Recurrence Risk Assessment for Treated Cancer Patients: A Comprehensive Probability Based on Clinical and Molecular Factors

Sabina Akelbek[1] **McKenzie Hebert**[2]

[1]Purdue University, [2]Arizona State University

SCHOOL OF PUBLIC HEALTH
BIOSTATISTICS
UNIVERSITY OF MICHIGAN

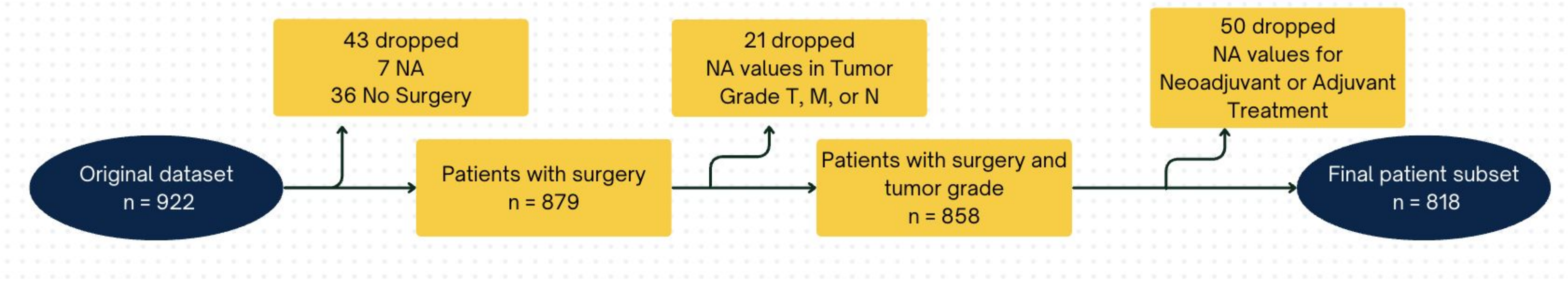
Background

Cancer Data Science

- The Oncotype Score is a genomic test that predicts recurrence risk and guides chemotherapy decisions for specific breast cancer types.
- The limitations of this score is that it can only be used for those with Luminal-A type cancer. Additionally, it only informs adjuvant chemotherapy decisions.

The Data

- Original clinical features dataset included 922 patients with 98 variables.
- The data subset used included 818 patients who underwent surgery, with features encompassing surgery type, adjuvant and neoadjuvant treatments (radiation, chemo, endocrine, anti-HER2 Neu), molecular subtype, and tumor grades (T, M, N).



Original dataset
n = 922

43 dropped
7 NA
36 No Surgery

Patients with surgery
n = 879

21 dropped
NA values in Tumor Grade T, M, or N

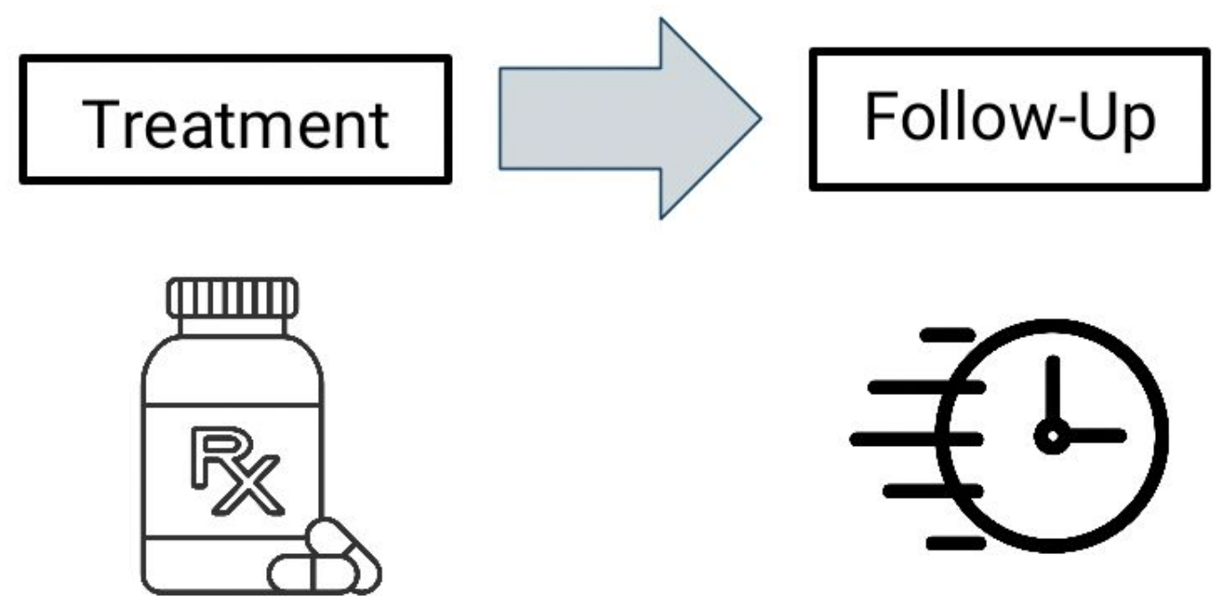
Patients with surgery and tumor grade
n = 858

