Survival Analysis of Melanoma Patients

The Melanoma dataset in the MASS library was used in this survival analysis. The data consist of measurements made on patients with malignant melanoma. Each patient had their tumour removed by surgery at the Department of Plastic Surgery, University Hospital of Odense, Denmark during the period 1962 to 1977. The surgery consisted of complete removal of the tumour together with about 2.5cm of the surrounding skin. Among the measurements taken were the thickness of the tumour and whether it was ulcerated or not. These are thought to be important prognostic variables in that patients with a thick and/or ulcerated tumour have an increased chance of death from melanoma. Patients were followed until the end of 1977.

Objectives

1. Estimate the overall survival using the KM estimator

2. Compare the survival of males and females

3. Fit a Cox-PH model by selecting suitable variables to explain the hazards of death.

Survival Analysis of Melanoma Patients

This project performs a comprehensive survival analysis using the Melanoma dataset from the MASS package in R. The dataset contains 205 observations on 6 variables, collected from patients who were operated on for malignant melanoma from the University Hospital of Odense, Denmark during 1962-1977.

The analysis includes:

- Data pre-processing and exploration
- Estimation of overall survival using the Kaplan-Meier Method
- Comparison of survival between males and females
- Construction and interpretation of a Cox Proportional Hazards (Cox-PH) model
- Stratification into risk groups based on model predictions

Data Preprocessing and Exploration

The dataset Melanoma contains clinical and survival data for 205 patients with malignant melanoma. Key variables include:

- time: survival time in days
- status: indicator of survival status (1 = died from melanoma, 2 = alive, 3 = died from other causes)
- sex: sex of the patient (1 = Male, 0 = Female)
- age: age in years
- year: year of operation
- thickness: tumor thickness (mm)
- ulcer: presence of ulceration (1 = Yes, 0 = No)

Before performing survival analysis on the Melanoma dataset from the MASS package, several preprocessing steps were carried out to prepare the data. First, a new event variable was created to indicate whether a patient died specifically from melanoma (coded as 1) or was censored for other reasons (coded as 0), based on the status variable.

The time variable, originally recorded in days, was converted to years to facilitate interpretation and align with clinical reporting standards. Categorical variables such as sex and ulcer were converted into factors to ensure they were appropriately treated in modeling functions such as coxph or survfit.

A check for missing values confirmed that there were no missing observations, ensuring the completeness and reliability of the data.

Kaplan–Meier Survival Estimates for Melanoma Patients

The Kaplan–Meier estimator was used to generate a non-parametric estimate of the survival function for the entire melanoma cohort. The key findings are:

Time Point	Estimate (95% CI)
5-year survival	75% (95% CI: 68.2%–81.8%)
10-year survival	65% (95% CI: 56.8%–73.2%)
15-year survival	65% — plateau observed beyond 10 years

