Final Project Report

BIOS:7210 Survival Data Analysis

Survival Prediction in Breast Cancer: A Cox Proportional Hazards Model and DeepSurv Neural Network Comparison

Student Name: **Samuella Boadi**

December 12, 2025

Abstract

Accurate prediction of breast cancer survival is essential for guiding clinical decision-making and identifying high-risk patients. Using data from breast cancer cases in the SEER registry, we evaluated prognostic factors associated with overall survival and compared the predictive performance of a Cox proportional hazards model with a DeepSurv neural network. Univariable and multivariable analyses identified age, race, tumor grade, N stage, tumor size, hormone receptor status, and lymph node involvement as significant predictors of survival. Violations of the proportional hazards assumption for estrogen and progesterone receptor status were addressed using stratification. The final Cox model demonstrated good discrimination (C-index = 0.691), outperforming the DeepSurv model (C-index = 0.658). These findings highlight the continued strength and interpretability of traditional survival models in structured clinical datasets, while also indicating areas where deep learning methods may require larger or richer data sources to achieve comparable performance.

Keywords: Survival analysis, Cox model, Deep learning, DeepSurv, Breast cancer, SEER, Prognostic modeling.

1 Introduction

Breast cancer remains the most commonly diagnosed malignancy among women worldwide and a leading cause of cancer-related mortality, accounting for more than 2.3 million new cases and nearly 685,000 deaths in 2020 alone [Cuthrell and Tzenios, 2023]. In the United States, breast cancer represents approximately 30% of all new cancer diagnoses in women each year . The disease is biologically heterogeneous, consisting of multiple histopathological and molecular subtypes that differ in prognosis, therapeutic response, and patterns of recurrence [Diana, 2025]. Survival analysis is essential in oncology for identifying prognostic factors and predicting patient outcomes. The Cox proportional hazards model is widely used because of its interpretability and ability to handle censored data, but its assumptions may limit performance in the presence of complex, nonlinear relationships. With the growth of high-dimensional clinical data, machine-learning approaches—particularly deep learning—have emerged as flexible alternatives for improving survival prediction [Baidoo and Rodrigo, 2025].

The emergence of deep learning methods has introduced new possibilities for survival modeling, with DeepSurv representing a notable advancement that combines the Cox proportional hazards framework with deep neural networks to model complex patient-treatment interactions [Katzman et al., 2018a, Baidoo and Rodrigo, 2025].

Although classical Cox regression remains the dominant analytical tool in survival modeling, recent advances in machine learning have led to the development of more flexible approaches capable of modeling complex nonlinear relationships and high-dimensional interactions. One such method is DeepSurv, a deep neural network extension of the Cox model that uses multilayer perceptrons to learn nonlinear risk functions while retaining the partial likelihood framework of Cox regression [Katzman et al., 2018b]. DeepSurv has demonstrated improved predictive accuracy in several biomedical applications; however, it sacrifices interpretability and may be sensitive to tuning parameters and training procedures. Furthermore, evaluation metrics such as the concordance index (C-index), while widely used, have notable limitations—especially in heavily censored datasets—regarding calibration and sensitivity to the censoring distribution [Hartman et al., 2023]. Thus, careful comparison between DeepSurv and classical Cox modeling is warranted.

In this project, we pursue two primary objectives. First, we conduct a comprehensive survival analysis using a Cox proportional hazards model on a training dataset comprising 90% of the eligible SEER patients. This includes model building, assessment of functional forms, proportional hazards evaluation, and careful interpretation of results. Second, we evaluate the predictive performance of the final Cox model against DeepSurv using the held-out 10% of patients. This comparison highlights the strengths and limitations of classical statistical modeling relative to modern deep learning approaches in survival data analysis.

2 Methods

Data Source and Preprocessing

Data for this study were obtained from the Surveillance, Epidemiology, and End Results (SEER) breast cancer registry. After importing the dataset into R, we performed standard preprocessing steps including variable renaming, missing data checks, and creation of an event indicator where `status_event = 1` represents death and `status_event = 0` represents censoring. A survival object was constructed as

`Surv(Survival.Months, status_event)`.

Categorical variables were converted to factors, and the data were randomly partitioned into a 90% training set and a 10% validation set to support the development and evaluation of predictive models.

Exploratory Analyses and Collinearity Assessment

Descriptive summaries and correlation analyses were conducted to understand relationships among predictors. Pearson correlations for continuous variables showed generally weak associations, with only Regional Nodes Examined and Regional Nodes Positive displaying a moderate correlation (r = 0.41). Cramér's V revealed redundancy among categorical staging variables: X6th Stage overlapped strongly with T Stage and N Stage, and `differentiate` was perfectly correlated with `Grade`. To avoid multicollinearity, X6th Stage, T stage and `differentiate` were removed. These evaluations indicated no remaining collinearity concerns among predictors.