50 dropped
NA values for Neoadjuvant or Adjuvant Treatment

Final patient subset
n = 818

Motivation and Goal

- Our project explores how to more effectively assess recurrence risk using a combination of clinical and molecular data for treatment planning.



Treatment

Follow-Up

- In order to address genomic test limitations (restricted eligibility, cost, chemotherapy focus), an interactive R Shiny application was developed as a tool for clinicians and patients.

Methods

Outcome Definition

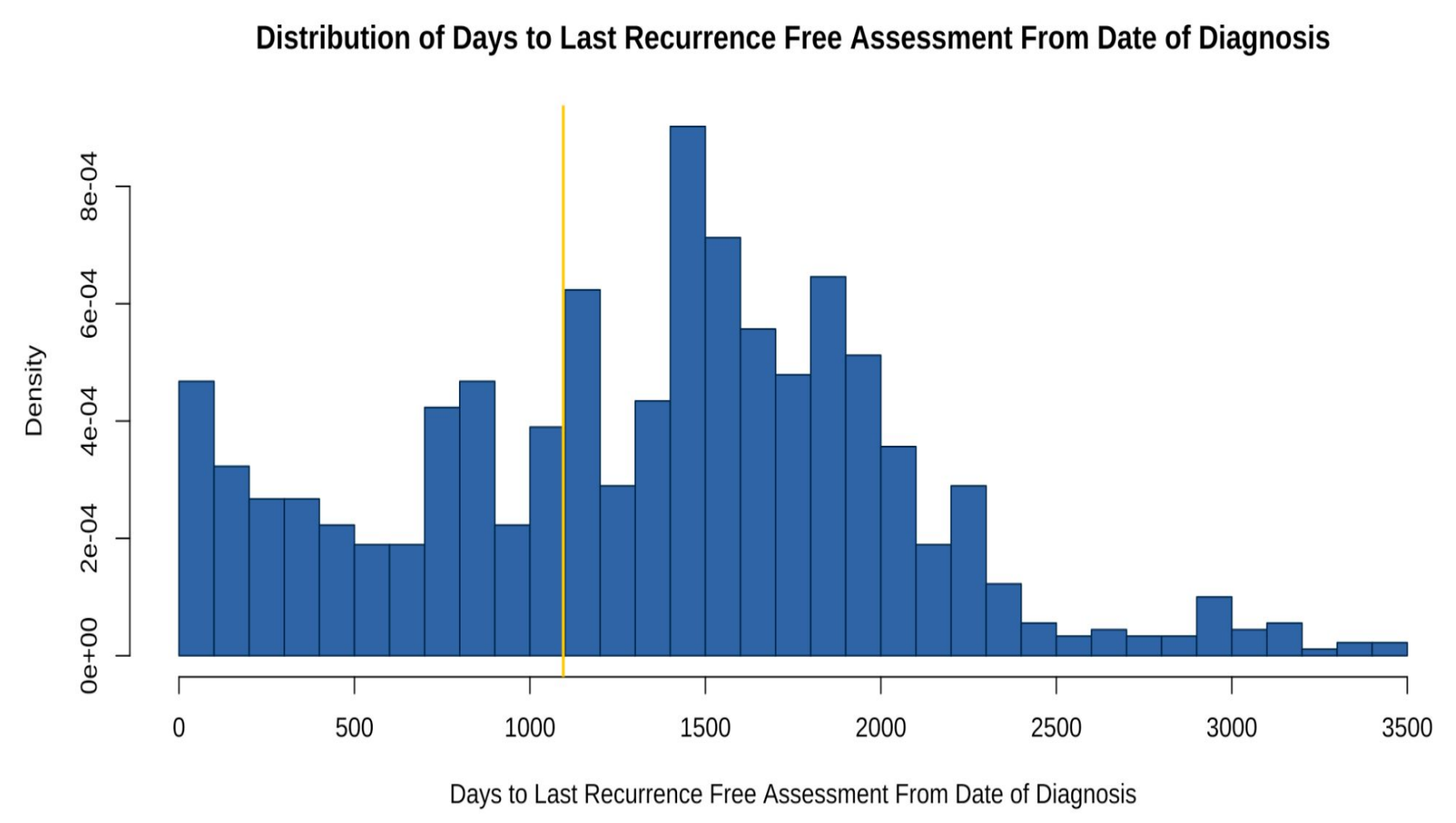


Figure 1. Distribution of patient follow-up time, with a yellow line representing the value we chose (3 years).

Figure 2. The table displays the percentage of patients followed-up at 1, 2, and 3 years. We face a trade off: more certainty in outcome versus more minority class events.

Time	Followed (%)	No Recurrence	Recurrence
1 Yr	91.6%	806	12
2 Yrs	83.5%	780	38
3 Yrs	69.4%	762	56
Total	100%	734	84

SMOTE

- Synthetic Minority Oversampling Technique
- Generates synthetic examples for the minority class, helping to balance imbalanced datasets and improve model performance.