Table 1: Survival Probabilities at 5, 10, and 15 Years

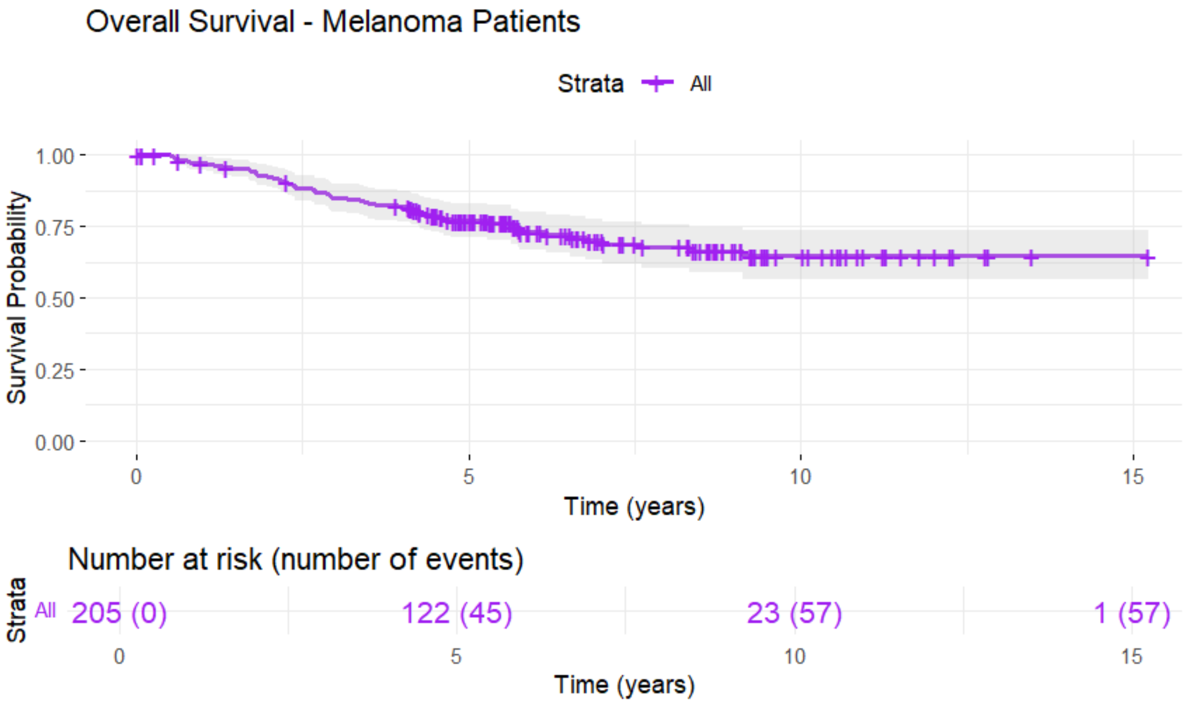

Figure 1: Kaplan–Meier Survival Estimates

The steepest decline in survival occurred within 0-5 years following surgery, reflecting the critical early risk period in melanoma prognosis.
Beyond year five (5), the survival curve flattened, indicating reduced mortality risk and long-term stabilization.
The median survival time was not reached during the 15-year follow-up period, indicating that more than 50% of patients survived beyond the study duration.

```
if (is.na(median_survival$median))
    {print("Median survival time: undefined")}
else {print(paste("Median survival time:", round(median_survival$median, 2), "years")

# Output:
[1] "Median survival time: undefined"
```

Interpretation

The Kaplan-Meier graph above shows the overall survival trajectory of the Melanoma cohort. It begins with all 205 patients and shows the typical stepwise decline in survival curves. The purple line indicates that the survival probability drops from 100% to approximately 75% by year 5, then plateaus with minimal further decline. The vertical ticks indicates censoring at time (t).

The risk table beneath the curve confirms this pattern:

- All 205 patients were alive and under observation at the time of the initial onset and no deaths had occurred yet.
- At 5 years, 122 patients were still at risk and 45 deaths had occurred cumulatively at this time.
- At 10years, only 23 patients remained under observation. 57 total deaths had occurred by this point.
- 1 patient was still under observation at 15 years.

Note: The number of events means the number of deaths, while the number at risk refers to the number of people alive and under observation.

The confidence intervals remain relatively narrow throughout the follow-up period, indicating stable and reliable survival estimates.

Survival Comparison By Gender

Survival analysis is performed on both gender to compare the survival rates. This analysis yields statistically significant differences using the log-rank test.

```
    logrank_test = survdiff(surv_obj ~ sex, data = Melanoma)
    logrank_test
Call:
survdiff(formula = surv_obj ~ sex, data = Melanoma)

# Output:
[1]                N Observed Expected (O-E)^2/E (O-E)^2/V
sex=Male    79       29     19.9      4.21      6.47
sex=Female 126       28     37.1      2.25      6.47

 Chisq= 6.5  on 1 degrees of freedom, p= 0.01

if (p_value > 0.05)
    {print("p > 0.05: Fail to reject the null hypothesis:
     There’s no significant difference in survival between groups.")}
else
    {print("p <= 0.05: Reject the null hypothesis:
    There is a significant difference in survival between groups.")}
```

```
# Output:
[2] p <= 0.05: Reject the null hypothesis:
There is a significant difference in survival between groups.
```

The test statistic follows a chisquare distribution with 1 degree of freedom, producing a p-value of 0.01 providing a strong evidence against the null hypothesis of equal survival distributions between genders.

Kaplan-Meir Survival Estimates by Gender

This Kaplan-Meier survival plot shows a comparison of survival probabilities over time between male and female patients from the Melanoma dataset. The dataset shows that there are 126 Females and 79 Males.