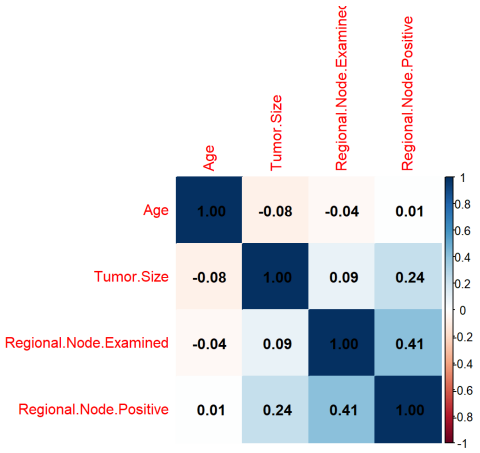

Figure 1: Correlation plot for continuous variables

Model Building

Univariable Cox models were first fitted to screen candidate predictors, and variables with $p < 0.20$ were included in the multivariable model. A full Cox model was then refined using backward elimination, supported by a confounding assessment in which variables were retained if their removal changed any remaining coefficient by 20% or more. This process resulted in a main-effects model including Age, Race, N Stage, Grade, Tumor Size, Estrogen Status, Progesterone Status, Regional Nodes Examined, and Regional Nodes Positive. Functional form was evaluated using penalized spline terms, which showed that all continuous predictors were approximately linear on the log-hazard scale, permitting linear specifications in the final model.

Interaction Assessment

A full two-way interaction Cox model was fitted to evaluate potential interactions. Because most interaction terms were non-significant, stepwise selection using Akaike's Information Criterion (AIC) was performed on the full interaction model. The resulting model retained several interaction terms; however, a likelihood ratio test comparing this model to the spline-based main-effects model showed no significant improvement in model fit ($p = 0.36$). Therefore, the main-effects model was retained for inference.

Proportional Hazards Assumption and Stratification

The proportional hazards assumption was evaluated using Schoenfeld residuals via the `cox.zph` test. Estrogen Status and Progesterone Status violated the PH assumption. To address this, the final Cox model was stratified on both variables. Stratification allows each stratum to have its own baseline hazard while estimating a common effect for the remaining covariates. After stratification, all PH tests were non-significant, and the global test ($p = 0.94$) indicated that the proportional hazards assumption was satisfied.

Influence diagnostics were assessed using DFBETA residuals. No extreme influential observations were identified.

DeepSurv Neural Network Model

To compare traditional Cox modeling with modern deep learning approaches, a DeepSurv model was trained using the `dnn` package. Continuous and categorical variables were converted into a numeric design matrix via `model.matrix`. A neural network with two hidden layers of 64 and 32 units was constructed, using ReLU activation in hidden layers and an identity activation in the output layer. The model was trained using the Cox partial likelihood loss for 2000 epochs with mini-batch optimization.

Predicted risk scores were obtained from the fitted model, and model discrimination was evaluated using the concordance index.

3 Results

Descriptive Statistics

The table below summarizes baseline characteristics of 3,621 breast cancer patients stratified by survival status (Alive vs. Dead). Patients who died were more likely to have advanced disease at diagnosis, as reflected by higher proportions in N2–N3 stages, T3–T4 tumor stages, and more aggressive histologic grades (Grade III–IV). They also had substantially worse differentiation and were more often negative for estrogen and progesterone receptors—factors associated with poorer prognosis. Additionally, non-survivors tended to be older (mean age 55.2 vs. 53.7 years), had larger tumors on average, and had more positive lymph nodes, indicating more advanced spread of disease. The table shows that demographic and clinical factors linked to more aggressive tumor biology and more advanced staging are more common among patients who died, consistent with their worse survival outcomes.

Characteristic	Alive (n = 3065)	Dead (n = 556)
Race		
White	2600 (84.8%)	462 (83.1%)
Black	205 (6.7%)	65 (11.7%)
Other	260 (8.5%)	29 (5.2%)
Marital Status		
Married	2050 (66.9%)	322 (57.9%)
Single	466 (15.2%)	93 (16.7%)
Divorced	353 (11.5%)	81 (14.6%)
Widowed	169 (5.5%)	46 (8.3%)
Separated	27 (0.9%)	14 (2.5%)
T Stage		
T1	1308 (42.7%)	140 (25.2%)
T2	1338 (43.5%)	273 (49.1%)
T3	368 (12.0%)	104 (18.7%)
T4	57 (1.9%)	39 (7.0%)
N Stage		
N1	2204 (71.9%)	243 (43.7%)
N2	594 (19.4%)	150 (27.0%)
N3	267 (8.7%)	163 (29.3%)
6th Stage		
IIA	1088 (35.5%)	88 (15.8%)
IIB	889 (29.0%)	120 (21.6%)
IIIA	777 (25.4%)	165 (29.7%)
IIIB	44 (1.4%)	20 (3.6%)
IIIC	267 (8.7%)	163 (29.3%)
Differentiation		
Well differentiated	445 (14.5%)	35 (6.3%)
Moderately differentiated	1841 (60.1%)	271 (48.7%)
Poorly differentiated	770 (25.1%)	242 (43.5%)
Undifferentiated	9 (0.3%)	8 (1.4%)
Grade		
Grade I	445 (14.5%)	8 (1.4%)
Grade II	1841 (60.1%)	271 (48.7%)
Grade III	770 (25.1%)	242 (43.5%)
Grade IV	9 (0.3%)	35 (6.3%)
A Stage		
Regional	3016 (98.4%)	525 (94.4%)
Distant	49 (1.6%)	31 (5.6%)
Estrogen Status		
Positive	2919 (95.2%)	460 (82.7%)
Negative	146 (4.8%)	96 (17.3%)
Progesterone Status		
Positive	2611 (85.2%)	180 (32.4%)
Negative	454 (14.8%)	376 (67.6%)
Continuous Variables (mean ± SD)		
Age (years)	53.7 ± 9.8	55.2 ± 9.8
Tumor Size	29.3 ± 20.4	37.4 ± 24.4
Nodes Examined	14.3 ± 8.0	15.1 ± 8.5
Nodes Positive	3.6 ± 4.4	7.2 ± 7.2

Table 1: Baseline characteristics of breast cancer patients by survival status.

