Parsa Moheban
Mattathias Klassen
Michelle Kojekine

Using a Machine Learning Model to Predict the Presence of Metastatic Tumors

Abstract

Metastasis describes the ability of a tumor cell to dislocate to other regions of a body, and start new growth colonies [4]. The 5-year survival rate of metastatic patients with Kidney Renal Clear Cell Carcinoma (KIRC) is less than 10%, and unlike other renal cell carcinoma types, KIRC has a much higher recurrence and metastasis rate [7]. A Kaplan-Meier estimator is used in plotting the survival rates of patients with metastatic tumors and those without to see how metastasis affects patient survivability directly. Our investigation involves using data provided by the TCGA program regarding KIRC, obtaining data of gene counts per patient, performing PCA, and logistic regression in order to determine whether metastatic data can be governed solely from gene levels. Our machine learning model is ~95% successful in its predictions, and can ideally be applied in determining metastatic rates in other samples, with the eventual goal of predicting metastasis before it occurs. A second model is also applied to the regional lymph nodes in order to better understand how gene expression of the patients can affect cancer data. Here, a much lower accuracy due to smaller sample data is seen.

Introduction

Cancer originating in the kidneys are some of the most commonly diagnosed tumors among humans [1]. Also referred to as renal cell carcinoma, it includes a variety of subtypes of cancer which originate at different sites within the urinary tubules (excluding renal interstitial tumors and renal pelvis tumors) [2]. Within these subtypes, by far the most common cancer is Kidney Renal Clear Cell Carcinoma (KIRC) - making up roughly 80% of all renal cell carcinoma cases [1]. Various clinical methods exist in preventing the ongoing spread of the cancer (mostly surgical means), but even so, roughly 60% of patients diagnosed with KIRC die within 3 years of diagnosis [2]; furthermore, less than 5% of patients with KIRC survive past 5 years [1].

The greatest contributor to death from cancer is the presence of metastatic tumors [3]. At a base level, metastasis describes the ability of tumor cells to break away from the origin site and dislocate to other regions in the body - there a new series of tumor cells can grow [4]. A tumor cell can begin metastasis via an invasion of the surrounding tissue of a primary tumor [3]; from there, the cells can enter the bloodstream, and traverse through it to find a new region to start growth [3]. After growth is initiated, the tumor cells can alter the cellular composition, immune status, supply of blood, and extracellular matrix in order to create a more ideal condition for tumorous growth [3]. Moreover, metastatic tumors are the way in which a cancer can dangerously spread. In the case of KIRC, this can lead to further reductions in survivability for a patient (will be seen in a later plot).

Although there are many ongoing investigations into treatment methods - such as new targeted therapy with promising results in patient survival rates [2], and new surgical methodologies - our investigation acts as a prelude to other research.

Where other researchers have found specific genes that impact the probability of KIRC [1], our research focuses on overall gene expression leading to an either metastatic or non-metastatic tumor.
Our goal is to design a machine learning algorithm that can interpret the gene expression of specific patients and predict whether or not the patient has a metastatic tumor with use of the TCGA KIRC clinical data, and RNA sequence data. We will then perform a secondary analysis on predicting regional lymph nodes with the same

algorithm to test its overall accuracy. Our overall research question is then whether or not we can predict the metastasis of a tumor via an RNA sequence alone.

The importance of this research is that if successful, solely based on a RNA sequence, we can predict how a tumor may develop and spread via the use of bioinformatic tools and algorithms. Ideally, this could lower patient death rates as cancer spread could be caught before developing to a deadly stage.

Methods

There are two main files sourced from The Cancer Genome Atlas pertaining to KIRC. The first file is a plethora of information that holds all of the data sourced from the patients. The second file is each patient's RNA sequence sampled from their tumor sites. The most important first step was reformatting the files to give consistent information when cross references between the two were made. We began by clearing both files of any duplicate, missing, and inconsistent entries between the files. This was immediately superseded by a transposition of the RNA sequence file in order to have all of the patients in the rows to match the clinical data file. This left both files having the exact same number of patient entries for a simple comparison when the data would be analyzed. The final step in preparing the data was to remove any columns in the transposed RNA matrix that had a variance of 0. Leaving any of these rows in would have created errors in the PCA and further analyses.