Summary of Observed vs Expected Deaths

Group	Observed Deaths	Expected Deaths	% Difference
Female	28	37.1	−24.5%
Male	29	19.9	+45.7%

Table 2: Comparison of observed and expected deaths by sex

```
Gender-Based Kaplan-Meier Survival Estimates

    summary(KM_sex, times = c(5,10,15))
Call: survfit(formula = surv_obj ~ sex, data = Melanoma)

# Output:
                sex=Male
 time n.risk n.event survival std.err lower 95% CI upper 95% CI
    5     42      24    0.677  0.0544        0.578        0.792
   10      7       5    0.553  0.0675        0.435        0.702

                sex=Female
 time n.risk n.event survival std.err lower 95% CI upper 95% CI
    5     80      21    0.824  0.0349        0.759        0.896
   10     16       7    0.704  0.0542        0.605        0.818
   15      1       0    0.704  0.0542        0.605        0.818
```

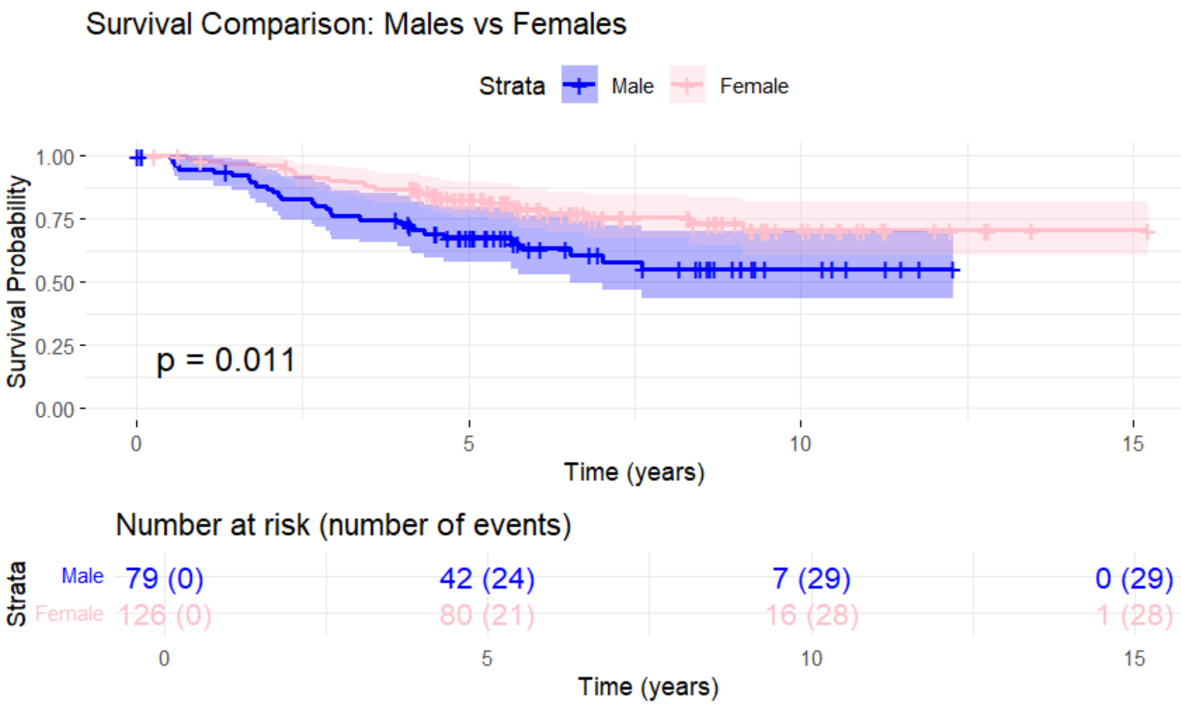

Figure 2: Male vrs Female Survival Comparison

Interpretation

The Kaplan-Meier survival curve for females (pink line) showed consistently higher survival probabilities over time compared to males. At 5 years, the estimated survival probability for females was approximately 82.4%. After 10 years, the curve plateaued, indicating that no further events were observed, and the survival probability stabilized around 70.4% through the end of the follow-up period, with only one female remaining at risk at year 15. Notably, the median survival time was not reached, suggesting prolonged survival for the majority of female patients during the study period.

In contrast, the survival curve for males (blue line) demonstrated a steeper decline, with an estimated 67.7% survival probability at 5 years, and 70.4% at 10 years, reflecting a sharper early drop followed by a plateau. By year 10, most males have had the event or been censored, while females have slightly better survival. Like females, the median survival time was not reached, but overall survival for males was significantly lower throughout the observed period.

