Survival Analysis — Comprehensive Reference

I. Fundamentals of Survival Data

A. Time-to-Event Data Structure

- **Survival time** T: Random variable $T \geq 0$ measuring time from start (e.g., treatment, diagnosis) to event (e.g., death, relapse, failure).
- **Three components**: (1) Starting point (when clock starts), (2) Endpoint (when clock stops), (3) Time unit (days, months, years).
- **Event**: Can be death, disease, recurrence, recovery, onset of symptoms, equipment failure. Often called "failure".

B. Censoring Mechanisms

- **Right-censoring** (most common): Event time > observed time. Causes:
  - End of study (administrative censoring)
  - Loss to follow-up
  - Withdrawal from study
  - Competing events

  Notation: Observe (Y, δ) where $Y = \min(T, C)$, $\delta = I(T \leq C)$. $\delta = 1$ (event), $\delta = 0$ (censored).
- **Left-censoring**: Event occurred before observation started. Example: child already knew task at study start.
- **Interval-censoring**: Event time known to lie in $(L, R]$. Example: disease detected between visits.
- **Key assumption**: Censoring is *non-informative* (independent of failure time). Violation leads to bias.

C. Core Functions

Survival Function: $S(t) = P(T > t) = 1 - F(t)$

- Properties: $S(0) = 1$, $S(\infty) = 0$, non-increasing.
- Interpretation: Probability of surviving beyond time t.
- 5-year survival rate: $S(5)$.

Probability Density Function (PDF): $f(t) = \dfrac{dF(t)}{dt} = -\dfrac{dS(t)}{dt}$

Hazard Function: $h(t) = \lim_{\Delta t \to 0} \dfrac{P(t \leq T < t + \Delta t \mid T \geq t)}{\Delta t} = \dfrac{f(t)}{S(t)}$

- Instantaneous failure rate per unit time among survivors at t.
- Units: 1/time (e.g., per year).
- Not a probability (can exceed 1).
- For large population: $h(t)\Delta t \approx \frac{\#\text{ deaths in } (t, t+\Delta t)}{\#\text{ alive at } t}$.

Cumulative Hazard: $H(t) = \int_0^t h(u)\, du$

Key Relationships:

$$S(t) = \exp[-H(t)] = \exp\left[-\int_0^t h(u)\, du\right]$$

$$h(t) = -\frac{d \log S(t)}{dt}$$

$$f(t) = h(t)S(t)$$

D. Comparing Survival Data

- Compare entire *functions* $S(t)$ or $h(t)$, not just means.
- Median survival often more robust than mean (especially with censoring).

II. Nonparametric Methods

A. Kaplan-Meier (KM) Estimator

Formula: At distinct event times $t_1 < t_2 < \cdots < t_m$:

$$\hat{S}(t) = \prod_{t_i \leq t} \left(1 - \frac{d_i}{n_i}\right)$$

where $d_i = \#$ events at t_i, $n_i = \#$ at risk just before t_i.

Properties:

- Step function (jumps only at event times).
- Censoring at t_j: subject removed from risk set at t_j^+.
- If largest time is censored, $\hat{S}(t)$ doesn't reach 0.
- $\hat{S}(0^-) = 1$.

Variance — Greenwood's Formula:

$$\widehat{\text{Var}}[\hat{S}(t)] = \hat{S}(t)^2 \sum_{t_i \leq t} \frac{d_i}{n_i(n_i - d_i)}$$

Confidence Intervals:

Plain CI (can go outside $[0, 1]$*):*

$$\hat{S}(t) \pm z_{\alpha/2} \sqrt{\widehat{\text{Var}}[\hat{S}(t)]}$$

Log-log transformation (preferred):

$$\sigma^2 = \frac{1}{[\log \hat{S}(t)]^2} \sum_{t_i \leq t} \frac{d_i}{n_i(n_i - d_i)}$$

$$C_u = \log[-\log \hat{S}(t)] + z_{\alpha/2}\sigma$$

$$C_l = \log[-\log \hat{S}(t)] - z_{\alpha/2}\sigma$$

$$95\%\text{ CI} = \left(\exp(-e^{C_u}), \exp(-e^{C_l})\right)$$

Guarantees CI $\in (0, 1)$. Better coverage for extreme $S(t)$.