Next, a survival analysis was conducted using the patient IDs, metastasis status, OS status, and OS life span. Using these factors, we can attribute patients' survivability rates to see the correlation of metastatic tumors to a patients life span. Conducting this helps solidify previous works and provides a solid foundation for the project. It was an important first analysis to conduct as it is the leading principle in the hypothesis. Using a Kaplan-Meier plot provided significant insight from the survival analyses of the patients based on their metastatic stage.

Following the survival analysis, a Principal Component Analysis was conducted. This was done on the transposed data matrix to have every row labeled as a patient and all the columns to be the different principal components. Conducting the PCA in this manner allows for simple traceability back to the clinical data set. Using a simple for loop, the 90% most significant PCs were found, however, were much too large of a value to be used for further analysis. Instead, the top 20 were chosen as they attributed to the highest cumulative proportion of the variance as seen in figure 2. With these significant PC values now in hand, they were added to a matrix along with each patient’s M stage, N stage, and OS status.

After the creation of the final matrix, logistic regression was used in order to create a machine learning algorithm to predict the metastatic state of each patient's tumors. Since the data describing metastasis is binary, logistic regression was the optimal selection (when compared to linear regression). A training data-set and a test data-set were set-up in order to train the algorithm and eventually test its capabilities. Next, any metastatic value that had the label “MX” was removed. If these values had been kept, it would have caused high inconsistencies in the data, since a lack of data cannot be predicted by the algorithm. After the values were removed, glm was called on the training set and then plugged into a predict function to test the model’s effectiveness with the test set. With a P-value that was higher than 0.5, the tumor would be classified as metastatic, and non-metastatic otherwise. A confusion matrix was then used to check to see if the results acquired had some significance. A graphical representation of the confusion matrix was plotted for better representation.

The final piece of the code was another logistic regression model for local migration of the tumors using the N stage of the data. The above steps were conducted again for the N stage status of the patients.

Results

Performing a survival analysis allowed us to see the trend in survivability probability for patients suffering from metastatic tumors, and those without.

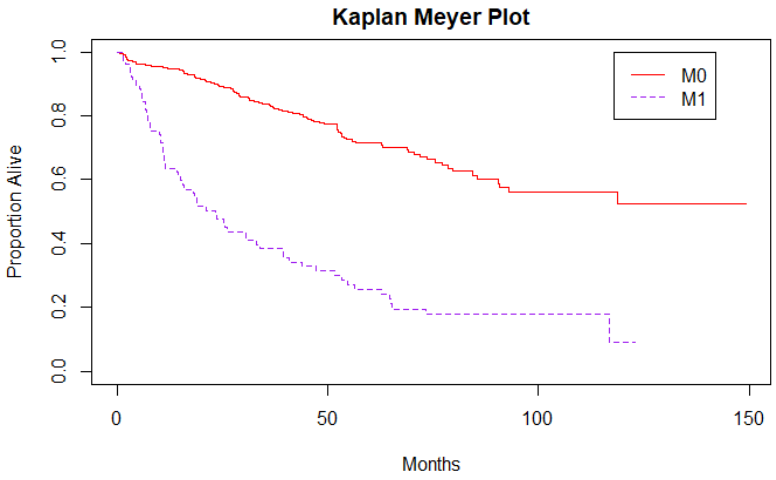

Fig. 1 Kaplan-Meier plot using estimator to showcase proportion of patients alive vs. survival time (in months) for patients with (M1) and without (M0) metastatic tumors.

The Kaplan–Meier plot seen in figure 1 solidified the understanding that in the presence of a metastatic tumor, a patient has a far lower probability of surviving compared to those with regionally static cancers. The finding further fuels the initial goal of determining the causes behind metastasis.