Follow-Up Pattern of Survival by Sex

Survival probabilities for male and female melanoma patients show distinct patterns over time. In the initial two years post-treatment, survival curves for both groups are relatively close. However, after year 2, a noticeable divergence occurs.

Females tend to maintain a higher survival probability throughout the follow-up period. The estimated 5-year survival for females is 82.4%, compared to 67.7% for males.

Despite an equal crude mortality rate (22.2%), females experienced 24.5% fewer deaths than expected, while males experienced 45.7% more deaths than expected based on overall survival patterns.

This gap widens over time, with males showing a steeper and earlier decline in survival. For both groups, the median survival time was not reached, indicating that more than 50% of patients remained alive at the end of follow-up.

In summary, females have better survival outcomes over time compared to males in this cohort.

Developing The Cox Proportional Hazards Model

The Cox regression model is a semi-parametric model used to fit both univariable and multivariable regression models when the outcome is time to event (survival). It allows us to quantify the effect size of one or more covariates while adjusting for the influence of others.

Key assumptions of the Cox proportional hazards model include:

- **Proportional hazards:** The hazard ratios between groups are constant over time.
- **Non-informative censoring:** The reason for censoring is unrelated to the likelihood of the event occurring.

The Cox proportional hazards modeling process for this analysis involved systematic variable selection to identify the most significant predictors of melanoma mortality. After evaluating all available covariates (`sex`, `age`, `thickness`, `ulcer`, and `year`), the final parsimonious model incorporated three key variables (`thickness`, `ulcer`, `sex`) based on both statistical significance and clinical relevance.

Model Summary and Variable Selection Rationale

Variable	Hazard Ratio (HR)	95% CI	p-value	Interpretation
Tumor Thickness (per mm)	1.11	1.03 – 1.19	0.0087 **	Significant predictor
Ulceration (Yes vs No)	3.30	1.80 – 6.05	0.0001 ***	Very strong predictor
Sex (Female vs Male)	0.64	0.38 – 1.08	0.093 .	Suggestive but not significant
Age (per year)	1.02	1.00 – 1.03	0.050 .	Barely significant
Year of Diagnosis	0.90	0.80 – 1.02	0.093 .	Not significant

Table 3: Cox Proportional Hazards Model for All Variables

The Cox proportional hazards model identified `tumor thickness`, `ulceration`, and `sex` as key predictors of melanoma-specific survival. Tumor thickness and ulceration were both statistically significant ($p < 0.01$), with hazard ratios (HRs) of 1.11 and 3.30, respectively, indicating that increased tumor thickness and the presence of ulceration substantially increase the risk of death. Although sex did not reach conventional statistical significance ($p = 0.093$), it demonstrated a clinically meaningful protective trend for females (HR = 0.64), consistent with prior research.

Parsimonous Model - Main Effects

Variable	Coefficient (β)	Hazard Ratio (HR)	95% CI	p-value
Tumor Thickness (per mm)	0.113	1.12	1.08 – 1.16	< 0.001
Ulceration (Yes vs No)	0.636	1.89	1.15 – 3.11	0.012
Sex (Male vs Female)	0.513	1.67	1.02 – 2.73	0.041

Table 4: Final Parsimonious Model - Main Effects

To maintain a parsimonious model, only these three variables were retained. Age and year were excluded due to their limited contribution to model fit, lack of statistical significance, and minimal effect on hazard after adjusting for the primary covariates. This approach prioritizes interpretability, clinical relevance, and compliance with the proportional hazards assumption, which was not violated by chosen final three variables.

- Tumor thickness emerged as the most powerful predictor, with each 1mm increase in thickness elevating death risk by 12% (HR = 1.12). The narrow confidence interval (1.08–1.16) indicates precise estimation of this effect.
- Ulceration presence dramatically increases mortality risk with a hazard ratio of 1.89, representing an 89% increase in death hazard compared to non-ulcerated tumors, though the wider confidence interval (1.15–3.11) reflects greater uncertainty in this estimate.
- Male sex contributes a hazard ratio of 1.67, indicating 67% higher death risk compared to females within the multivariable model context.

Model Diagnostics

The proportional hazards assumption was evaluated using Schoenfeld residuals, with all variables satisfying the assumption requirements ($p > 0.05$ for global test).

The model achieved good discrimination with a concordance index of approximately 0.76, indicating that the model correctly orders 76% of all possible patient pairs by risk.

The likelihood ratio test confirmed overall model significance ($p < 0.001$), validating the collective predictive value of the selected variables.