B. Nelson-Aalen (NA) Estimator

Cumulative Hazard:

$$\hat{H}(t) = \sum_{t_i \leq t} \frac{d_i}{n_i}$$

Survival Function:

$$\tilde{S}(t) = \exp[-\hat{H}(t)]$$

- Asymptotically equivalent to KM.
- Useful for direct cumulative hazard estimation.
- Variance: $\widehat{\text{Var}}[\hat{H}(t)] = \sum_{t_i \leq t} \frac{d_i}{n_i^2}$.

C. Quantile Estimation

General pth quantile:

$$\hat{t}_p = \min\{t_j : \hat{S}(t_j) < 1 - p\}$$

Special case: if $\hat{S}(t_j) = 1 - p$ exactly, use $\hat{t}_p = (t_j + t_{j+1})/2$.

Common quantiles:

- Median: $\hat{t}_{0.5} = \min\{t_j : \hat{S}(t_j) < 0.5\}$.
- First quartile: $\hat{t}_{0.25} = \min\{t_j : \hat{S}(t_j) < 0.75\}$.
- Third quartile: $\hat{t}_{0.75} = \min\{t_j : \hat{S}(t_j) < 0.25\}$.

Confidence Interval (Brookmeyer-Crowley): Set of all t satisfying

$$\frac{\log[-\log \hat{S}(t)] - \log[-\log(1-p)]}{\sqrt{\widehat{\text{Var}}(\log[-\log \hat{S}(t)])}} \in [-z_{\alpha/2}, z_{\alpha/2}]$$

Limitation: Only estimate quantiles within observed range of $\hat{S}(t)$. If last observation is censored and $\hat{S}(t_{\max}) > 0.1$, cannot estimate 90th percentile.

D. Mean Survival Time

Unrestricted mean:

$$\hat{\mu} = \int_0^\infty \hat{S}(t)\, dt = \sum_{i=1}^{m} \hat{S}(t_i)(t_{i+1} - t_i)$$

Issue: If largest time is censored, $\hat{\mu}$ underestimates true mean.

Restricted Mean Survival Time (RMST):

$$\hat{\mu}(\tau) = \int_0^\tau \hat{S}(t)\, dt$$

- Choose τ as maximum follow-up or clinically relevant time.
- More robust with heavy censoring.
- Compare $\Delta\hat{\mu}(\tau)$ between groups.
- Variance: $\widehat{\text{Var}}[\hat{\mu}(\tau)] = \sum_{t_i \leq \tau} \left[\int_{t_i}^\tau \hat{S}(u)\, du\right]^2 \frac{d_i}{n_i(n_i - d_i)}$.

III. Comparing Survival Curves

A. Hypothesis Testing Framework

Test $H_0 : S_1(t) = S_0(t)$ for all t vs. $H_a : S_1(t) \neq S_0(t)$ for some t.

General Weighted Test Statistic:

At each event time t_j, construct 2×2 table:

Group	Events	At Risk	Expected
1	d_{1j}	n_{1j}	$e_{1j} = \frac{n_{1j} d_j}{n_j}$
0	d_{0j}	n_{0j}	$e_{0j} = \frac{n_{0j} d_j}{n_j}$
Total	d_j	n_j	d_j

Variance:

$$v_{1j} = \frac{n_{1j} n_{0j} d_j (n_j - d_j)}{n_j^2 (n_j - 1)}$$

Weighted statistic:

$$Z = \frac{\sum_j w_j (d_{1j} - e_{1j})}{\sqrt{\sum_j w_j^2 v_{1j}}} \sim N(0,1)$$

or $\chi^2 = Z^2 \sim \chi^2_1$.

