“I hereby declare that the work represented in this report represents my individual effort. I

understand that I am encouraged to seek advice and guidance on any course material from my

professor, teaching assistants and fellow classmates, however, I am responsible, solely, for this

written document and my oral presentation of components of this report.”

\_\_\_\_\_\_\_\_\_Sean Vieau\_\_\_\_\_\_\_\_\_\_ \_\_\_\_\_\_\_10/06/2024\_\_\_\_\_\_\_

(name) (date)

**Project 3**

**Introduction**

The aim of the current study is to investigate whether MRI image features extracted from baseline kidney images can enhance the prediction of disease progression in young Autosomal Dominant Polycystic Kidney Disease (ADPKD) patients. This study is important because recent literature has demonstrated the inclusion of MRI image features improves the prognostic accuracy of models in adults, but this has not yet been demonstrated in children. Improving prediction models could greatly assist in diagnosing and predicting kidney disease outcomes in children, which is more challenging than in adults and could thereby lead to improved treatment.

The two clinical hypotheses for this study are that including MRI image features will improve model performance compared to using baseline kidney volume alone in models predicting 1) percentage change of total kidney volume growth and 2) classification of a patient as having fast or slow progression of the disease. Data was received as a .csv file and was pre-processed and cleaned by the investigator.

**Method**

**Study Design**

This project recruited 71 young patients with ADPKD and collected MRI data at baseline and after 3 years. Additionally, the height corrected kidney volume for each patient was collected at baseline and 3 years by a physician, and the percentage change calculated. Patients were classified as having slow or fast progression of the disease based on this percentage change. Image features were extracted from the baseline MRI images including 2 image features on kidney geometric information, 5 features based on Gabor transform, 2 features based on gray level co-occurrence matrix, 5 features based on image textures, and 5 features based on local binary pattern.

**Statistical Hypotheses**

1. A linear regression model predicting percentage change of total kidney volume growth including MRI image features and baseline kidney volume will have better performance than a model with baseline kidney volume alone, as determined by specificity, sensitivity, PPV, NPV, accuracy, and AUC.
2. A logistic regression model predicting classification of disease progression as slow or fast including MRI image features and baseline kidney volume will have better performance than a model with baseline kidney volume alone, as determined by specificity, sensitivity, PPV, NPV, accuracy, and AUC.

**Data Management**

All data analysis and management was performed in R version 4.4.1. Data was pre-cleaned by the Dr. Fuyong Xing.

*Outliers****.*** Boxplots and histograms of all MRI image features and kidney volumes were made to assess for potential outliers. Outliers were then identified using jackknife residuals. Data points +/- 3 SD of the mean with high leverage and influence were considered outliers and removed from the data set as determined by Cook’s D, DFITS, DFBETAS, and hat-values.

*Missingness.* There were no missing values in this data set, and thus missing data mechanisms did not need to be assessed.

*Feature Scaling.* 5-fold cross validation was performed, and z-score normalization executed on the features of each training set. The mean and standard deviation of each training set was then used to normalize the respective test set.

*Feature Engineering and Selection.* Scatterplots were made to assess the relationship of each MRI image feature with the percentage change in total kidney volume. Features were then squared for plots that appeared quadratic to better capture non-linear relationships.

*Feature Selection.*  A correlation matrix was performed to assess which MRI image features were most strongly correlated with percentage change in total kidney volume and slow or fast disease progression. Individual univariate analyses (linear regression for total kidney volume, logistic regression for slow or fast disease progression) were then conducted to assess if each MRI image feature was a significant predictor of the target variables. To keep with the “Rule of 10”, the top 5-6 most promising features were chosen for the final model.

**Data Analysis**

To answer the first research question, a linear regression was performed on the percentage change in total kidney volume over 3 years with the 6 features chosen during initial screening using 5-fold cross validation.

To answer the second research question, a logistic regression was performed predicting classification of slow or fast disease progression with the 6 features chosen during initial screening using 5-fold cross validation.

Several plots were made to ensure the data met the assumptions for a linear regression. Q-Q plots and histograms of the jackknife residuals were created to assess normality. A scatterplot of the jackknife residuals was made to ensure errors were centered around zero. Homoscedasticity was assessed by plotting the jackknife residuals for each treatment group,, along with Bartlett’s test of Homogeneity of Variances.