Group B: TJ Gardner-Souza, Amanda Huang, Luc Ravenelle, Erik Wilder, and Muskan Yadav
Professor Ding
MATH 7343
April 25, 2022

A Statistical Analysis of Heart Disease & Stroke

Objective

Heart Disease and Stroke are some of the leading causes of death in America. We investigated an open-source stroke dataset from Kaggle that contains information on various patients' health metrics and whether or not they suffered a stroke. The objective was to find variables that were correlated with a patient who had suffered a stroke, and to determine if an accurate classification could be built to predict whether a new patient would be likely to have a stroke or not.

Study Design

Target Population: Global human population (all ages, with no distinct region specified)
Sampled Population: Those patients with data recorded (unclear, source of data is confidential)
Sample: 5,111 patients
Parameter in question: Proportion of patients who have experienced stroke

The dataset chosen for analysis involves patients from an unknown source, regardless the dataset is intended for a global population. People of all ages with various pre-existing health conditions and statuses have been recorded in this dataset with a final, exhaustive parameter indicating their stroke diagnosis. According to the World Health Organization (WHO) stroke is the second largest cause of death among people worldwide, with its share standing at about eleven percent of deaths each year. This dataset's purpose and our goal is to find other health status input parameters that may be strong predictors of stroke, and make a prediction for a patient.

The dataset hosts a sample of 5,111 patients of unknown origin or location. Although the sample is large, we must call attention to the fact that randomness cannot be guaranteed, because the methods of collection and source are unknown. It must also be stated that this study is observational in nature, as the parameters and their respective statistics cannot be collected in a proper or controlled means. Therefore, the outcome of this analysis will only show association and not make any causal conclusions. All variables that may or may not show statistical significance in our analysis will infer an association or non-association between themselves and stroke.

Data Analysis

The dataset was imported from the Kaggle repository. It contains 10 predictor variables (7 categorical and 3 numerical) and a binary variable indicating whether or not the patient had a stroke. To prepare the data, null values in the predictor variables were replaced so that R could understand the data set. We then normalized the numerical data such as glucose, BMI, and age. The categorical variables include gender, hypertension, heart disease, work type, urban versus rural residence, smoking status, and stroke history. For regression formatting, we created 'dummy' variables from the categorical variables and up-sampled the minority class.

We used contingency tables to determine the frequency of an individual categorical variable against stroke; specifically, stroke versus hypertension, stroke versus heart disease, and stroke versus smoking status. Results of the contingency tests can be pictured below:

```
            stroke
hypertension    0    1
           0 4309  149
           1  391   60
            stroke
hypertension          0          1
           0 0.96657694 0.03342306
           1 0.86696231 0.13303769

        Pearson's Chi-squared test

data:  cont_table_hyp
X-squared = 99.704, df = 1, p-value < 2.2e-16
```