B. Common Weight Choices

1. **Log-rank test:** $w_j = 1$
   - Equal weight to all times.
   - Most powerful under proportional hazards (PH).
   - Default in most software.
   - $\chi^2_{\text{LR}} = \frac{[\sum_j (d_{1j} - e_{1j})]^2}{\sum_j v_{1j}}$.
2. **Wilcoxon (Gehan-Breslow):** $w_j = n_j$
   - Weight by # at risk.
   - Emphasizes early differences.
   - More powerful when hazards cross or differ early.
3. **Tarone-Ware:** $w_j = \sqrt{n_j}$
   - Compromise between log-rank and Wilcoxon.
4. **Peto-Peto, Fleming-Harrington:** $w_j = \tilde{S}(t_{j-1})$ or $w_j = \tilde{S}(t_{j-1})^p [1 - \tilde{S}(t_{j-1})]^q$
   - Flexible family; choose p, q to emphasize early, late, or middle differences.

C. Multiple Group Comparisons ($K > 2$)

- Test H_0: all K survival curves equal.
- χ^2 statistic with $K - 1$ df.
- **Pairwise comparisons**: use Bonferroni adjustment. For $m = \binom{K}{2}$ pairs, reject at α/m.
- SAS: `strata group / adjust=bon;`

D. Stratified Tests

Control for confounders (e.g., age, sex):

- Within each stratum s, compute $(O_{1s} - E_{1s})$ and V_{1s}.
- Pool: $Z = \frac{\sum_s (O_{1s} - E_{1s})}{\sqrt{\sum_s V_{1s}}}$.
- Assumes common treatment effect across strata.

IV. Hazard Function & Proportional Hazards

A. Hazard Interpretation

- $h(t) = 0$: no risk at t; $S(t)$ flat.
- Large $h(t)$: rapid decline in $S(t)$.
- $h(t)$ can be constant (exponential), increasing (Weibull $\beta > 1$), decreasing (Weibull $\beta < 1$), or non-monotonic (log-normal, log-logistic).

B. Proportional Hazards (PH) Assumption

For two groups with hazards $h_1(t)$ and $h_0(t)$:

$$\frac{h_1(t)}{h_0(t)} = \text{HR} = \text{constant} \quad \forall t$$

Implications:

- $h_1(t) = \text{HR} \cdot h_0(t)$
- $H_1(t) = \text{HR} \cdot H_0(t)$
- $S_1(t) = [S_0(t)]^{\text{HR}}$
- HR is instantaneous relative risk, constant over time.

Checking PH graphically:

- Plot $\log[-\log \hat{S}(t)]$ vs. $\log t$ (or t). Under PH, curves should be parallel.
- Plot $\log \hat{H}(t)$ vs. $\log t$ or t. Should be parallel under PH.

V. Cox Proportional Hazards Model

A. Model Specification

Univariable:

$$h(t, x, \beta) = h_0(t) \exp(\beta x)$$

Multivariable:

$$h(t, \mathbf{x}, \boldsymbol{\beta}) = h_0(t) \exp(\beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p)$$

or

$$\log \left[\frac{h(t, \mathbf{x})}{h_0(t)} \right] = \boldsymbol{\beta}^\top \mathbf{x}$$

Key features:

- **Semi-parametric:** $h_0(t)$ unspecified (nonparametric baseline).
- **No intercept:** absorbed into $h_0(t)$.
- **PH:** HR independent of t.

Baseline hazard: $h_0(t) = h(t, \mathbf{x} = \mathbf{0})$. Baseline survival: $S_0(t) = \exp[-H_0(t)]$ where $H_0(t) = \int_0^t h_0(u)\, du$.