To continue the analysis on RNA seq data, we organized our data via transposing the RNA sequence. After removing data with negligible variance, and running PCA analysis via prcomp, the result was a list of dataframes regarding the standard deviation the function found, rotation, averages, and scaled values. The most important data was the “x” that encapsulated a matrix of the patients and the PCs associated with the genes. Taking the 90% significance of the PCs for dimensionality reduction resulted in far too many principal components being considered “significant” - in reality, of the 406 significant PCs, only the top 20 were carried forward for the model with the rest being discarded. As an aside, taking a larger number of principal components resulted in the model having strangely skewed results and being far too computationally straining.

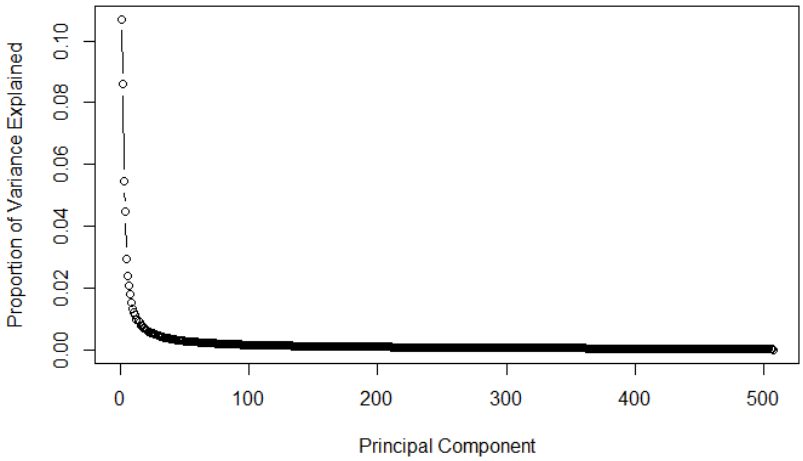

Fig. 2 Proportion of Variance Vs. each principal component shows that roughly the first twenty PCs have the highest variance.

Variance throughout the principal components was measured and plotted as seen in figure 2. The highest levels of variance can be seen in the first 20-30 principal components which further solidified our selection in only using the top 20 PCs for modeling. A cumulative proportion of variance plot was also made as seen in figure 3 indicating similar results. The highest cumulative proportion is attributed to the first 20-30 PCs.

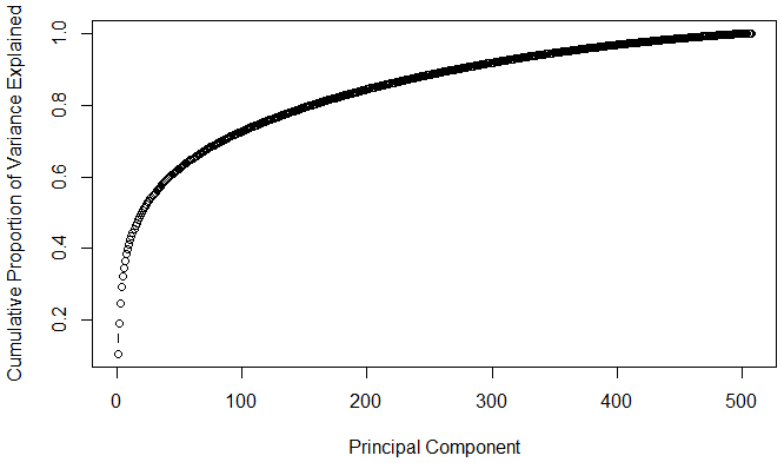

Fig. 3 Cumulative distribution of variance for each principal component. The highest proportion seen for the first ~30 PCs.

The third plot describes the principal components and compares how PCA transforms the data into a new space. The observations corresponding to the first two important PCs are graphed and can be seen in figure 4. The three levels in metastatic data are coloured and plotted respectively, with a 4th color representing where the holes in the data are, which are later addressed and removed.