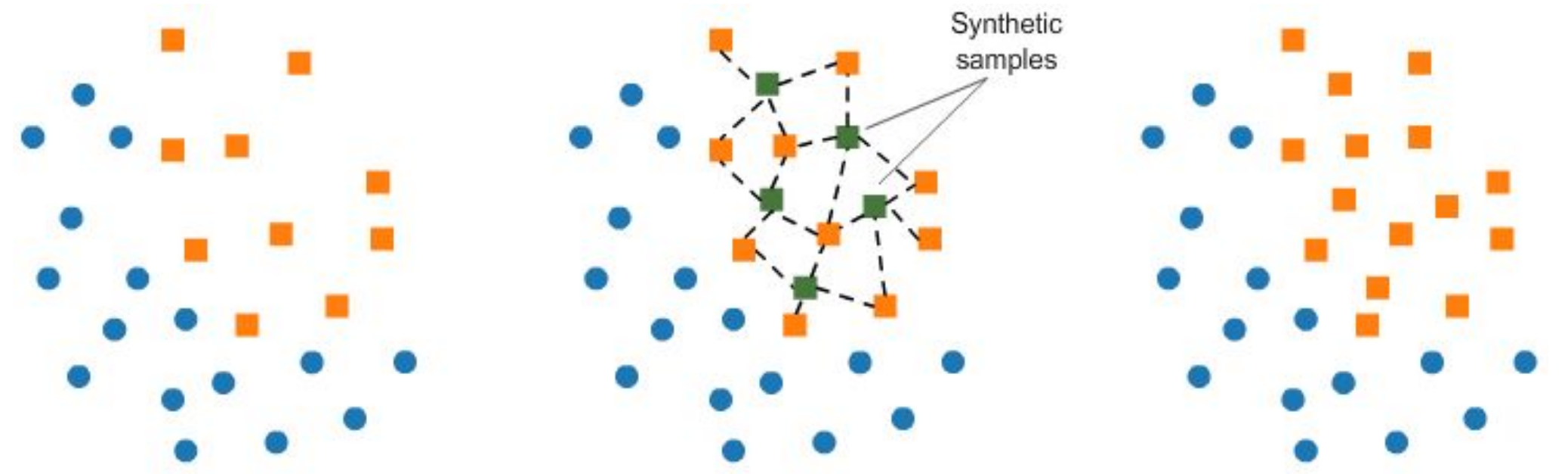


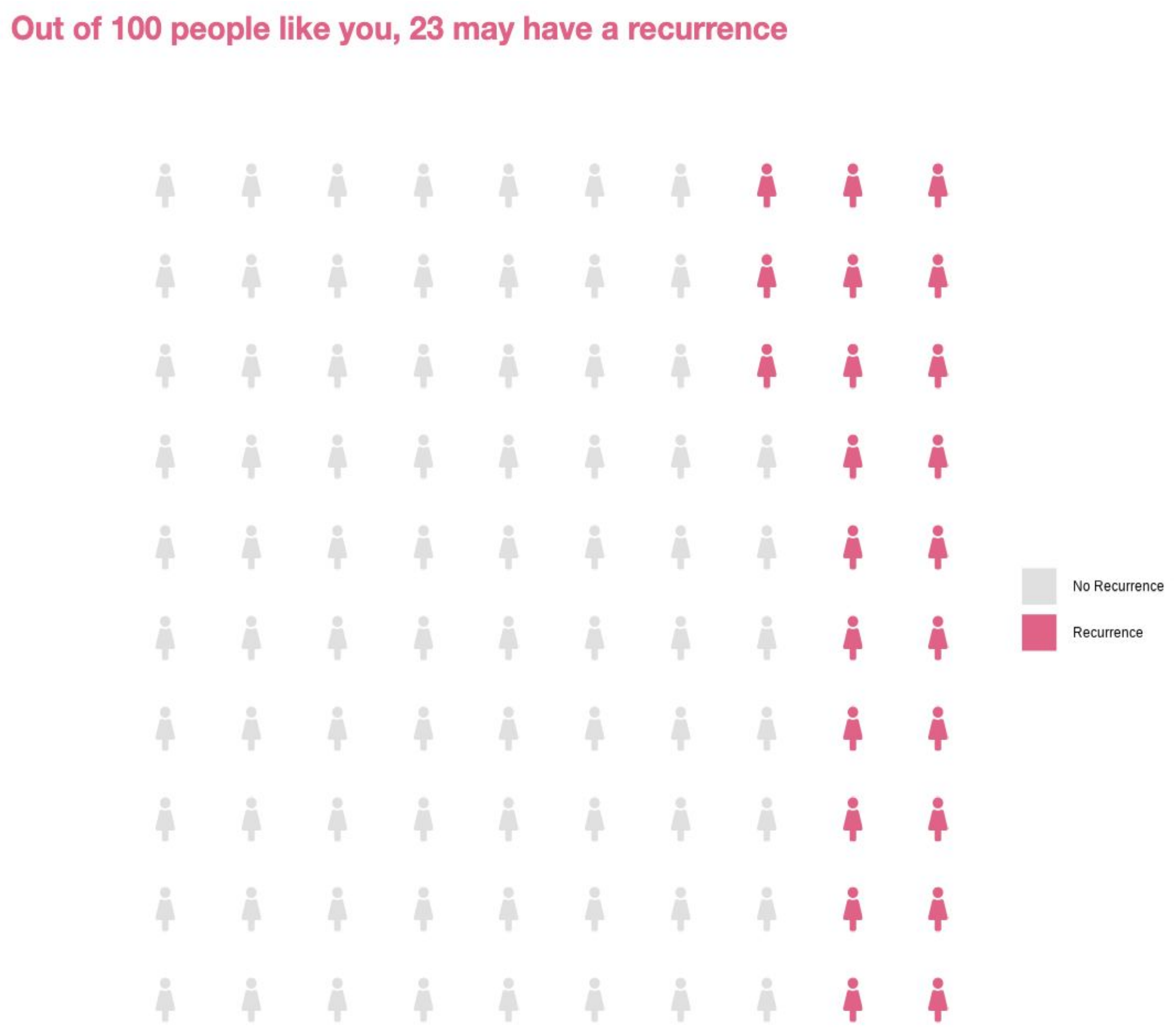
Figure 3. A visual representation of the way that SMOTE creates new synthetic instances.