Individual survival:

$$S(t, \mathbf{x}) = [S_0(t)]^{\exp(\boldsymbol{\beta}^\top \mathbf{x})}$$

B. Hazard Ratio (HR)

Definition:

$$\text{HR}(\mathbf{x}_1 : \mathbf{x}_0) = \frac{h(t, \mathbf{x}_1)}{h(t, \mathbf{x}_0)} = \exp[(\mathbf{x}_1 - \mathbf{x}_0)^\top \boldsymbol{\beta}]$$

Interpretation by covariate type:

1. **Binary x (0/1):** $\text{HR} = e^\beta$
   - $\beta > 0$: $x = 1$ has higher hazard (worse survival).
   - $\beta < 0$: $x = 1$ has lower hazard (better survival).
   - Example: $\hat{\beta} = 0.555 \Rightarrow \widehat{\text{HR}} = 1.742$. "Experimental group has 1.742 times the death rate of control (74.2% increase)."
   - Example: $\hat{\beta} = -0.684 \Rightarrow \widehat{\text{HR}} = 0.505$. "Experimental group has 0.505 times the death rate of control (49.5% reduction or 50.5% of control rate)."
2. **Continuous x (per 1-unit):** $\text{HR}(x + 1 : x) = e^\beta$
   - Often report HR for clinically meaningful change (e.g., 5-year age increase).
   - $\text{HR}(x + k : x) = e^{k\beta}$.
   - Example: $\hat{\beta}_{\text{age}} = 0.046$. For 5-year increase: $\widehat{\text{HR}} = e^{5 \times 0.046} = 1.259$. "Death rate increases 25.9% per 5-year age increase."
3. **Categorical x (reference cell coding):**
   - Create $K - 1$ dummies for K levels. Example: 4 age groups $\Rightarrow$ 3 dummies.
   - HR(level j : ref) $= e^{\beta_j}$.
   - HR(level j : level k) $= e^{\beta_j - \beta_k}$.
   - Variance: $\widehat{\text{Var}}(\hat{\beta}_j - \hat{\beta}_k) = \widehat{\text{Var}}(\hat{\beta}_j) + \widehat{\text{Var}}(\hat{\beta}_k) - 2\widehat{\text{Cov}}(\hat{\beta}_j, \hat{\beta}_k)$.
   - 95% CI: $\exp\left[(\hat{\beta}_j - \hat{\beta}_k) \pm z_{\alpha/2} \sqrt{\widehat{\text{Var}}(\hat{\beta}_j - \hat{\beta}_k)}\right]$.

C. Estimation (Partial Likelihood)

No ties: Cox's partial likelihood:

$$L_p(\boldsymbol{\beta}) = \prod_{j=1}^{m} \frac{\exp(\boldsymbol{\beta}^\top \mathbf{x}_j)}{\sum_{k \in R(t_j)} \exp(\boldsymbol{\beta}^\top \mathbf{x}_k)}$$

where $m = \#$ events, $R(t_j) =$ risk set at t_j.

Log partial likelihood:

$$\ell_p(\boldsymbol{\beta}) = \sum_{j=1}^{m} \left[\boldsymbol{\beta}^\top \mathbf{x}_j - \log \sum_{k \in R(t_j)} \exp(\boldsymbol{\beta}^\top \mathbf{x}_k) \right]$$

MLE: Solve $\frac{\partial \ell_p}{\partial \boldsymbol{\beta}} = \mathbf{0}$ (Newton-Raphson).

Variance: $\widehat{\text{Var}}(\hat{\boldsymbol{\beta}}) = \left[-\frac{\partial^2 \ell_p}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^\top} \Big|_{\hat{\boldsymbol{\beta}}} \right]^{-1} = I(\hat{\boldsymbol{\beta}})^{-1}$ (observed information).

Handling ties: When multiple events at t_j:

- **BRESLOW (fast, default):** Approximates exact likelihood. Less accurate with many ties.
- **EFRON (recommended):** Better approximation, moderate computation. Use this in course examples.
- **EXACT/DISCRETE:** Computationally intensive but exact. Use for small samples or many ties.