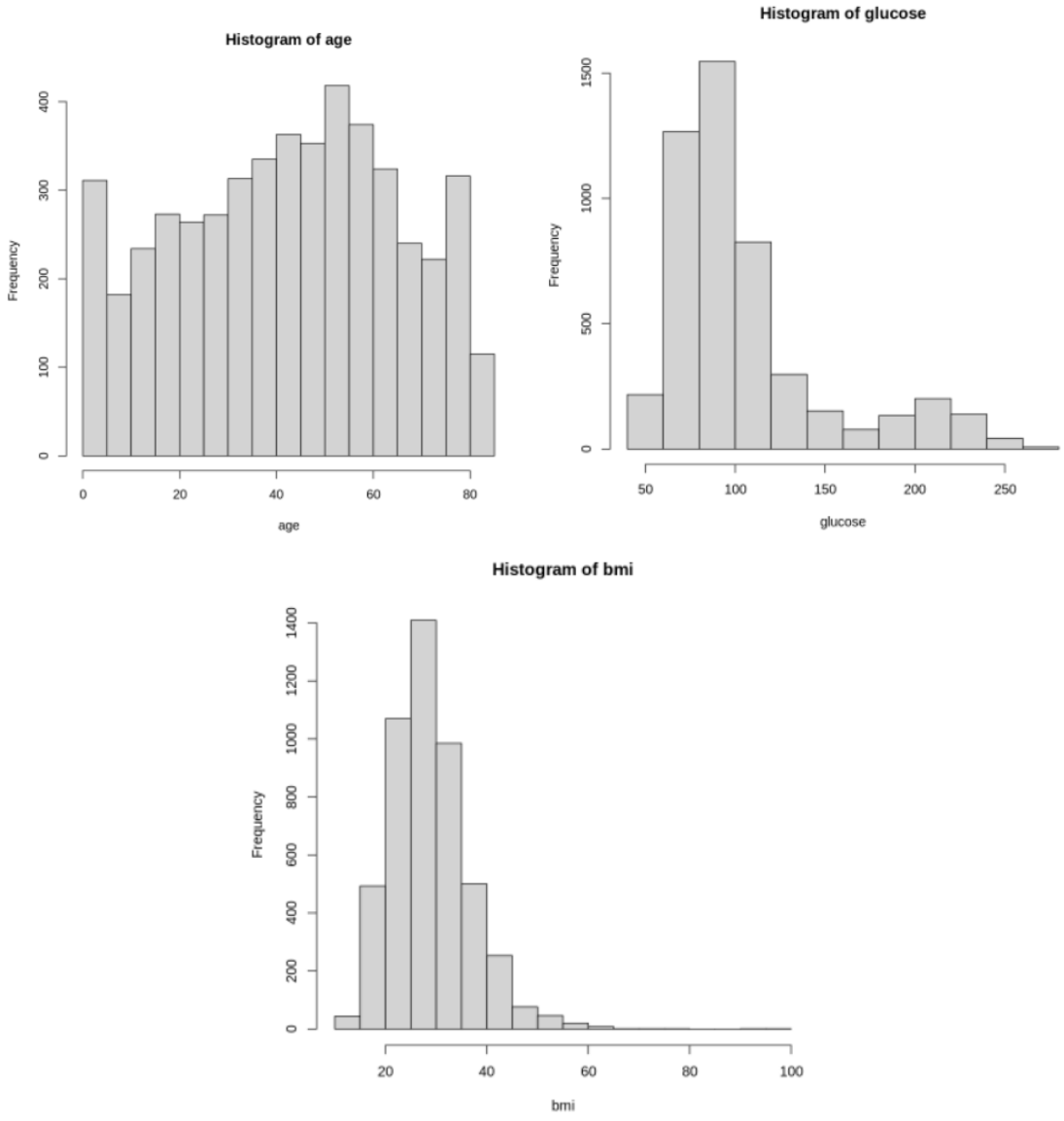
```
             stroke
heart_disease    0    1
           0 4497  169
           1  203   40
            stroke
heart_disease          0          1
           0 0.96378054 0.03621946
           1 0.83539095 0.16460905

        Pearson's Chi-squared test

data:  cont_table_hd
X-squared = 93.403, df = 1, p-value < 2.2e-16
```

```
                        stroke
smoking                    0    1
  formerly smoked        780   57
  never smoked          1768   84
  smokes                 698   39
  Unknown               1454   29
                        stroke
smoking                      0          1
  formerly smoked   0.93189964 0.06810036
  never smoked      0.95464363 0.04535637
  smokes            0.94708277 0.05291723
  Unknown           0.98044504 0.01955496

        Pearson's Chi-squared test

data:  cont_table_smoke
X-squared = 34.943, df = 3, p-value = 1.252e-07
```

For each of the Chi-square tests performed above, the null hypothesis is that there is no relationship between the predictor variables (hypertension, heart disease, and smoking status) and the target variable (stroke). If this were the case, we would see the same distribution of stroke vs. non-stroke patients across the different categories of the independent variables. However, in this case, we see that the p-value for each of the tests performed is less than the level of significance, $\alpha = 0.05$. Therefore, we can reject the null hypothesis that no relationship

exists between each of the predictor variables and the target variable. Thus, we conclude that there is sufficient evidence to suggest that hypertension, heart disease, and smoking status are all correlated with stroke.