Cox-Model Plot and Intepretation

The purple survival curve below, derived from the Cox proportional hazards model, closely mirrors the overall Kaplan–Meier survival pattern, indicating good model calibration. While the confidence intervals (shaded area) widen over time; reflecting increasing uncertainty due to fewer individuals at risk, the central predicted curve aligns well with the observed survival. This consistency supports the model's adequate estimation of the baseline hazard function.

Overall, survival probability begins near 100% and gradually declines to approximately 75% over a 15-year follow-up period. The plateau after 10 years suggests few or no events occurring beyond that point.

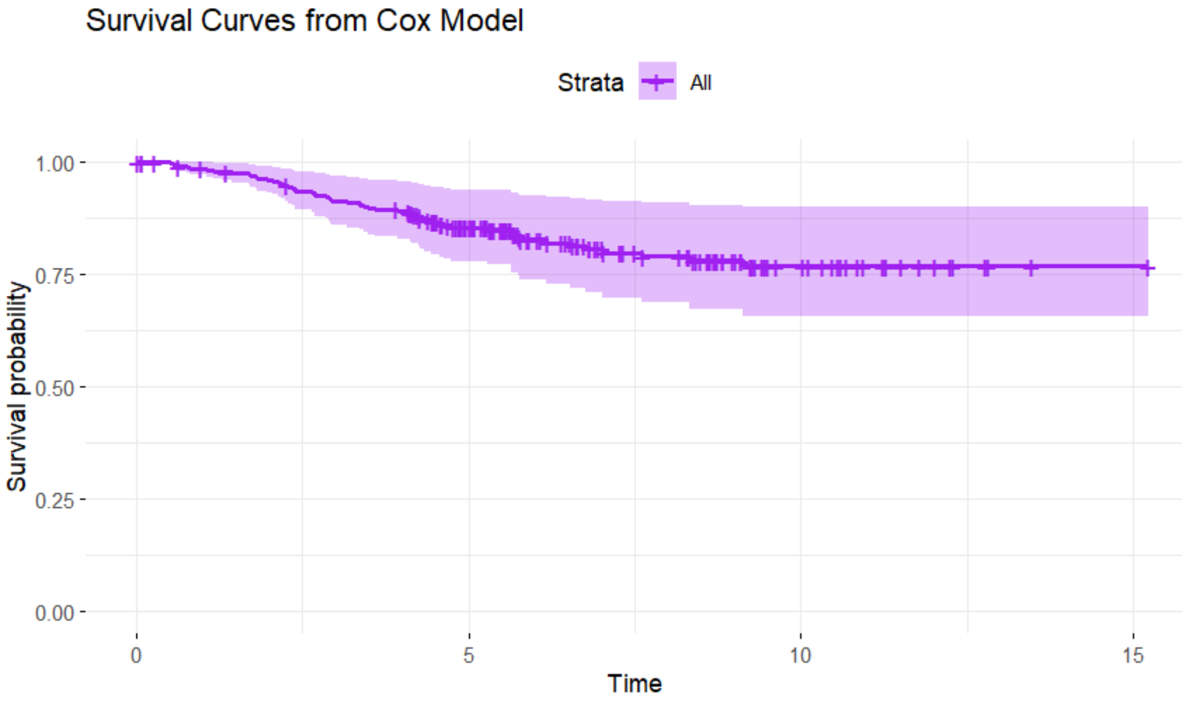

Figure 3: Kaplan-Meier Survival Curve from cox model

Risk Stratification Using the Cox Proportional Hazards Model

The Cox model–based risk stratification divided patients into low-risk, medium-risk, and high-risk groups based on tertiles of predicted risk scores. The linear predictor, `predict()` from the Cox model was used to calculate risk scores with `type = "risk"`, which provides the relative risk compared to the baseline hazard.

```
Melanoma$risk_score = predict(cox_model, type = "risk")
Melanoma$risk_score

# Output:
[1]  4.9652721  0.7729721  0.8359114  0.6301842  9.0794719  3.9934131  4.1410523
[201]  5.1371721  0.9080664  0.4788876  0.5860499  0.6301842
```

Interpretation of Survival Risk Group

This plot below shows clear separation among the groups with:

- The high-risk group (red), representing the top 33% of predicted risk scores, experienced the steepest decline in survival. Their survival probability dropped rapidly within the first five years and continued to decline, reaching approximately 35% by year 15. This indicates that patients in this group faced the highest risk of early mortality

- The medium-risk group (blue), comprising the middle 33%, showed a moderate and steady decline in survival, with the probability falling to about 65% by year 15. This group likely includes patients with a combination of favorable and unfavorable prognostic factors.