XGBoost

- XGBoost is a machine learning method that builds a series of decision trees, with each new tree learning from the mistakes of the previous ones, ultimately combining their predictions to give you a probability score for classification.
- Advantages of XGBoost:
  - ○ Works well with highly imbalanced datasets, built in class weighting functions to address that issue.
  - ○ Has built in feature importance function which is helpful with interpretability and understanding the model's drivers.
  - ○ Built-in LASSO regularization. Important for preventing overfitting.
- Our model combines SMOTE with hyperparameter tuning to enhance the performance of the model.
- We performed our hyperparameter tuning using a GridSearch with a 5-fold cross validation. This method searches over every possible grid to maximize a chosen evaluation metric.

R Shiny Application

- An interactive web application was developed to provide clinicians with a tool for inputting patient diagnosis data and neoadjuvant treatments.
- Generates a recurrence risk probability.
  - ○ Clinicians can use the score to determine adjuvant treatment course.
  - ○ Patients can use the score to visualize their risk.
- Features a human pictogram display to make the recurrence risk score easy to understand for both patients and clinicians.



Out of 100 people like you, 23 may have a recurrence

Figure 4. R Shiny App Recurrence Risk Pictogram. This figure illustrates the user-friendly output of our R Shiny application, translating patient data into an easily digestible visual of recurrence probability to aid treatment discussions between clinicians and patients.

Results

Evaluation Metrics

- Calibration plot: a plot that displays the predicted probabilities on the x axis and the observed outcomes on the y axis.
- AUC: a single number that measures the area under the ROC curve. Models aim to be close to 1, as 0.5 represents random chance.
- Brier Score: measures the accuracy of probabilistic predictions for binary outcomes, calculated as the mean squared difference between predicted probabilities and actual outcomes; a lower score indicates better calibration and accuracy.