The plot below shows that patients who died were generally older, with a higher median age and a slightly wider interquartile range compared to survivors. The "Dead" group also shows several older observations extending into the upper whisker, indicating that mortality was more common among older patients. In contrast, the "Alive" group tends to cluster at younger ages with a lower median age.

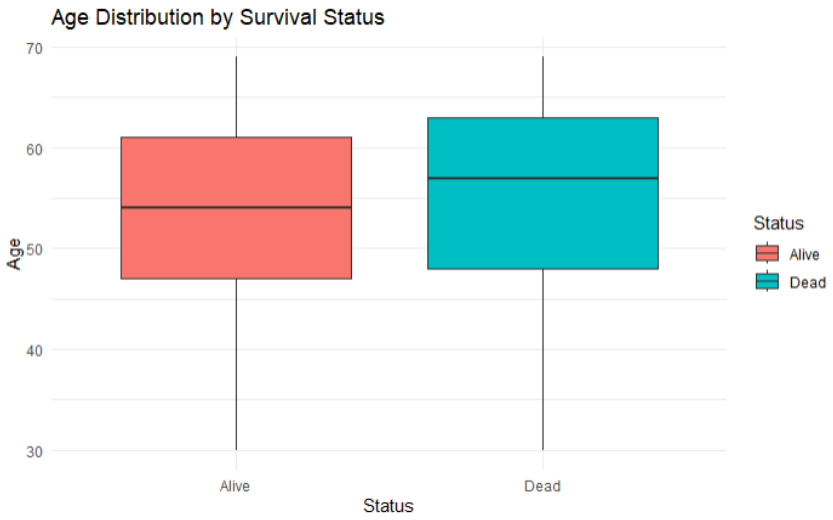

Figure 2: age distribution of patients who were alive versus

Continuous predictors demonstrated approximately linear relationships with the log-hazard, supporting their linear specification in the final model. The penalized spline plots demonstrate that Tumor Size, Regional Node Examined, and Regional Node Positive exhibit approximately linear relationships with the log-hazard across their observed ranges, supporting their use as linear terms in the Cox model. Age shows a slight curvature at the extremes, but the degree of nonlinearity is mild and does not meaningfully change the hazard direction or interpretation. Therefore, for simplicity and interpretability, age was retained as a linear covariate. Assumptions were all met.

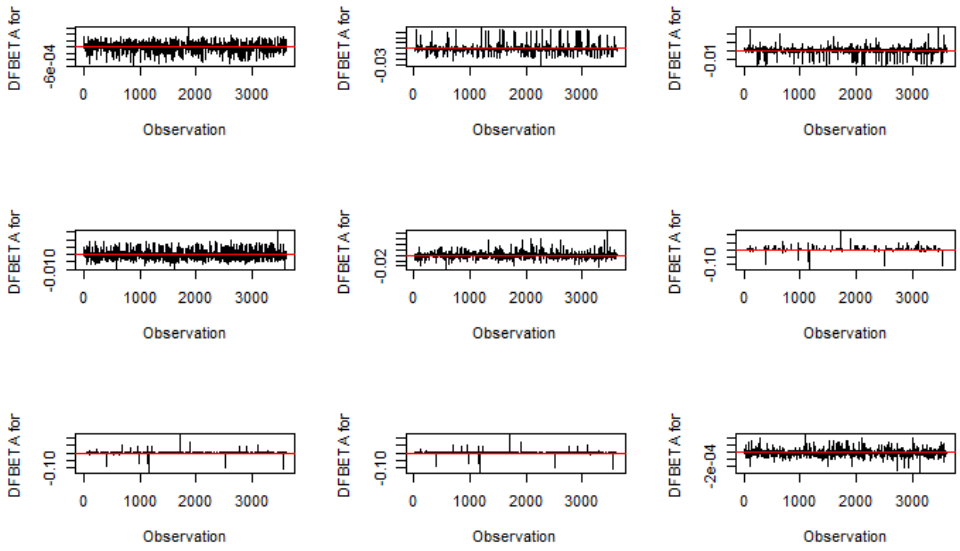

Figure 3: age distribution of patients who were alive versus

The plots above assess how much each individual observation influences the estimated regression coefficients in the stratified Cox model. Across all coefficients, the plotted values are centered closely around

zero, with only small fluctuations and no extreme spikes. This pattern indicates that no single observation exerts undue influence on the parameter estimates.