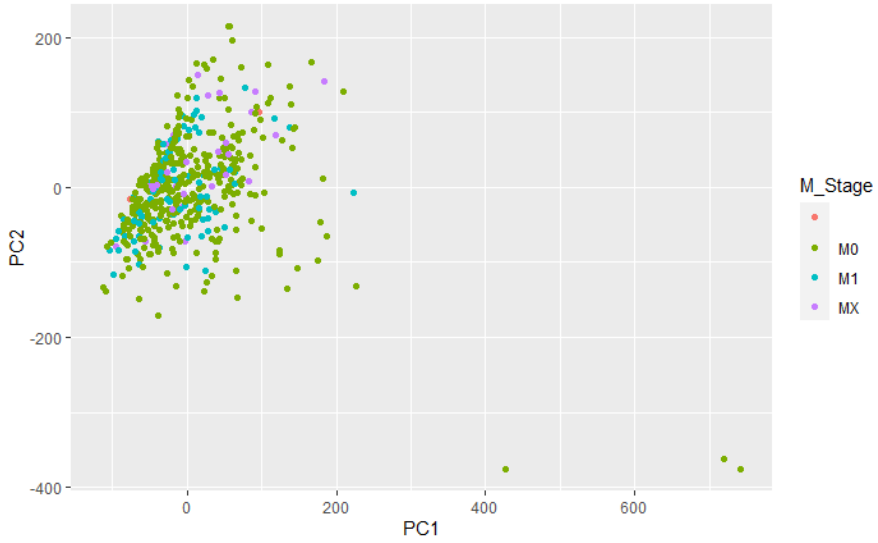

Fig. 4 Observations corresponding to the first two important PCs with respect to the different levels in metastatic data.

The seemingly random variability seen in the corners of the plot represent extraneous data and unfiltered extrema. Overall, the spread of the data is heavily skewed towards the upper bounds of PC2.

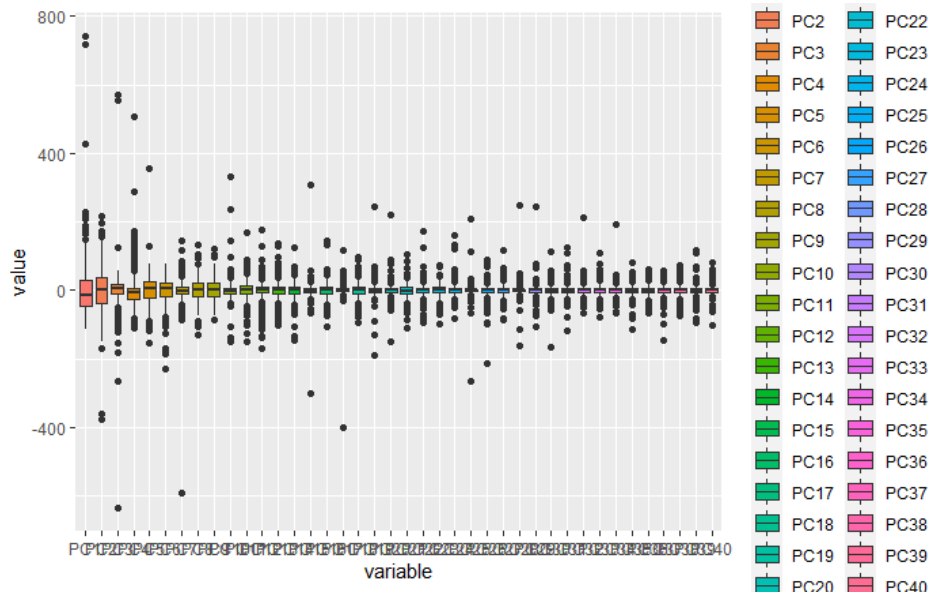

Fig. 5 Variability in the first 40 PCs represented as 40 individual box plots colored accordingly. First ~20 PCs have greatest value minimums and maximums, and highest interquartile range.

A fourth and final plot for PCA describes the variability in our first 40 PCs (double the number used for the model). The first 40 were considered in order to understand to what scale our data can be reduced. Since the greatest range was seen in the first twenty PCs, those were used for modeling.

After PCA, the training and test sets were set up, missing data was removed, and a generalized linear model (GLM) was used with the “binomial” argument to clarify that logistic regression is used - logistic regression is widely more applicable here since metastasis data is binary: either the tumor is metastatic (M1) or not (M0).