Before performing tests on the numerical variables, histograms were implemented to categorize the type of distribution and to determine the correct type of test that should be executed. The outputs of the histograms received for age, glucose and BMI are pictured below:



Normally distributed data looks similar to a bell shaped curve with one peak and symmetric around the mean. Based on the histograms above, age, glucose and BMI are not normally distributed. Thus, warranting the Wilcoxon rank-sum test.

The Wilcoxon rank-sum is a nonparametric test for unpaired data that can be used to compare two samples from independent populations and does not require for the underlying populations to be normally distributed or that their variances be equal. For each of the Wilcoxon tests – comparing stroke to numeric data: age, average glucose level, and BMI– the results return p-values that are significantly less than the level of significance, $\alpha = 0.05$. With the p-values less than the level of significance for all three tests, we concluded that we reject our null hypothesis for these tests and conclude that there is evidence that the medians of the two tested populations are not equal. The outputs of the Wilcoxon tests are pictured below:

```
        Wilcoxon rank sum test with continuity correction

data:  stroke$age and nonstroke$age
W = 821280, p-value < 2.2e-16
alternative hypothesis: true location shift is not equal to 0

        Wilcoxon rank sum test with continuity correction

data:  stroke$avg_glucose_level and nonstroke$avg_glucose_level
W = 614278, p-value = 8.168e-10
alternative hypothesis: true location shift is not equal to 0

        Wilcoxon rank sum test with continuity correction

data:  stroke$bmi and nonstroke$bmi
W = 569022, p-value = 0.0001026
alternative hypothesis: true location shift is not equal to 0
```

Results

To determine if it is possible to build an accurate classifier for identifying a new patient as “at-risk” for having a stroke, we have performed a series of logistic regressions. Our first attempt used all the variables and split the data into 75% training and 25% testing data. Numerical data were normalized and categorical data were placed into ‘dummy’ columns to achieve the best results. The ‘Formerly Smoked’ Category had only the ‘0’ class, so this variable was omitted from all regression attempts. The prediction result of this model is pictured below:

stroke.pred	0	1
0	1175	53

Due to the imbalance of stroke patients and non-stroke patients, the model did not make any ‘stroke’ predictions. To troubleshoot this, we used an over-sampling technique to increase the

size of the minority class. Another logistic regression was performed with the same input variables and a balanced target class. The results are as given below:

```
Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.7603  -0.7749  -0.2240   0.8107   2.6343

Coefficients: (4 not defined because of singularities)
                Estimate Std. Error z value Pr(>|z|)
(Intercept)    -12.01092  198.36696  -0.061 0.951718
male1           -0.01136    0.08587  -0.132 0.894740
female1               NA         NA      NA       NA
age              1.36938    0.06518  21.010  < 2e-16 ***
hypertension1    0.77725    0.11602   6.699 2.10e-11 ***
heart_disease1   0.74617    0.14513   5.141 2.73e-07 ***
married1         0.21629    0.13288   1.628 0.103591
unmarried1            NA         NA      NA       NA
private1        10.58041  198.36696   0.053 0.957463
govt1           10.27298  198.36699   0.052 0.958698
self1           10.61543  198.36698   0.054 0.957322
children1       11.36336  198.36714   0.057 0.954319
never_worked1         NA         NA      NA       NA
urban1           0.13695    0.08214   1.667 0.095467 .
rural1                NA         NA      NA       NA
glucose          0.13422    0.03241   4.141 3.46e-05 ***
bmi              0.16358    0.04639   3.526 0.000421 ***
smokes1          0.34374    0.12596   2.729 0.006354 **
never_smoked1   -0.13400    0.10622  -1.262 0.207118
unknown_smoker1 -0.07198    0.12930  -0.557 0.577760
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 5100.1  on 3680  degrees of freedom
Residual deviance: 3661.7  on 3665  degrees of freedom
AIC: 3693.7

Number of Fisher Scoring iterations: 12
```

This logistic regression identifies six variables to be correlated with stroke. Age, hypertension, heart disease, glucose, BMI, and smoking status. Since all had p-values less than the level of significance, $\alpha = 0.05$. These results are in line with our previous Chi-square and Wilcoxon tests, so we are confident there exists a correlation between these six variables and stroke. Classification results are as shown below:

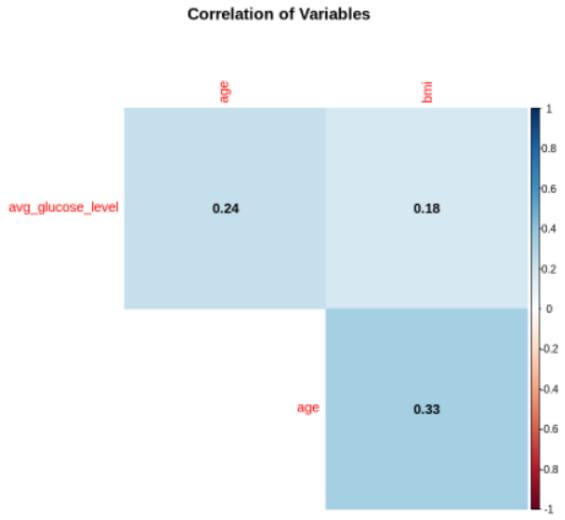
```
                          stroke.pred   0   1
                                    0 989  18
                                    1 186  35

Call:
accuracy.meas(response = test$stroke, predicted = balanced_result)

Examples are labelled as positive when predicted is greater than 0.5

precision: 0.158
recall: 0.660
F: 0.128
```

This model has a precision of 0.158, which can be interpreted as that around 85% of positive classifications were incorrect. The recall is 0.660, meaning that the model classified 66% of true positives correctly. The F-score is an overall accuracy measure for the model, which in our case is low, suggesting that this model does not perform well overall.

As a measure to improve the performance, we omitted variables that were not found to be significant. Results from this model are as pictured below:

```
Coefficients:
               Estimate Std. Error z value Pr(>|z|)
(Intercept)    -1.27066    0.06496 -19.561  < 2e-16 ***
age             1.36228    0.05748  23.701  < 2e-16 ***
hypertension1   0.77463    0.11273   6.872 6.35e-12 ***
heart_disease1  0.74139    0.14348   5.167 2.38e-07 ***
glucose         0.13898    0.03195   4.349 1.37e-05 ***
bmi             0.14997    0.04503   3.331 0.000866 ***
smokes1         0.41448    0.10549   3.929 8.53e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 5100.1  on 3680  degrees of freedom
Residual deviance: 3681.6  on 3674  degrees of freedom
AIC: 3695.6

Number of Fisher Scoring iterations: 5

stroke.pred   0   1
          0 994  16
          1 181  37

Call:
accuracy.meas(response = test$stroke, predicted = balanced_result_refined)

Examples are labelled as positive when predicted is greater than 0.5

precision: 0.170
recall: 0.698
F: 0.137
```

All variables have a p-value less than the level of significance, $\alpha = 0.05$, and hence all of them are correlated with stroke, as was expected. We observe slight improvements in precision, accuracy, and recall as well.



Noticeably, none of the numerical variables show a high correlation with each other, so we do not need to remove any of these from the model. Additionally, to test correlation among the categorical variables, we performed a series of chi-square tests among them:

```
             heart_disease
hypertension    0    1
           0 4273  185
           1  393   58
             heart_disease
hypertension          0          1
           0 0.95850157 0.04149843
           1 0.87139690 0.12860310

        Pearson's Chi-squared test

data:  cont_table_hyp
X-squared = 66.045, df = 1, p-value = 4.407e-16
```

```
             smoking
heart_disease formerly smoked never smoked smokes Unknown
            0             767         1771    682    1446
            1              70           81     55      37
             smoking
heart_disease formerly smoked never smoked    smokes   Unknown
            0       0.1643806    0.3795542 0.1461637 0.3099014
            1       0.2880658    0.3333333 0.2263374 0.1522634

        Pearson's Chi-squared test

data:  cont_table_hd
X-squared = 50.919, df = 3, p-value = 5.09e-11


                    hypertension
smoking                0    1
  formerly smoked  727  110
  never smoked    1636  216
  smokes           655   82
  Unknown         1440   43
                    hypertension
smoking                      0          1
  formerly smoked 0.86857826 0.13142174
  never smoked    0.88336933 0.11663067
  smokes          0.88873813 0.11126187
  Unknown         0.97100472 0.02899528

        Pearson's Chi-squared test

data:  cont_table_smoke
X-squared = 102.89, df = 3, p-value < 2.2e-16
```