3.1 Final model and Interpretation

The hazard ratio for age (HR = 1.022) in the table below indicates that each additional year of age is associated with a 2.2% increase in the hazard of death. Although this effect appears modest at the unit scale, it implies a clinically meaningful increase in mortality risk across decades of life. Compared with Black patients (reference group), those classified as "Other" race experienced a substantially lower hazard of death (HR = 0.45), and White patients also had a significantly reduced hazard (HR = 0.69). These findings suggest that racial differences may reflect disparities in tumor biology, access to care, comorbid conditions, or socioeconomic factors captured indirectly by race.

Nodal involvement was among the strongest predictors of survival. Patients with N2 disease had a 64% higher hazard of death compared with those with N1 disease, and those with N3 disease had nearly double the hazard (HR = 1.96). This pattern reflects the well-established prognostic importance of regional lymphatic spread in breast cancer, where increasing nodal burden corresponds to more aggressive disease and a greater likelihood of distant metastasis.

Table 2: Final Stratified Cox Proportional Hazards Model Results

Variable	**HR (exp(coef))**	**Lower 95% CI**	**Upper 95% CI**	**p-value**
Age	1.0223	1.0124	1.0322	< 0.0001
Race: Other	0.4532	0.2920	0.7036	< 0.0001
Race: White	0.6924	0.5326	0.9000	0.0060
N Stage: N2	1.6390	1.3577	1.9815	< 0.0001
N Stage: N3	1.9551	1.3557	2.8195	< 0.0001
Grade 1	0.2219	0.0964	0.5107	< 0.0001
Grade 2	0.3459	0.2153	0.5544	0.0036
Grade 3	0.5123	0.3073	0.8541	0.0650
Tumor Size	1.0071	1.0047	1.0100	< 0.0001
Regional Nodes Examined	0.9713	0.9584	0.9843	< 0.0001
Regional Nodes Positive	1.0557	1.0371	1.0745	< 0.0001

Tumor grade was also a significant prognostic factor. Relative to Grade IV tumors (reference), Grade I and Grade II tumors were associated with markedly lower hazards of death (78% and 65% reductions, respectively), reflecting the more favorable biology of well- and moderately-differentiated tumors. Grade III showed a weaker, borderline association with hazard reduction, consistent with its intermediate biological behavior. These results reaffirm the predictive value of histologic differentiation in breast cancer.

Tumor size, measured in millimeters, demonstrated a positive association with mortality (HR = 1.007). Each additional millimeter increased the hazard by approximately 0.7%, and the cumulative effect becomes clinically meaningful when considering tumors that differ in size by several centimeters. Lymph node counts provided additional insight: each additional node examined was associated with a 2.9% reduction in hazard, potentially reflecting stage migration, more thorough surgical clearance, or improved classification of disease extent. Conversely, each additional positive lymph node increased the hazard by 5.6%, emphasizing the prognostic impact of metastatic spread within the lymphatic system.

Together, these findings demonstrate that tumor burden (tumor size, nodal stage, and number of positive nodes), age at diagnosis, histologic grade, and race play major roles in predicting survival among women with breast cancer. The direction and magnitude of effects align with established clinical knowledge, supporting the validity of the final Cox model and its relevance for prognostic assessment.