The prediction model that was used on the test set resulted in a confusion matrix with ~90-95% accuracy at identifying metastatic tumors and roughly 75% accuracy at predicting the presence of regional lymph nodes. The results of the confusion matrix, as well as the individual summaries for the logistic regression model can be seen in the appendix.

An AUC - ROC curve describes performance at various thresholds for our classifications [5]. ROC represents the probability curve itself, while AUC represents the degree of separability [5]. Generally, higher AUC values mean the model that's being plotted is better at distinguishing between positives and negatives in the data [5]. For cancer and genomics data, the high degree of accuracy is crucial for correct modeling. A high AUC is demonstrated by a sharper curve of the plot.
ROC curves for both metastasis and regional lymph nodes were plotted in figures 6 and 7 respectively.

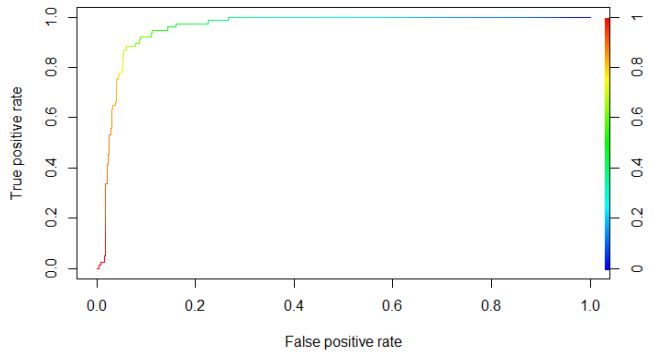

Fig. 6 ROC curve for metastatic tumor model with AUC ~ 0.95.

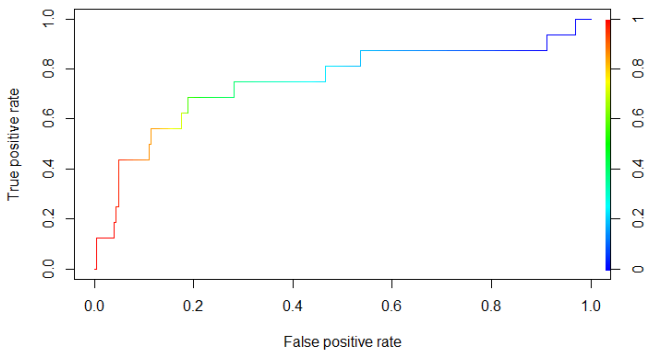

Fig. 7 ROC curve for regional lymph nodes model with AUC ~ 0.75.

The ROC curve seen in figure 6 for the metastatic tumor model describes the model as demonstrating high separability of the data due to the sharp high AUROC, with an AUC near 0.95. The model has high discrimination capacity to distinguish between positive and negative classes.

The ROC curve seen in figure 7 for the regional lymph nodes model describes the model as demonstrating moderately low separability of the data due to the jagged AUROC. Since the AUC is much closer to a value in the range of 0.75, the model has relatively low discrimination capacity to distinguish between positive and negative classes. Overall, the model is much better at predicting the presence of a metastatic tumor than a regional lymph node.

Discussion

In studies that show 5-year survival data regarding various cancers, a metastatic disease is nearly always a terminal illness [6]. In a study performed by Steeg et al. (2016), a localized melanoma cancer with a near 100% 5-year survival rate drops to 20% when the disease is no longer localized [6]. A similar trend is seen with our cancer investigation, where the 5-year survival rate of metastatic patients with KIRC is less than 10% [7]; furthermore, unlike other renal cell carcinoma types, KIRC has a much higher recurrence and metastasis rate [7]. The survival rate of metastatic patients diagnosed with KIRC versus non-metastatic patients is also shown in a survival analysis using a Kaplan-Meier curve seen in the results. The data further illustrates the idea that metastasis is a terminal illness in itself. Being able to then correctly identify a disease that is metastatic, or vastly

more useful: becoming metastatic, is a monumental aid to patients suffering from kidney renal clear cell carcinoma, and other varieties of the disease.

