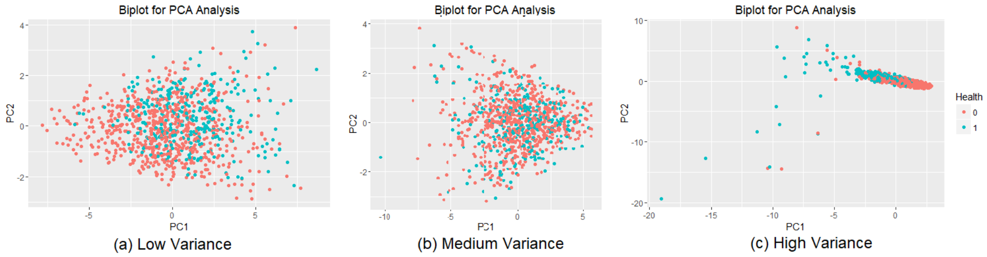
2.4 Random Forest Model for Classification and Regression

The random forest algorithm is a statistical model used for classification and regression. The Random Forest consists of an ensemble of decision trees, where the outputs of the individual trees are ultimately utilized to produce a final output. The algorithm initially randomly selects k features from the dataset for each decision tree, and among the features, the best node to split the data on is calculated using metrics like Gini index. This process involves selecting nodes and splitting data on the to maximize differences in the data after the split. After the iterative process of splitting data, predictions can be made by each decision tree. The final output of the algorithm is the majority-voted decision for classification or the average prediction for regression. Random Forest minimizes over-fitting if there are enough decision trees in the forest. Also, the classifier of Random Forest is robust enough to handle missing values as well as categorical data. Because of these factors, the model accuracy of Random Forest is generally higher than other methods such as logistic and linear regression. We utilized the random forest classifier and random forest regressor from Python's scikit-learn library to make classification predictions of mortality and regression predictions of survival probability. A classifier model and regression model were created for each of three simulated datasets, each created with a different variance value.

3 Results

3.1 PCA Analysis

We performed PCA on all of our datasets, and note that PC1 corresponds to the outcome of the patient. In Figure 2, we see that there is horizontal shift between patients that survived (pink) and patients that did not (blue). However, we note that there are no obvious clusters that are formed, and therefore it is difficult to classify patient outcomes based on PCA analysis.



(a) Low Variance

(b) Medium Variance

(c) High Variance

Figure 3: Biplot of PCA analysis for data with low variance, where Health indicates the outcome for the patient; 0 indicated death, and 1 indicates survival. We see a significant shift to the right, indicating that PC1 predicts patient outcomes.

3.2 Random Forest Model

We simulated data for features from distributions that used three different parameter values for variance (to simulate noise). When running our random forest model against this data, we found improvements in its predictive capacity as the variance parameter decreased. This aligned well with our simulation method; as the variance in our data decreased, model accuracy should have theoretically increased as the prediction process became more deterministic. We ran a random forest classifier to predict a binary mortality status based on our simulated features. We proceeded to assess our models' classification performance using two main forms of data visualization: confusion matrix heatmaps and ROC-AUC curves. These AUC values were not very potent - they achieved AUC values just above 50% for each variance parameter value, signifying that our model's ability to distinguish false positives from true positives was only marginally better than random chance. This was not an ideal result, as it rendered our classification model relatively useless.