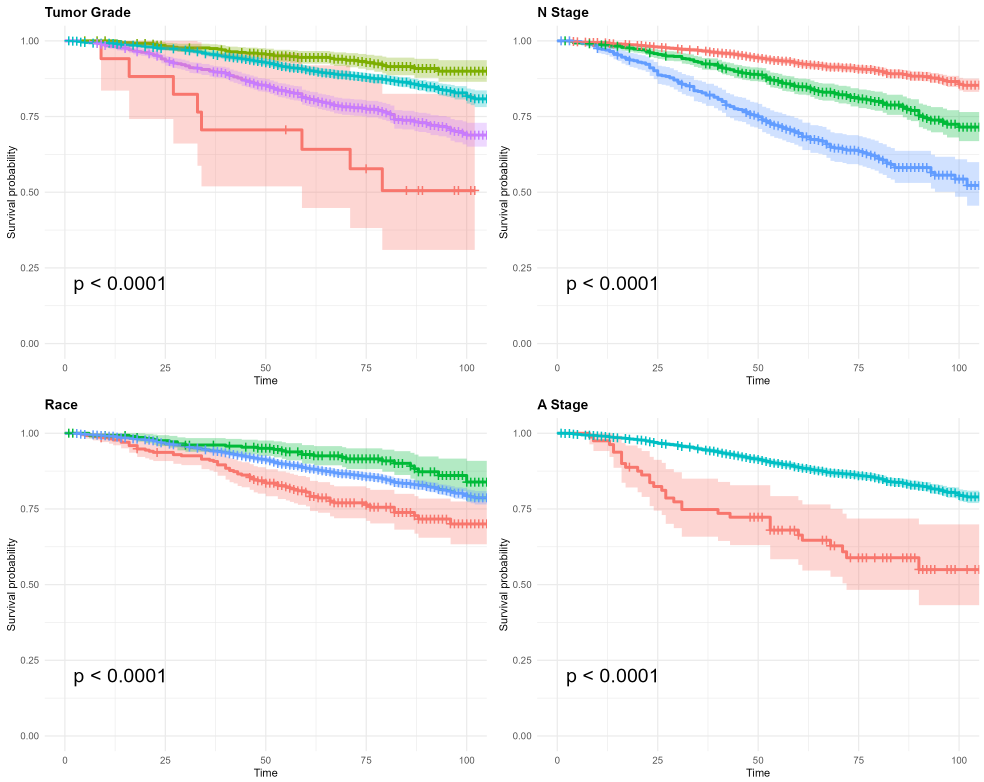

Figure 4: KM Plots

The plots above displays Kaplan–Meier survival curves for four key predictors in the dataset: tumor grade, N stage, race, and A stage. Across all variables, the log-rank tests indicated highly significant differences in survival ($p < 0.0001$). Survival decreased consistently with increasing disease severity. Patients with poorly differentiated or anaplastic tumors (Grade III/IV) experienced markedly lower survival compared to those with well or moderately differentiated tumors. Similarly, more advanced nodal involvement showed a strong gradient: patients with N3 disease had the poorest survival, followed by N2 and N1.

Differences by race were also evident, with Black patients showing lower survival probabilities than White and Other racial groups throughout follow-up. Finally, patients with distant A-stage disease demonstrated substantially reduced survival relative to those with regional disease. Together, these patterns indicate that tumor biology (grade), disease extent (nodal and A stage), and demographic factors (race) are strongly associated with overall survival and visually reinforce the variables identified as important predictors in the Cox model.

Part II: Comparison between Cox model and DeepSurv

Using concordance index (C-index) for evaluation, DeepSurv achieved

$$C_{\text{DeepSurv}} = 0.658,$$

whereas the stratified Cox model achieved a higher value of

$$C_{\text{Cox}} = 0.691.$$

This indicates that the Cox model was better able to distinguish between patients with higher versus lower survival risk in the validation dataset. The superior performance of the Cox model suggests that the underlying relationships between covariates and the hazard were largely linear and well-captured by the proportional hazards structure. Consequently, the additional nonlinear modeling flexibility offered by DeepSurv did not translate into improved prediction accuracy for this dataset.

3.2 Pros and Cons of DeepSurv vrs Cox Model

Although DeepSurv offers the ability to model nonlinear effects and interactions without explicit specification and provides a flexible architecture capable of capturing complex patterns in high-dimensional data while retaining the proportional hazards structure, it also has important limitations, including the need for large sample sizes, extensive hyperparameter tuning, and computational intensity, as well as reduced interpretability in clinical settings. In contrast, the traditional Cox model provides easily interpretable hazard ratios, is computationally efficient and robust in moderate sample sizes, and performs well when covariate effects are approximately linear, with the option to incorporate splines, stratification, or interactions when needed. However, the Cox model also assumes proportional hazards and may fail to capture complex nonlinear relationships without additional modeling. Overall, despite the theoretical flexibility of DeepSurv, the stratified Cox model demonstrated superior predictive performance in this dataset and remains the more interpretable and reliable choice for these clinical data.

4 Discussion

This study identified several important clinical predictors of breast cancer survival and demonstrated that a carefully specified Cox proportional hazards model can provide strong and interpretable prognostic information. However, several limitations should be noted. The SEER dataset does not include treatment details, comorbidities, or molecular subtype information, all of which are known to influence survival and could lead to residual confounding in the model. Some covariates, particularly the staging variables, exhibited substantial redundancy, requiring removal to avoid multicollinearity. Additionally, the initial model failed the proportional hazards assumption for estrogen and progesterone receptor status, indicating time-varying effects that could not be captured without stratification. While stratification resolved this issue, it prevents direct estimation of hazard ratios for these variables.

4.1 Conclusion

Model performance comparisons showed that the stratified Cox model outperformed the DeepSurv neural network on the validation dataset, suggesting that the underlying relationships in this dataset are primarily linear and well handled by traditional survival methods. DeepSurv may also require larger sample sizes, more tuning, or richer covariate information to demonstrate clear advantages. Overall, despite these limitations, the final Cox model provided stable estimates and strong discrimination, supporting its usefulness for clinical interpretation and prediction in this setting.

References

Theophilus Gyedu Baidoo and Hansapani Rodrigo. Data-driven survival modeling for breast cancer prognostics: A comparative study with machine learning and traditional survival modeling methods. *PloS one*, 20(4):e0318167, 2025.

Kimberly Morton Cuthrell and Nikolaos Tzenios. Breast cancer: updated and deep insights. *International Research Journal of Oncology*, 6(1):104–118, 2023.

Anukhanova Diana. Breast cancer epidemiology molecular subtypes and diagnostic advancements. , 4(2 (83)):576–586, 2025.

Nathaniel Hartman, SeungJun Kim, Kaiming He, and John D Kalbfleisch. Pitfalls of the concordance index for survival outcomes. *Statistics in Medicine*, 42(13):2179–2190, 2023.

Jared L Katzman, Uri Shaham, Alexander Cloninger, Jonathan Bates, Tingting Jiang, and Yuval Kluger. Deepsurv: personalized treatment recommender system using a cox proportional hazards deep neural network. *BMC medical research methodology*, 18(1):24, 2018a.