By analyzing a select amount of principal components derived from a series of patients' RNA genes, we were able to devise a machine learning model that can predict metastasis. Ideally, this model could be applied to a patient's genes over several screenings at different time periods. As the expression levels change, the algorithm would theoretically be able to predict the presence of distant metastasis - as was the intention. The results match literature research that discusses the phenomenon of cancer metastasis being directed by gene expression programs within the tumor cells [8].

A major limitation of the study was the data and hardware that was available. Principal component analysis of our transposed RNA data took vast amounts of time to process. Performing the PCA on the un-transposed data was not at all feasible to the vast amount of genes present. A future goal would be limiting these genes down to the most significant via further research and study into KIRC. Once the highest significance genes are sorted, performing our analysis on the new data may yield more accurate and discernible results. In our case, we severely limited the number of PCs we took and conducted our logistic regression on the smaller subset, although it may not be entirely correct. We also took the results from our PCA and attempted to create a second model that could predict the presence of regional lymph nodes. This second model was unsuccessful, potentially due to the large amounts of missing data among the patients. Future research could focus on collecting more patient data regarding lymph node values, and testing the model against the data again to test its accuracy. In our case, we had results with high inaccuracy.

Conclusion

We sought to find a link connecting KIRC becoming metastatic in a patient, with a patient's respective RNA gene counts via a machine learning model employing logistic regression. By creating a dataframe with our principal components and designing a model, we were able to predict metastasis in a patient with an accuracy of ~95%. We ran a similar model against the presence of regional lymph nodes and were not as successful, as there were a high degree of false positives and false negatives - as determined by an ROC curve.
Overall, the research question was addressed and our metastatic model was successful in interpreting data among a variety of patients. We can conclude that we can accurately predict metastatic data among patients via the use of an RNA sequence alone.

Contributions

Parsa designed the project idea, while he and Matt planned the analysis equally. Parsa and Matt both worked on the code base for the project, with Matt focusing on the PCA and Parsa working on the Logistic Regression. They both worked equally in providing the relevant plots. Michelle worked on the survival analysis of the data, and determining the survival rates of patients with metastatic tumors. All three members contributed equally to the progress report. Matt wrote the methods section of the final report, while Parsa wrote the abstract, introduction, results, discussion, and conclusion. Parsa organized the slides, and Matt designed the slides for the methods section. Parsa did the remaining slides for the introduction, project goal, results, and discussion. Matt organized the code and wrote comment statements describing the various methods and their functions. Michelle wrote the descriptions for the Kaplan-Meier block, and its respective functions.

Parsa Moheban
Mattathias Klassen
Michelle Kojekine

Appendix

```
glm.fit: fitted probabilities numerically 0 or 1 occurred
Call:
glm(formula = as.factor(OS_Status) ~ ., family = binomial, data = kirc.dataset.train)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.4222  -0.6990  -0.4370   0.5737   2.0293

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -2.168e+00  1.197e+00  -1.811 0.070084 .
T_StageT1A   1.726e-02  1.256e+00   0.014 0.989032
T_StageT1B   2.922e-01  1.238e+00   0.236 0.813448
T_StageT2    1.317e+00  1.253e+00   1.051 0.293088
T_StageT2A  -1.438e+01  1.665e+03  -0.009 0.993108
T_StageT2B   2.394e-01  1.697e+00   0.141 0.887774
T_StageT3   -1.480e+01  2.400e+03  -0.006 0.995078
T_StageT3A   1.414e+00  1.226e+00   1.153 0.248970
T_StageT3B   1.865e+00  1.253e+00   1.489 0.136495
T_StageT3C   1.866e+01  2.400e+03   0.008 0.993795
T_StageT4    3.984e+00  2.315e+00   1.721 0.085220 .
M_StageM1    2.382e+00  6.187e-01   3.849 0.000118 ***
N_StageN1    6.247e-01  9.372e-01   0.667 0.505050
N_StageNX   -3.310e-01  3.745e-01  -0.884 0.376764
PC1         -3.339e-02  1.860e-02  -1.795 0.072688 .
PC2          1.499e-02  1.148e-02   1.306 0.191460
PC3         -4.207e-02  2.765e-02  -1.522 0.128129
PC4         -1.035e-02  1.158e-02  -0.893 0.371627
PC5         -6.307e-03  1.018e-02  -0.619 0.535756
PC6          1.610e-02  1.498e-02   1.075 0.282527
PC7         -3.291e-02  4.404e-02  -0.747 0.454833
PC8          1.360e-03  9.452e-03   0.144 0.885573
PC9          1.444e-03  8.233e-03   0.175 0.860792
PC10        -8.167e-03  2.462e-02  -0.332 0.740063
PC11         1.636e-02  1.453e-02   1.126 0.259991
PC12         3.143e-02  3.422e-02   0.918 0.358459
PC13         2.206e-02  1.930e-02   1.143 0.253136
PC14         2.960e-02  2.549e-02   1.161 0.245623
PC15         5.596e-02  5.662e-02   0.988 0.322939
PC16        -1.515e-02  1.332e-02  -1.137 0.255560
PC17        -2.147e-02  6.470e-02  -0.332 0.740056
PC18        -1.743e-02  1.401e-02  -1.244 0.213434
PC19        -2.481e-02  2.191e-02  -1.132 0.257479
PC20        -7.132e-03  1.214e-02  -0.588 0.556796
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 318.15  on 242  degrees of freedom
Residual deviance: 220.72  on 209  degrees of freedom
AIC: 288.72

Number of Fisher Scoring iterations: 15
```

