Why Choose Complete Case Analysis?

- **Guarantees Data Integrity**: in a risk prediction setting, our target population is the population with the complete set of features observed. It is the population in which eventually our model will be implemented.
- **Avoids Assumptions**: Does not depend on assumptions about why data is missing.
- **Preserves True Variance**: Imputed data mimics existing patterns without adding new variability, risking the reinforcement of trends.
- **Less than 5%** of the data was missing.

Right Censoring

Summary Statistics for Time-to-Event Data (in Years)

Min (years)	Q1 (years)	Median (years)	Mean (years)	Q3 (years)	Max (years)	SD (years)
0	2.89117	6.310746	7.787924	10.88022	36.72827	6.619472

- **Cases**: A case (`disease_status == 1`) is assigned `1` if their event occurred within the respective timeframe (2 years or 5 years). If the event happened after the timeframe, they are assigned `0` but remain tracked in the dataset.
- **Controls**: A control (`disease_status == 0`) is always assigned `0` for both timeframes (2 years and 5 years), as they do not experience the event by definition.

- Yes, the outcome is treated as a binary variable (Yes/No) for each time horizon, such as 2 years or 5 years.
- If a participant's person-time is censored before reaching the 5-year mark (e.g., censored at 3 years), they are considered censored for the 5-year outcome. This means they contribute to the analysis up until the point of censoring (3 years), but their status after 3 years is unknown.
- For the 5-year outcome, such a participant would be excluded from contributing either a "Yes" or "No" beyond their censoring point.
- This ensures the integrity of the survival analysis by only including data up to the point where it is observed or can be validly interpreted!!!

Harrell's C Statistic

Mathematical Definition of Harrell's C Statistic:

Harrell's C statistic measures the concordance between predicted risk scores from a survival model and observed survival times. It is formally defined as:

$$C = \frac{\text{Number of Concordant Pairs} + 0.5 \times \text{Number of Tied Pairs}}{\text{Number of Comparable Pairs}}$$

Key Components:

1. **Concordant Pairs**:
   - A pair (i, j) is concordant if:

     $$\text{Predicted Risk}_i > \text{Predicted Risk}_j \quad \text{and} \quad \text{Time}_i < \text{Time}_j$$

     where Time_i and Time_j are observed survival times, and Predicted Risk_i and Predicted Risk_j are the predicted risk scores for individuals i and j.
2. **Tied Pairs**:
   - A pair (i, j) is tied if:

     $$\text{Predicted Risk}_i = \text{Predicted Risk}_j$$
   - In this case, 0.5 is added to the concordance count.
3. **Comparable Pairs**:
   - A pair (i, j) is comparable if the survival times for i and j can be compared, meaning neither is censored in a way that prevents the comparison.
4. **Censored Data Handling**:
   - For censored data, Harrell's C uses pairs where at least one individual has an observed event time that can be ranked relative to the other.

PH Assumption

1. Schoenfeld Residuals Test: Tests whether the Schoenfeld residuals are independent of time.

- **How to Use**:
  - After fitting a Cox model, check for a significant correlation between residuals and time.
- **Key Tool**: cox.zph() function in R (from the survival package).
- **Interpretation**:
  - **p > 0.05**: No evidence of a violation of the PH assumption.
  - **p ≤ 0.05**: Suggests the PH assumption may be violated for the corresponding covariate.

2. Graphical Methods: Schoenfeld Residual Plots:

- Plot Schoenfeld residuals against time for each covariate.
- If the residuals show no clear trend over time, the PH assumption holds.

Baseline Hazard Function

The **Nelson-Aalen estimator** was used to estimate the baseline hazard, as it aligns with the non-parametric nature of the Cox model and is the standard approach for deriving the baseline survival probability.

Alternative and difference:

Kaplan-Meier Estimator:

- Estimates the **survival function**, which represents the probability of surviving past time t.
- Focuses on the survival probability directly rather than the hazard.

Nelson-Aalen Estimator:

- Estimates the **cumulative hazard function** H(t), which represents the accumulated risk of experiencing the event up to time t.
- Focuses on modeling the cumulative hazard as a step function.