Jared L Katzman et al. Deepsurv: personalized treatment recommender system using a cox proportional hazards deep neural network. *BMC Medical Research Methodology*, 18:24, 2018b.

Appendix: R Code for Survival and DeepSurv Analysis

```
## -----------------------------
## 1. Load packages and data
## -----------------------------
library(survival)
library(survminer)
library(dplyr)
library(broom)
library(ggplot2)
library(tableone)
library(splines)
library(glmnet)
library(MASS)
library(corrplot)
library(vcd)
library(patchwork)
library(dnn)
library(survcomp)

set.seed(7210)

## Read SEER data
seer <- read.csv("C:/Users/boadi/OneDrive - University of Iowa/Documents/fall2025/Fall2025/Survival/SEE

## Fix name
seer <- seer %>%
  rename(Regional.Node.Positive = Reginol.Node.Positive)

## Create event indicator and factors
seer <- seer %>%
  mutate(
    status_event = ifelse(Status == "Dead", 1, 0),
    surv_obj     = Surv(time = Survival.Months, event = status_event),
    Race                = factor(Race),
    Marital.Status      = factor(Marital.Status),
    T.Stage             = factor(T.Stage),
    N.Stage             = factor(N.Stage),
    X6th.Stage          = factor(X6th.Stage),
    differentiate       = factor(differentiate),
    Grade               = factor(Grade),
    A.Stage             = factor(A.Stage),
    Estrogen.Status     = factor(Estrogen.Status),
    Progesterone.Status = factor(Progesterone.Status)
```

```
  )

## -----------------------------
## 2. Train / validation split
## -----------------------------
n         <- nrow(seer)
train_idx <- sample(seq_len(n), size = floor(0.9 * n))
seer_train <- seer[train_idx, ]
seer_valid <- seer[-train_idx, ]

## -----------------------------
## 3. Tumor size vs T stage (ANOVA)
## -----------------------------
boxplot(Tumor.Size ~ T.Stage, data = seer_train,
        main = "Tumor Size by T Stage",
        ylab = "Tumor Size", xlab = "T Stage")
summary(aov(Tumor.Size ~ T.Stage, data = seer_train))

## -----------------------------
## 4. Baseline table (Table 1)
## -----------------------------
cat_vars  <- c("Race", "Marital.Status", "T.Stage", "N.Stage",
               "X6th.Stage", "differentiate", "Grade", "A.Stage",
               "Estrogen.Status", "Progesterone.Status")

cont_vars <- c("Age", "Tumor.Size",
               "Regional.Node.Examined", "Regional.Node.Positive")

vars_for_table <- c(cat_vars, cont_vars)

tab1 <- CreateTableOne(vars = vars_for_table,
                       strata = "Status",
                       data = seer_train,
                       factorVars = cat_vars)

## As data frame with N in first row
tab_df <- as.data.frame(print(tab1, showAllLevels = TRUE,
                              quote = FALSE, noSpaces = TRUE,
                              printToggle = FALSE))
N_alive <- sum(seer_train$Status == "Alive")
N_dead  <- sum(seer_train$Status == "Dead")
tab_df$N <- c(paste0("Alive = ", N_alive,
                     ", Dead = ", N_dead),
```

```
              rep("", nrow(tab_df) - 1))

tab_df  ## (export from R if needed)

## ----------------------------
## 5. Correlation among continuous vars
## ----------------------------
num_vars <- seer_train %>%
  dplyr::select(Age, Tumor.Size,
                Regional.Node.Examined, Regional.Node.Positive)

cor_matrix <- cor(num_vars, use = "complete.obs")
corrplot(cor_matrix, method = "color", addCoef.col = "black")

## ----------------------------
## 6. Cramer's V among categorical vars
## ----------------------------
cat_vars_cv <- seer_train %>%
  dplyr::select(Race, Marital.Status, T.Stage, N.Stage,
                X6th.Stage, A.Stage, differentiate, Grade)

cramer_results <- matrix(NA,
                         ncol = ncol(cat_vars_cv),
                         nrow = ncol(cat_vars_cv),
                         dimnames = list(names(cat_vars_cv),
                                         names(cat_vars_cv)))

for (i in 1:ncol(cat_vars_cv)) {
  for (j in 1:ncol(cat_vars_cv)) {
    tbl <- table(cat_vars_cv[[i]], cat_vars_cv[[j]])
    crv <- assocstats(tbl)$cramer
    cramer_results[i, j] <- crv
  }
}

cramer_results

## Keep variables after collinearity assessment
cat_vars  <- c("Race", "Marital.Status", "N.Stage", "Grade", "A.Stage",
               "Estrogen.Status", "Progesterone.Status")
cont_vars <- c("Age", "Tumor.Size",
               "Regional.Node.Examined", "Regional.Node.Positive")
```

```