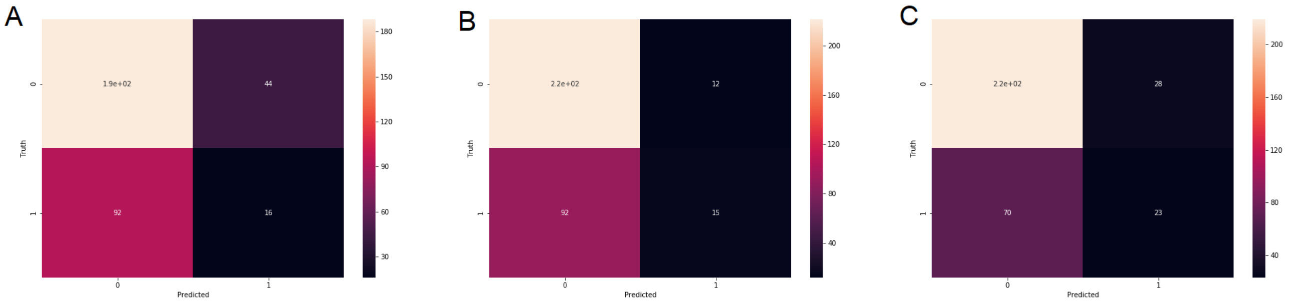
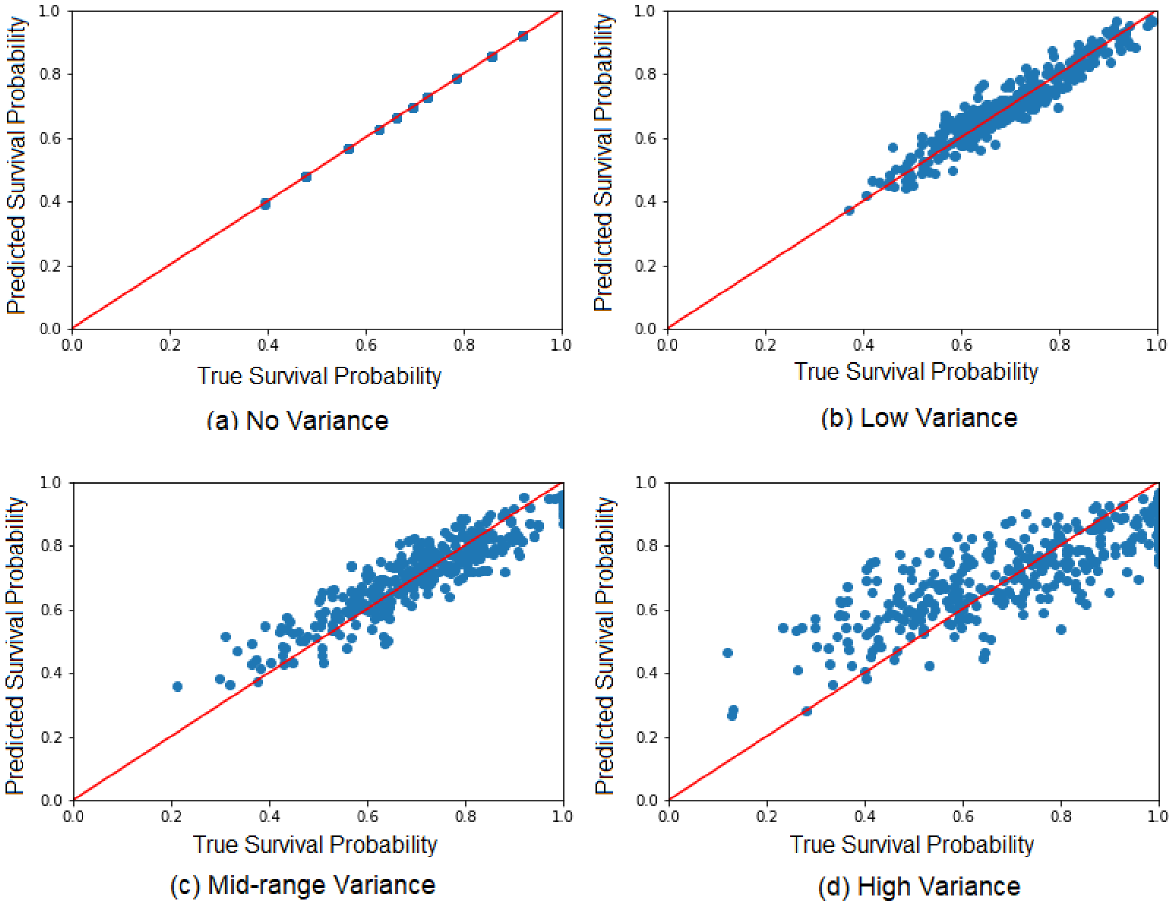


Figure 4: Confusion Matrices from Random Forest Classification for the three levels of variance in the data. Panel A is no variance, Panel B is low variance, and Panel C is mid-range variance. The classification was not done on the high variance data as the classifier was performing poorly on the first three datasets.

We found that, given the amount of noise and variation in our data, discretizing our outcomes to binary dead/alive classifications led to a lack of predictive power, as well as a general loss of information. As a result, we decided to additionally train random forest regression models in order to predict survival probability as a continuous percentage. We found our new models had greatly improved performance, achieving scores ranging from 99% at the lowest variance to 61% at the highest variance.



(a) No Variance

(b) Low Variance

(c) Mid-range Variance

(d) High Variance

Figure 5: Plots of predicted versus true values of the probability of a patient's survival generated by the Random Forest Regression model. Each data point (patient) is in blue, and the red line is the true value. Panel (a) is the dataset with no variance, Panel (b) is low variance, and Panel (c) is mid-range variance, and Panel (d) is high variance.

When we examined features and their importance (characterized by Gini Index), we did not find any salient trends between different variance models. We expect this is a result of the simulation and pulling from distributions, though in the case of analyzing real data, there may be more noticeable trends. Some of the leading indicators in our models were age, blood measurements, and cholesterol levels.