Appendix 1A glm summary for metastatic tumor model

```
glm_pred  M0  M1
      M0 190   4
      M1   8  35

glm_pred_train  M0  M1
            M0 193   5
            M1  14  31
Confusion Matrix and Statistics

          Reference
Prediction  M0  M1
        M0 190   4
        M1   8  35

               Accuracy : 0.9494
                 95% CI : (0.9132, 0.9736)
    No Information Rate : 0.8354
    P-Value [Acc > NIR] : 7.12e-08

                  Kappa : 0.8231

 Mcnemar's Test P-Value : 0.3865

            Sensitivity : 0.8974
            Specificity : 0.9596
         Pos Pred Value : 0.8140
         Neg Pred Value : 0.9794
             Prevalence : 0.1646
         Detection Rate : 0.1477
   Detection Prevalence : 0.1814
      Balanced Accuracy : 0.9285

       'Positive' Class : M1

[1] "Accuracy for the Test set: 0.949367088607595"
```

Appendix 1B Confusion matrix and accuracy summary for metastatic tumor model

```
glm.fit: fitted probabilities numerically 0 or 1 occurred
Call:
glm(formula = as.factor(OS_Status) ~ ., family = binomial, data = kirc.dataset.train)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.0348  -0.6674  -0.1412   0.4926   2.2043

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -1.823e+01  3.192e+03  -0.006 0.995442
T_StageT1A   1.747e+01  3.192e+03   0.005 0.995633
T_StageT1B   1.731e+01  3.192e+03   0.005 0.995673
T_StageT2    1.624e+01  3.192e+03   0.005 0.995941
T_StageT2A  -1.340e+00  4.833e+03   0.000 0.999779
T_StageT2B   6.830e-02  5.504e+03   0.000 0.999990
T_StageT3   -3.133e+00  7.262e+03   0.000 0.999656
T_StageT3A   1.753e+01  3.192e+03   0.005 0.995619
T_StageT3B   1.625e+01  3.192e+03   0.005 0.995938
T_StageT4    3.492e+01  4.184e+03   0.008 0.993341
M_StageM1    3.076e+00  9.076e-01   3.389 0.000703 ***
M_StageMX   -1.445e+01  7.464e+03  -0.002 0.998455
N_StageN1    2.092e+00  1.202e+00   1.740 0.081904 .
PC1          6.759e-03  2.647e-02   0.255 0.798449
PC2         -4.138e-03  1.705e-02  -0.243 0.808269
PC3          1.855e-02  3.665e-02   0.506 0.612814
PC4          4.453e-02  5.444e-02   0.818 0.413412
PC5         -6.370e-02  4.223e-02  -1.508 0.131437
PC6         -8.750e-02  6.594e-02  -1.327 0.184552
PC7         -2.512e-01  2.497e-01  -1.006 0.314411
PC8         -1.878e-02  2.996e-02  -0.627 0.530758
PC9         -2.060e-02  1.453e-02  -1.418 0.156205
PC10        -1.262e-01  1.290e-01  -0.978 0.327935
PC11         3.685e-03  6.183e-02   0.060 0.952477
PC12         7.410e-02  1.544e-01   0.480 0.631296
PC13         9.049e-02  6.738e-02   1.343 0.179247
PC14        -3.681e-03  5.094e-02  -0.072 0.942397
PC15         9.978e-03  1.085e-01   0.092 0.926702
PC16         4.007e-02  3.859e-02   1.039 0.299002
PC17        -3.710e-02  1.322e-01  -0.281 0.779065
PC18         2.574e-02  3.991e-02   0.645 0.519004
PC19         2.185e-02  5.805e-02   0.376 0.706608
PC20        -4.271e-02  3.298e-02  -1.295 0.195290
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 160.723  on 120  degrees of freedom
Residual deviance:  93.011  on  88  degrees of freedom
AIC: 159.01

Number of Fisher Scoring iterations: 17
```