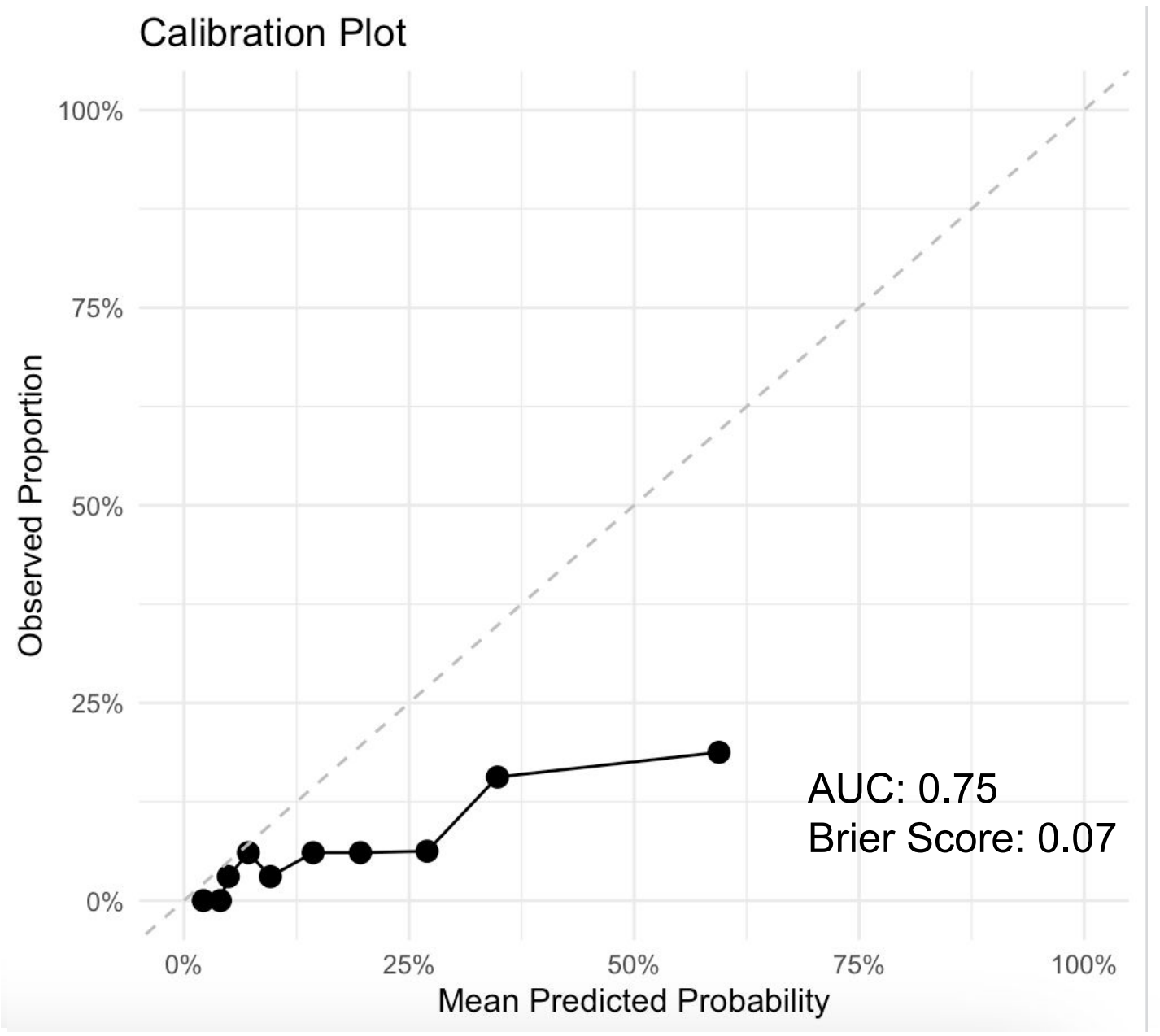


Figure 5. A calibration plot for our model with the AUC and Brier Score

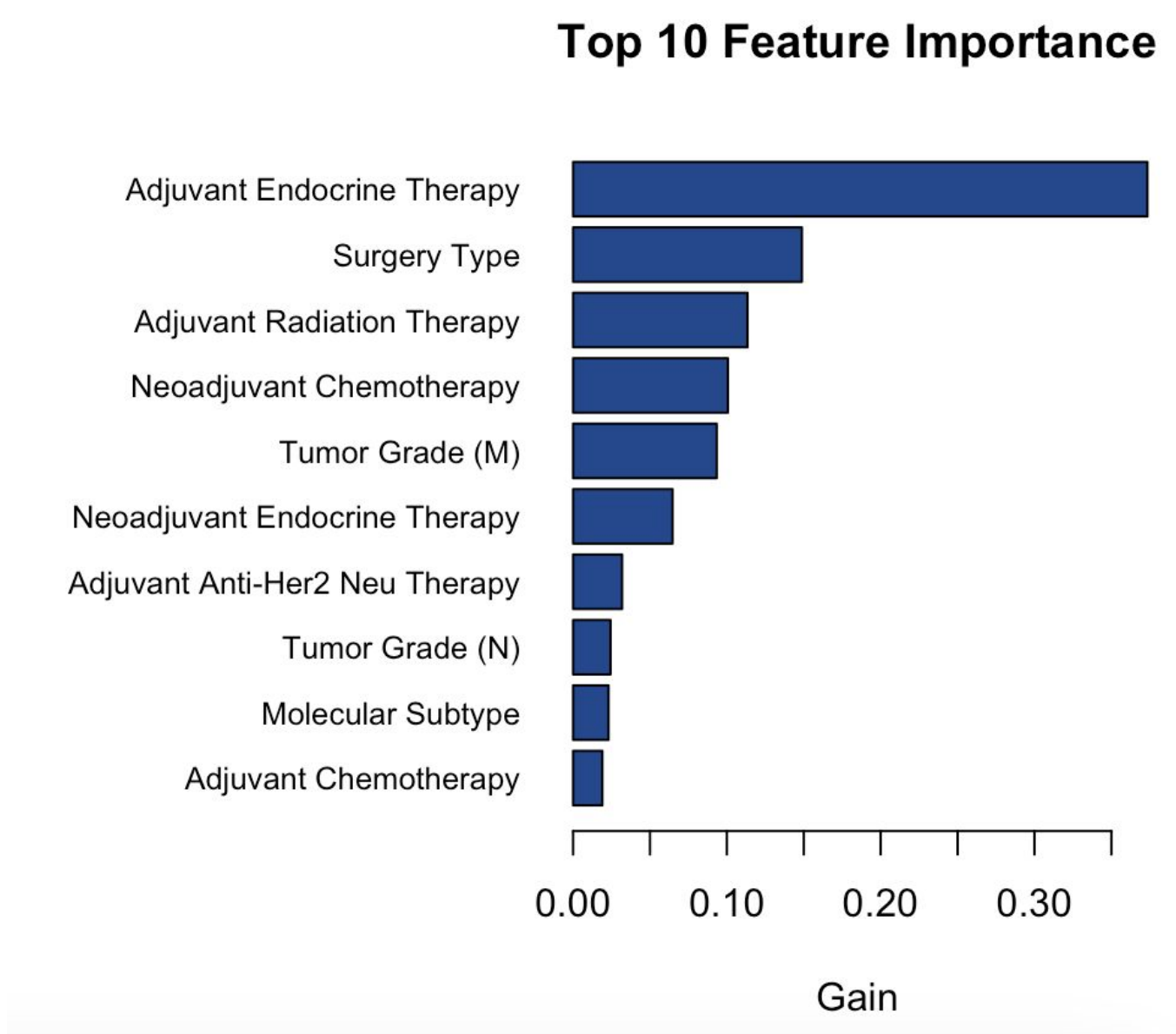


Figure 6. A bar chart displaying the top 10 most important features

Limitations

- Our model treats recurrence as a binary outcome. This is limiting because it cannot capture the details and information about when recurrence will occur which is an important factor in treatment decisions
- For clinical use, other predictors such as sociodemographic and comorbidity data should be incorporated.
- Predictions are limited to a 3-year timeline. As seen in Figure 2, we are missing events that occur past 3 years as well as potentially missing recurrence events that occur after the study was completed.
- Only 69.4% of the patients in our dataset were followed up with to the 3 year mark, representing a portion of uncertainty in their outcome.

Future Work

- To further explore the performance of our model, we could compare it to oncotype score predictability in a prospective trial.
- Expand R Shiny Functionality: Develop new features, such as the ability to compare predicted outcomes for different hypothetical treatment scenarios.
- Time-to-Event Prediction: Adapt the model to predict specific time-to-recurrence probabilities rather than just binary outcomes.

Acknowledgement: We would like to thank Dr. Krithika Suresh, Dr. Nicholas Hartman, and Grant Carr for their generous help and support throughout this project, as well as Dr. Matthew Zawistowski and the entire staff for the opportunity to be part of the Big Data Summer Institute.

References

1. Saha, A., Harowicz, M. R., Grimm, L. J., Kim, C. E., Ghate, S. V., Walsh, R., & Mazurowski, M. A. (2018). A machine learning approach to radiogenomics of breast cancer: a study of 922 subjects and 529 DCE-MRI features. British Journal of Cancer, 119(4), 508-516. https://doi.org/10.1038/s41416-018-0185-8
2. The Cancer Imaging Archive (TCIA). (n.d.). Duke Breast Cancer MRI. Retrieved from https://wiki.cancerimagingarchive.net/pages/viewpage.action?pageId=70226903