4 Discussion

Our main finding was that classification was a more robust predictor of mortality, as it preserved resolution in the data. Simulation was a useful tool in creating preliminary models and understanding trends in our data. We had initially attempted to utilize the MIMIC-III public database, but we had trouble with generating enough data after filtering for relevant patients. We decided to pivot and simulate data for the input features we wanted to use, and purely used this data for assessing the accuracy and performance of our models. In the future, we envision the use of our random forest models trained with real clinical data to predict patient mortality. Since our regression models performed significantly better than our classification models, we believe that predicting continuous features (such as survival probability) would serve higher utility than predicting binary features such as mortality. Since health data for patients varies over time and visits, we would aim to construct a model to assess the role of temporal dynamics. Implemented successfully, these kinds of models can aid clinical decisions relating to patient care.

References

1. Beckerman, James. “Understanding Cholesterol Levels: LDL, HDL, Total Cholesterol, and Triglyceride Levels.” WebMD, WebMD, 6 July 2018, www.webmd.com/cholesterol-management/guide/understanding-numbers.

2. Bleyer, Anthony. “Urine Output and Residual Kidney Function in Kidney Failure.” UpToDate, 3 May 2018, www.uptodate.com/contents/urine-output-and-residual-kidney-function-in-kidney-failure/print.

3. “Blood Urea Nitrogen (BUN) Test.” Mayo Clinic, Mayo Foundation for Medical Education and Research, 2 July 2019, www.mayoclinic.org/tests-procedures/blood-urea-nitrogen/about/pac-20384821.

4. Chiavone, Paulo Antonio, and Yvoty Alves dos Santos Sens. “Evaluation of APACHE II System among Intensive Care Patients at a Teaching Hospital.” Sao Paulo Medical Journal, Associação Paulista De Medicina, 1 Nov. 2002, www.scielo.br/scielo.php?script=sci_arttextamp;pid=S1516-31802003000200004.

5. “Complete Blood Count (CBC).” Mayo Clinic, Mayo Foundation for Medical Education and Research, 19 Dec. 2018, www.mayoclinic.org/tests-procedures/complete-blood-count/about/pac-20384919?page=0amp;citems=10.

6. “Diabetes.” Mayo Clinic, Mayo Foundation for Medical Education and Research, 8 Aug. 2018, www.mayoclinic.org/diseases-conditions/diabetes/diagnosis-treatment/drc-20371451.

7. Hall, Margaret Jean, et al. “Trends in Inpatient Hospital Deaths: National Hospital Discharge Survey, 2000–2010.” Centers for Disease Control and Prevention, Centers for Disease Control and Prevention, 24 May 2017, www.cdc.gov/nchs/products/databriefs/db118.htm.

8. Hill, Brian, Robert Brown, Eilon Gabel, Christine Lee, Maxime Cannesson, Loes Olde Loohuis, Ruth Johnson, et al. 2018. “Preoperative Predictions of in-Hospital Mortality Using Electronic Medical Record Data.” doi:10.1101/329813. “Hypoxemia (Low Blood Oxygen).” Mayo Clinic, Mayo Foundation for Medical Education and Research, 1 Dec. 2018, www.mayoclinic.org/symptoms/hypoxemia/basics/definition/sym-20050930.

9. “Low White Blood Cell Count Causes.” Mayo Clinic, Mayo Foundation for Medical Education and Research, 30 Nov. 2018, www.mayoclinic.org/symptoms/low-white-blood-cell-count/basics/causes/sym-20050615

10. MIMIC-III, a freely accessible critical care database. Johnson AEW, Pollard TJ, Shen L, Lehman L, Feng M, Ghassemi M, Moody B, Szolovits P, Celi LA, and Mark RG. Scientific Data (2016).Sissons, Claire. “Low Creatinine Levels: Causes, Symptoms, and Treatments.” Medical News Today, MediLexicon International, 2 Nov. 2017, www.medicalnewstoday.com/articles/319892.

11. Nabili, Siamak N. “Complete Blood Count (CBC) Normal Ranges, Test Results, Chart.” MedicineNet, MedicineNet, 22 Nov. 2019, www.medicinenet.com/complete_blood_count/ article.html.

12. “Normal Calcium Levels.” UCLA Endocrine Center, UCLA Health, www.uclahealth.org/endocrine-center/normal-calcium-levels.

13. “Prothrombin Time Test.” Mayo Clinic, Mayo Foundation for Medical Education and Research, 6 Nov. 2018, www.mayoclinic.org/tests-procedures/prothrombin-time/about/pac-20384661?page=0citems=10.

14. “Pulse &amp; Heart Rate.” Cleveland Clinic, Cleveland Clinic, 18 Feb. 2018, my.clevelandclinic.org/health/diagnostics/17402-pulse–heart-rate.

15. Qi, Yanjun. “Random Forest for Bioinformatics.” SpringerLink, Springer, Boston, MA, 1 Jan. 1970, link.springer.com/chapter/10.1007/978-1-4419-9326-7_11.

16. “Understanding Blood Pressure Readings.” Www.heart.org, American Heart Association, 30 Nov. 2017, www.heart.org/en/health-topics/high-blood-pressure/understanding-blood-pressure-readings.