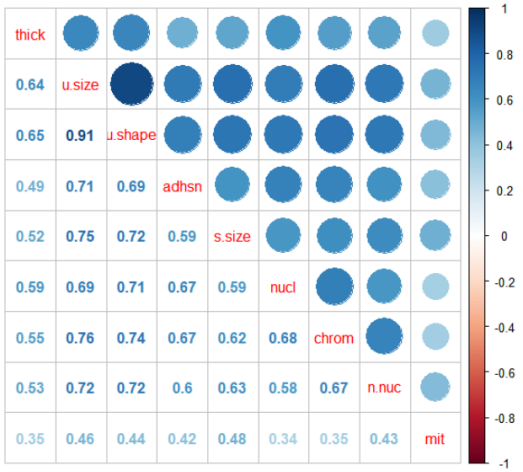
Data – The data set chosen contains data on 699 breast tumor biopsies collected from Dr. William H. Wolberg from the University of Wisconsin Hospitals. The data set can be accessed from the MASS library within the r statistical package version 3.5.1[1].

Tumors are either benign or malignant based on 9 independent variables as follows:

Thickness, size, shape, adhesion, epithelial cell size, nuclei, chromatin, nucleoli and mitosis

Each of the nine variables is an integer ranked from 1 to 10 in order of severity thus giving each variable a quantitative value that can be applied to a logistic regression function to determine a binary outcome of either benign (0) or malignant (1). Within the data set, 444 observations resulted in a benign outcome while 239 observations resulted in a malignant outcome.

To determine if there was a risk of multicollinearity between the independent variables a simple correlation matrix was used in conjunction with the variance inflation factor to determine if the level of collinearity was mild enough to move forward with the regression. The variance inflation factor compares the accuracy of a model with multiple variables to a model with only a single variable to determine the level of variability between the independent variables in the data set.



Variable	VIF
Thickness	1.2352
Size	3.2488
Shape	2.8304
Adhesion	1.3022
Cell size	1.6357
Nuclei	1.3729
Chromatin	1.5235
Nucleoli	1.3431
Mitosis	1.0597

The maximum VIF value was calculated to be 3.2488 which is under the suggested upper limit of 5.

Problem - Although a medical diagnosis tends to be quite accurate especially after a second opinion, under today's processes there is always bound to be some degree of human error in making such a decision. If data can accurately be collected with minimal to no human bias, determining a binary outcome such as a tumor being benign or malignant can be made more accurate through a mathematically enhanced approach such as logistic regression. Being able to systematically calculate an outcome using the relationship between many independent variables is a job best suited for statistical learning since human error is reduced in the equation.

Techniques – Since the predictor variables were all quantitative and the dependent variable was binary, it was appropriate to apply logistic regression to the biopsy data set. Logistic regression models use several

[1] P. M. Murphy and D. W. Aha (1992). UCI Repository of machine learning databases. [Machinereadable data repository]. Irvine, CA: University of California, Department of Information and Computer Science

independent variables to calculate the probability of a binary outcome. In this case, the independent variables listed above were all used to calculate the binary outcome of a tumor being either benign or malignant.

Since the data also contained the outcome or dependent variable for each observation it allowed me to split the data set into train and test sets that were then used to train and test the model's accuracy. I used 70% of the original data set to train the model and the remaining 30% of the data to test the model's accuracy. After cleaning the data and removing missing entries, which only accounted for 2.2% of the total data set, the model was now ready to be created using 683 observations. The glm function in r was used to create our binomial logistic regression model.

Results – After building the model using the train sample data set I applied my model to the remaining test data to determine the model's accuracy in predicting whether a patient's medical examination suggested they have either a benign or malignant tumor.

After reducing the test data set from 683 observations down to 209 observations the results were as follows. 139 benign tumors were accurately predicted to be benign tumors and 65 malignant tumors were accurately predicted to be malignant tumors based on the 9 independent variables describing each observation. That means that there were only 5/209 observations that resulted in an incorrect categorization of ither benign or malignant by the model leaving us a prediction accuracy 97.6%

Simply put, if a patient gets a pre-existing breast tumor medically examined the probability that this model will accurately determine whether the tumor is benign or malignant is 97.6%.

Conclusion - After applying logistic regression using the glm function in r on the biopsy data set the model was able to predict outcomes to an accuracy of 97.6%. Although this is accurate it means that the model gives incorrect outcomes 2.4% of the time and unlike getting a second opinion from another doctor the model will always produce the same output if it is given the same data. Therefore, efforts should be made using other statistical processes to increase the accuracy of the model. For instance, it is possible that the model is over fit and as a result produces incorrect output. Since the data set includes the dependent variable along with the independent variables a technique that could be used to further increase the accuracy of the model and reduce overfitting would be cross-validation. This would split the data into more than one train and test subsets and would average the results with the aim at increasing accuracy.

In addition to producing a model with a high level of accuracy for differentiating between benign and malignant tumors, this approach also suggests that the traditional process of diagnosing patients could benefit from the introduction of regression to mathematically keep track of the relationship between independent variables and their association with the dependent variable. Although data on the average doctor's ability to determine whether a tumor is malignant or benign is not readily available, a process such as this that is based heavily on numerical measurements is sure to benefit from the application of logistic regression.