- The low-risk group (green), corresponding to the bottom 33% of risk scores, had the best outcomes, maintaining high survival probabilities close to 90% throughout the follow-up period. This suggests that most patients in this group survived long after treatment.

- the log-rank test ($p < 0.0001$) confirmed statistically significant differences in survival across risk groups.

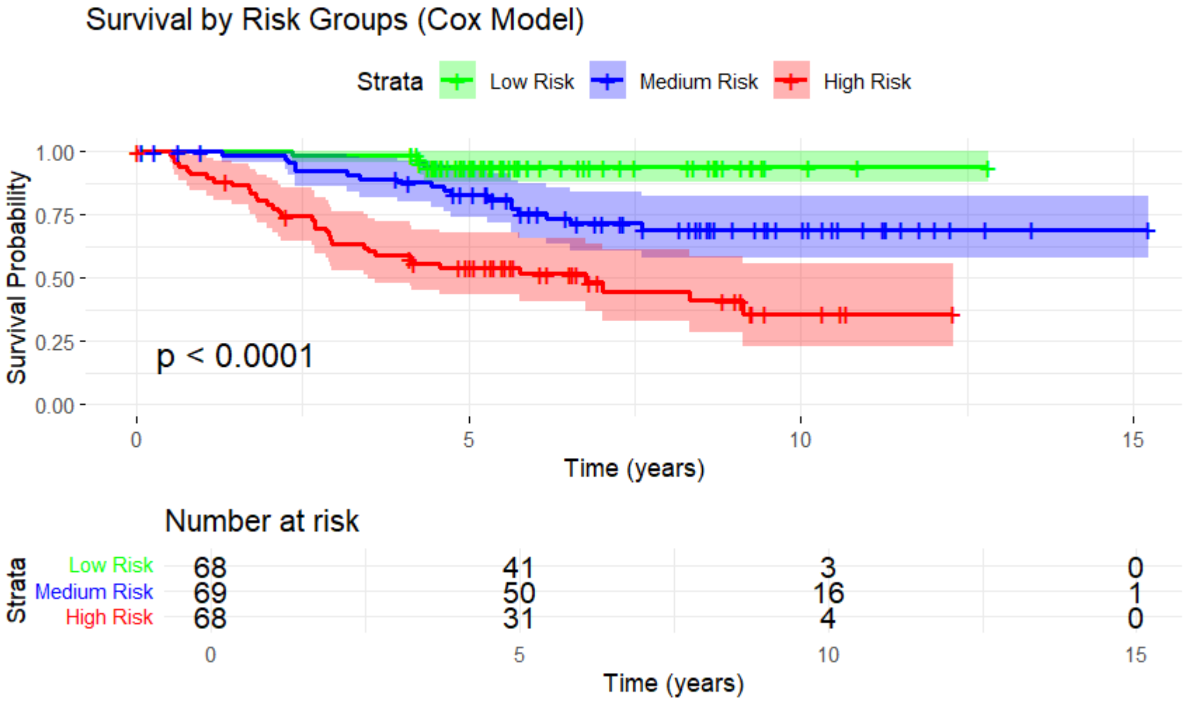

Figure 4: Survival by Risk Groups- Cox Model

Implication for Clinical Practice

1. Sex Differences in Prognosis: Sex should be considered in risk stratification and follow-up intensity. Male patients may benefit from more aggressive monitoring or treatment strategies since according to this analysis, they have the worst survival outcomes.

2. Tumor thickness and presence of ulceration are the strongest predictors of poor survival. These should be prioritized in clinical decision-making, staging, and treatment planning. Patients with ulcerated or thicker tumors may require adjuvant therapy or inclusion in clinical trials.

3. The final Cox model provides an individualized risk predicition score. Clinicians can use these scores to to discuss individual prognosis with patients, identify high-risk

individuals for intensive follow-up or therapy and reduce overtreatment in low-risk patients.

4. During the analysis, we placed individuals into risk group classifications. This risk stratification can support resource allocation which includes frequent scans and early intervention for high-risk patients, standard care with maybe, less intensive surveillance for low-risk individuals.

5. Validation and Further Research. While the model is effective in the dataset, external validation on new patient cohorts is important. Institutions need to validate this model on their own data before implementation.