We consider that the null hypothesis for each of these tests is that the variables are not correlated. Since each test has a p-value less than the level of significance, $\alpha = 0.05$, we reject the assumed null hypothesis that these variables are not correlated. Hence there is sufficient evidence to suggest that a correlation exists amongst all of the categorical predictor variables. For our final regression model, we kept all numerical variables but only selected the most significant categorical predictor value, hypertension. The results from the final model are as presented below:

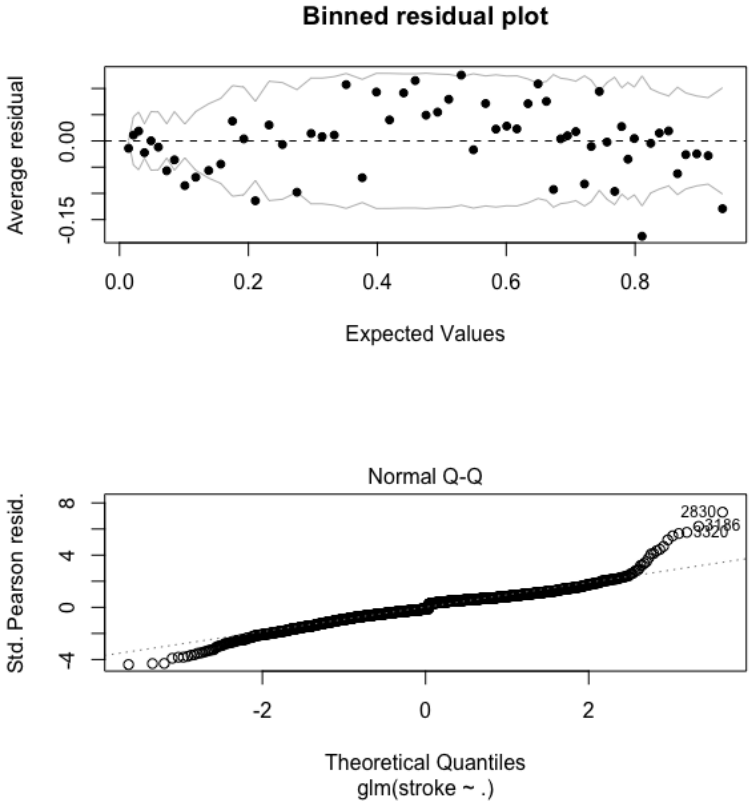
```
Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept)   -1.14407    0.05960 -19.196  < 2e-16 ***
age            1.39567    0.05596  24.940  < 2e-16 ***
hypertension1  0.78408    0.11225   6.985 2.84e-12 ***
glucose        0.16784    0.03137   5.350 8.80e-08 ***
bmi            0.15564    0.04483   3.472 0.000517 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 5100.1  on 3680  degrees of freedom
Residual deviance: 3736.7  on 3676  degrees of freedom
AIC: 3746.7

Number of Fisher Scoring iterations: 5

stroke.pred   0   1
          0 987  16
          1 188  37

Call:
accuracy.meas(response = test$stroke, predicted = balanced_result_correlated)

Examples are labelled as positive when predicted is greater than 0.5

precision: 0.164
recall: 0.698
F: 0.133
```

Removing correlated categorical features did not result in a significant improvement. However since this model is seen to be the least complex, we select this to be the best model. It was noted that 16.4% of positive classifications were correct, and 70% of true positives were identified correctly. The F-score of 0.133 suggests that this model does not generally perform well.