Appendix 2A glm summary for regional lymph nodes model

```
glm_pred N0 N1
       N0 92  5
       N1 19  1

glm_pred_train N0 N1
            N0 86  5
            N1 24  6
Confusion Matrix and Statistics

          Reference
Prediction N0 N1
        N0 92  5
        N1 19  1

               Accuracy : 0.7949
                 95% CI : (0.7103, 0.8639)
    No Information Rate : 0.9487
    P-Value [Acc > NIR] : 1.000000

                  Kappa : -0.0021

 Mcnemar's Test P-Value : 0.007963

            Sensitivity : 0.166667
            Specificity : 0.828829
         Pos Pred Value : 0.050000
         Neg Pred Value : 0.948454
             Prevalence : 0.051282
         Detection Rate : 0.008547
   Detection Prevalence : 0.170940
      Balanced Accuracy : 0.497748

       'Positive' Class : N1

[1] "Accuracy for the Test set: 0.794871794871795"
```

Appendix 2B Confusion matrix and accuracy summary for regional lymph nodes model

Parsa Moheban
Mattathias Klassen
Michelle Kojekine

Bibliography

[1] Z. Hu et al, "lncRNA MSC‐AS1 activates Wnt/β‐catenin signaling pathway to modulate cell proliferation and migration in kidney renal clear cell carcinoma via miR‐3924/WNT5A," Journal of Cellular Biochemistry, vol. 121, (10), pp. 4085-4093, 2020.

[2] C. Wei et al, "ZNF668: a new diagnostic predictor of kidney renal clear cell carcinoma," Anti-Cancer Drugs, vol. Publish Ahead of Print, 2021.

[3] J. W. Pollard, Metastasis: Mechanism to Therapy. Cold Spring Harbor, New York: Cold Spring Harbor Laboratory Press, 2019.

[4] Anonymous "metastasis," in Anonymous Berlin, Heidelberg: Springer Berlin Heidelberg, pp. 1069-1069.

[5] S. Narkhede, “Understanding AUC - roc curve,” Medium, 15-Jun-2021. [Online]. Available: https://towardsdatascience.com/understanding-auc-roc-curve-68b2303cc9c5. [Accessed: 08-Dec-2021].

[6] P. S. Steeg, "Targeting metastasis," Nature Reviews. Cancer, vol. 16, (4), pp. 201-218, 2016.

[7] K. Gong et al, "Comprehensive analysis of lncRNA biomarkers in kidney renal clear cell carcinoma by lncRNA-mediated ceRNA network," PloS One, vol. 16, (6), pp. e0252452-e0252452, 2021.

[8] F. Hartung et al, "A core program of gene expression characterizes cancer metastases," Oncotarget, vol. 8, (60), pp. 102161-102175, 2017.