## -----------------------------
## 7. Univariable Cox models
## -----------------------------
uni_vars <- c(cat_vars, cont_vars)

uni_cox <- function(var, data) {
  fml <- as.formula(paste("surv_obj ~", var))
  fit <- coxph(fml, data = data)
  tidy(fit, exponentiate = TRUE, conf.int = TRUE) %>%
    mutate(variable = var)
}

uni_results <- lapply(uni_vars, uni_cox, data = seer_train) %>%
  bind_rows() %>%
  dplyr::select(variable, term, estimate, conf.low, conf.high, p.value)

uni_summary <- uni_results %>%
  group_by(variable) %>%
  summarise(min_p = min(p.value, na.rm = TRUE), .groups = "drop") %>%
  arrange(min_p)

vars_selected <- uni_summary %>%
  filter(min_p < 0.20) %>%
  pull(variable)

vars_selected

## -----------------------------
## 8. Initial multivariable model
## -----------------------------
cox_full <- coxph(
  surv_obj ~ Age +
    Race + Marital.Status + N.Stage + Grade + A.Stage +
    Tumor.Size + Estrogen.Status + Progesterone.Status +
    Regional.Node.Examined + Regional.Node.Positive,
  data = seer_train
)
summary(cox_full)

## Remove Marital.Status and check confounding
cox_1 <- coxph(
  surv_obj ~ Age +
    Race + N.Stage + Grade + A.Stage +
```

```
    Tumor.Size + Estrogen.Status + Progesterone.Status +
    Regional.Node.Examined + Regional.Node.Positive,
  data = seer_train
)

terms_keep <- setdiff(
  names(coef(cox_full)),
  grep("^Marital.Status", names(coef(cox_full)), value = TRUE)
)

per_change_M <- data.frame(
  term       = terms_keep,
  per_change = round(
    abs((coef(cox_full)[terms_keep] - coef(cox_1)[terms_keep]) /
          coef(cox_full)[terms_keep]), 4)
)
per_change_M

## Remove A.Stage and check confounding
cox_3 <- coxph(
  surv_obj ~ Age +
    Race + Grade + Tumor.Size + N.Stage +
    Estrogen.Status + Progesterone.Status +
    Regional.Node.Examined + Regional.Node.Positive,
  data = seer_train
)

terms_keep2 <- setdiff(
  names(coef(cox_1)),
  grep("^A.Stage", names(coef(cox_1)), value = TRUE)
)

per_change_A <- data.frame(
  term       = terms_keep2,
  per_change = round(
    abs((coef(cox_1)[terms_keep2] - coef(cox_3)[terms_keep2]) /
          coef(cox_1)[terms_keep2]), 4)
)
per_change_A

## -----------------------------
## 9. Preliminary main-effects model
## -----------------------------
```

```
cox_main <- coxph(
  surv_obj ~ Age +
    Race + N.Stage + Grade +
    Tumor.Size + Estrogen.Status + Progesterone.Status +
    Regional.Node.Examined + Regional.Node.Positive,
  data = seer_train
)
summary(cox_main)

## -----------------------------
## 10. Check continuous functional form (psplines)
## -----------------------------
cox_main_spline <- coxph(
  surv_obj ~ pspline(Age) +
    Race +
    N.Stage + Grade +
    pspline(Tumor.Size) +
    Estrogen.Status + Progesterone.Status +
    pspline(Regional.Node.Examined) +
    pspline(Regional.Node.Positive),
  data = seer_train
)
summary(cox_main_spline)

par(mfrow = c(2, 2))
termplot(cox_main_spline, se = TRUE, terms = 1, ylabs = "Log Hazard")
termplot(cox_main_spline, se = TRUE, terms = 5, ylabs = "Log Hazard")
termplot(cox_main_spline, se = TRUE, terms = 8, ylabs = "Log Hazard")
termplot(cox_main_spline, se = TRUE, terms = 9, ylabs = "Log Hazard")

## -----------------------------
## 11. Interaction model and AIC step
## -----------------------------
cox_int <- coxph(
  surv_obj ~ (Age +
    Race + N.Stage + Grade + A.Stage +
    Tumor.Size + Estrogen.Status + Progesterone.Status +
    Regional.Node.Examined + Regional.Node.Positive)^2,
  data = seer_train
)

cox_step <- stepAIC(cox_int, direction = "both", trace = TRUE)
summary(cox_step)
```

```
lrtest(cox_main_spline, cox_step)  ## compare spline main-effects vs interaction

## -----------------------------
## 12. Final main-effects model
## -----------------------------
cox_final <- coxph(
  surv_obj ~ Age +
    Race + N.Stage + Grade + Tumor.Size +
    Estrogen.Status + Progesterone.Status +
    Regional.Node.Examined + Regional.Node.Positive,
  data = seer_train
)
summary(cox_final)

## -----------------------------
## 13. PH diagnostics and stratified model
## -----------------------------
test_ph_final <- cox.zph(cox_final, transform = "rank")
test_ph_final
plot(test_ph_final)

## Stratify on hormone receptor status
cox_strat <- coxph(
  surv_obj ~ Age + Race + N.Stage + Grade + Tumor.Size +
    Regional.Node.Examined + Regional.Node.Positive +
    strata(Estrogen.Status) +
    strata(Progesterone.Status),