To add to the analysis of the model's performance, we can plot the residuals and see how they compare to our fitted values. R can give several plot options through which we can examine the errors of our model. When using logistic regression, using a residual vs fitted plot is not very useful. Because of the discrete nature of the data, we cannot gather significant information from them. Instead, we have chosen to include a binned plot that places residuals into bins based on their fitted values. The first image below is plotting the average residual vs. the average fitted value for each bin. The gray lines represent the confidence limits in which 95% of all values should fit into, depending on the accuracy of the model. Our model does not seem to stick to that requirement, which further suggests a poor efficacy.

We can also observe from the binned plot that our model does well with positive residuals, but not negative ones. Since all of the outliers in this plot are negative. We know that in this plot the residual is derived from the observed minus the fitted value. Therefore, in these cases, not only is the observed less than the fitted but the difference is drastic enough for it to

land outside of the 95% interval. This indicates that our model is over predicting stroke which can be confirmed by our data referenced earlier. The precision of the model was 0.164 which confirms this indication, as 83% of predicted strokes were incorrect.



Examining residuals is an effective way to identify incorrect inferences in data and possible peculiarities that could lead us there. One such peculiarity shown in our residual plots is that both the plots show several data points that are classified as outliers. These points are shown to be outside of the expected range of possible values and could be impacting our model's efficacy. In this case, a naive statistician would simply remove them in order to improve the prediction of the model. However, this removal needs to be justified. Upon examining our outliers, we concluded that removal of these outlier data points would be inappropriate and would compromise the legitimacy of the model. Let's consider the data point #1341, for instance. This patient has a fairly high average glucose level of 178.76 with a below average BMI of 24.1. The patient registers as an outlier in our residual plot, but the data is conceivable. The patient is a middle aged male, with the preexisting condition of heart disease and there is no justifiable reason to remove him from the data set. After examining a few of these outliers for possible removal, we decided that it is best to not do so.

Discussion

The goal of this project was to identify variables that were associated with stroke and investigate if it is possible to build an accurate classifier to categorize a patient as at-risk for having a stroke. The Chi-square tests and Wilcoxon tests showed that age, heart disease, hypertension, smoking status, BMI, and average glucose level are all significant indicators. However, we weren't able to build an accurate classifier to prove the same.

To improve this model there are a number of things that could be explored if we were to take this further. One improvement we could make would be to try to refine the regression further. Our methodology was to take an initial regression baseline and then refine it by removing all variables that did not show statistical significance in our other exploratory tests (Wilcoxon and Chi-Squared). Another method we could have used would be to follow through the entire backwards stepwise procedure and examine each variable and its effect on the results, one at a time. We also could have explored our remaining variables in the refined regression models for interaction. If the variables were interacting and impacting the result of our prediction, there would be no way of knowing without stepping through each permutation of variable interaction combinations. If interaction were to be found, we could remove either variable to see if removal helped our results. This method was not pursued in lieu of our time constraint, but could be added in future iterations.

Finally, there are times in which the classifier can only perform so well given the data and methodology. When the model will not perform at the desired level it can sometimes be a good idea to revisit the available data and perhaps rethink the model. If precision is a critical outcome of this study a different prediction model could be suggested, instead of logistic regression. Or, it could be that the features collected are not substantial or indicative of the overall goal and a new study could be conducted where the data collection is refocused on other variables which could produce a better result.

Work Cited

Fedesoriano. (Updated a year ago, 2021). Stroke Prediction Dataset, Version 1. Retrieved March 2022 from https://www.kaggle.com/datasets/fedesoriano/stroke-prediction-dataset.

Group Member Contributions

TJ Gardner-Souza: Contributed to the logistic regression and wilcoxon tests coding and write-up

Amanda Huang: Contributed by setting up Zoom meetings and booking rooms for meet ups. For the project portion: formulated the contingency tables, wrote the data analysis portion of the paper, and presented the background, data set preparation, and data analysis slides, formatted slides, and edited paper for grammar issues.

Luc Ravenelle: Analysis contributions include regression, contingency tables as well as data formatting. Focused on the content of data and how that would translate to our study design, residual interpretations and how to improve the study going further.

Erik Wilder: Contributed by assisting in cleaning data and creation of correlation table, assisted with paper, assisted with presentation.

Muskan Yadav: Contributed by analyzing the correlation tables and interpreting the results, editing the paper and putting the slides together.