  data = seer_train
)
summary(cox_strat)

test_ph_strat <- cox.zph(cox_strat, transform = "rank")
test_ph_strat
plot(test_ph_strat)

## -----------------------------
## 14. DFBETA influence diagnostics
## -----------------------------
dfb <- residuals(cox_strat, type = "dfbeta")
index.obs  <- 1:nrow(dfb)
coef_names <- colnames(dfb)
```

```
## On-screen plots
par(mfrow = c(3, 3))
for (j in 1:ncol(dfb)) {
  plot(index.obs, dfb[, j], type = "h",
       xlab = "Observation",
       ylab = paste("DFBETA for", coef_names[j]),
       main = coef_names[j])
  abline(h = 0, col = "red")
}

## Save combined DFBETA plots
png("dfbeta_plots.png", width = 1800, height = 1800, res = 200)
par(mfrow = c(3, 3), mar = c(4, 4, 3, 1))
for (j in 1:ncol(dfb)) {
  plot(index.obs, dfb[, j], type = "h",
       xlab = "Observation",
       ylab = paste("DFBETA for", coef_names[j]),
       main = coef_names[j])
  abline(h = 0, col = "red")
}
dev.off()

## -----------------------------
## 15. Kaplan{Meier plots for key covariates
## -----------------------------
fit_grade  <- survfit(Surv(Survival.Months, status_event) ~ Grade,
                      data = seer_train)
fit_nstage <- survfit(Surv(Survival.Months, status_event) ~ N.Stage,
                      data = seer_train)
fit_race   <- survfit(Surv(Survival.Months, status_event) ~ Race,
                      data = seer_train)
fit_astage <- survfit(Surv(Survival.Months, status_event) ~ A.Stage,
                      data = seer_train)

km_theme <- theme_minimal(base_size = 9) +
  theme(
    legend.position = "none",
    plot.title = element_text(size = 10, face = "bold"),
    axis.title = element_text(size = 8),
    axis.text  = element_text(size = 7)
  )

p_grade <- ggsurvplot(
```

```
  fit_grade, data = seer_train,
  pval = TRUE, conf.int = TRUE,
  title = "Tumor Grade", ggtheme = theme_minimal()
)$plot + km_theme

p_nstage <- ggsurvplot(
  fit_nstage, data = seer_train,
  pval = TRUE, conf.int = TRUE,
  title = "N Stage", ggtheme = theme_minimal()
)$plot + km_theme

p_race <- ggsurvplot(
  fit_race, data = seer_train,
  pval = TRUE, conf.int = TRUE,
  title = "Race", ggtheme = theme_minimal()
)$plot + km_theme

p_astage <- ggsurvplot(
  fit_astage, data = seer_train,
  pval = TRUE, conf.int = TRUE,
  title = "A Stage", ggtheme = theme_minimal()
)$plot + km_theme

km_grid <- (p_grade | p_nstage) /
           (p_race  | p_astage)

km_grid
ggsave("KM_4panel_patchwork.pdf", km_grid,
       width = 10, height = 8)
ggsave("KM_4panel_patchwork.png", km_grid,
       width = 10, height = 8, dpi = 300)

## ----------------------------
## 16. DeepSurv model and C-index comparison
## ----------------------------
set.seed(2025)

cox_formula <- Surv(Survival.Months, status_event) ~
  Age + Race + Marital.Status + N.Stage + Grade + A.Stage +
  Tumor.Size + Estrogen.Status + Progesterone.Status +
  Regional.Node.Examined + Regional.Node.Positive

## Design matrices
```

```
X_train <- model.matrix(cox_formula, data = seer_train)[, -1]
p <- ncol(X_train)

## Deep neural network structure
ds_model <- dNNmodel(
  units       = c(64, 32, 1),
  activation  = c("relu", "relu", "idu"),
  input_shape = p
)

## Scale continuous covariates for DeepSurv
seer_train_scaled <- seer_train %>%
  mutate(
    Tumor.Size             = scale(Tumor.Size),
    Age                    = scale(Age),
    Regional.Node.Examined = scale(Regional.Node.Examined),
    Regional.Node.Positive = scale(Regional.Node.Positive)
  )

## Fit DeepSurv
fit_deepsurv <- deepSurv(
  formula    = cox_formula,
  model      = ds_model,
  data       = seer_train_scaled,
  epochs     = 3000,
  batch_size = 32,
  lr_rate    = 1e-3,
  alpha      = 0.7,
  lambda     = 1,
  verbose    = 1
)

## C-index for stratified Cox model on validation data
lp_cox_valid <- predict(cox_strat, newdata = seer_valid, type = "risk")

c_index_cox <- concordance.index(
  x          = lp_cox_valid,
  surv.time  = seer_valid$Survival.Months,
  surv.event = seer_valid$status_event
)
c_index_cox$c.index   ## ~ 0.69

## C-index for DeepSurv on validation data
```

```
X_valid  <- model.matrix(cox_formula, data = seer_valid)[, -1]
pred_ds  <- predict(fit_deepsurv, X_valid, type = "risk")
risk_ds  <- as.numeric(pred_ds$risk)

c_index_ds <- concordance.index(
  x          = risk_ds,
  surv.time  = seer_valid$Survival.Months,
  surv.event = seer_valid$status_event
)
c_index_ds$c.index    ## ~ 0.66
```