D. Inference

Single coefficient β_j:

Wald test: $z = \frac{\hat{\beta}_j}{\widehat{\text{SE}}(\hat{\beta}_j)} \sim N(0,1)$ or $\chi^2_W = z^2 \sim \chi^2_1$.

95% CI for β_j: $\hat{\beta}_j \pm z_{\alpha/2} \widehat{\text{SE}}(\hat{\beta}_j)$.

95% CI for HR: $\exp\left[\hat{\beta}_j \pm z_{\alpha/2} \widehat{\text{SE}}(\hat{\beta}_j)\right]$.

Multiple coefficients (overall test):

Test $H_0 : \beta_1 = \cdots = \beta_p = 0$ vs. H_a: not all zero.

1. **Likelihood Ratio (LR):** $G = -2[\ell_p(\mathbf{0}) - \ell_p(\hat{\boldsymbol{\beta}})] \sim \chi^2_p$.
   - Most reliable.
   - For nested models: $G = -2[\ell_p(\hat{\boldsymbol{\beta}}_{\text{small}}) - \ell_p(\hat{\boldsymbol{\beta}}_{\text{large}})] \sim \chi^2_{\Delta \text{df}}$.
2. **Wald:** $\hat{\boldsymbol{\beta}}^\top [\widehat{\text{Var}}(\hat{\boldsymbol{\beta}})]^{-1} \hat{\boldsymbol{\beta}} \sim \chi^2_p$.
   - Easy to compute from output.
   - Can be anti-conservative.
3. **Score (efficient score):** Based on $U(\mathbf{0}) = \frac{\partial \ell_p}{\partial \boldsymbol{\beta}} \big|_{\mathbf{0}}$.
   - Doesn't require $\hat{\boldsymbol{\beta}}$.
   - Used in some diagnostic tests.

Linear combinations & CIs:

For $L = \mathbf{c}^\top \boldsymbol{\beta} = c_1\beta_1 + c_2\beta_2 + \cdots + c_p\beta_p$ (e.g., contrasts, pairwise comparisons):

Point estimate: $\hat{L} = \mathbf{c}^\top \hat{\boldsymbol{\beta}}$.

Variance:

$$\widehat{\text{Var}}(\hat{L}) = \mathbf{c}^\top \widehat{\text{Var}}(\hat{\boldsymbol{\beta}})\mathbf{c} = \sum_i c_i^2 \widehat{\text{Var}}(\hat{\beta}_i) + 2\sum_{i<j} c_i c_j \widehat{\text{Cov}}(\hat{\beta}_i, \hat{\beta}_j)$$

95% CI for L: $\hat{L} \pm z_{\alpha/2}\sqrt{\widehat{\text{Var}}(\hat{L})}$.

95% CI for e^L (HR): $\exp\left[\hat{L} \pm z_{\alpha/2}\sqrt{\widehat{\text{Var}}(\hat{L})}\right]$.

Wald test: $z = \frac{\hat{L}}{\sqrt{\widehat{\text{Var}}(\hat{L})}} \sim N(0,1)$ or $\chi^2 = z^2 \sim \chi^2_1$.

SAS: Use `estimate` statement or extract `covb` from `ods output CovB=covmat;`

E. Baseline & Conditional Survival Estimation

Breslow estimator for $H_0(t)$:

$$\hat{H}_0(t) = \sum_{t_j \le t} \frac{d_j}{\sum_{k \in R(t_j)} \exp(\hat{\boldsymbol{\beta}}^\top \mathbf{x}_k)}$$

Baseline survival:

$$\hat{S}_0(t) = \exp[-\hat{H}_0(t)]$$

Conditional survival for x:

$$\hat{S}(t \mid \mathbf{x}) = [\hat{S}_0(t)]^{\exp(\hat{\boldsymbol{\beta}}^\top \mathbf{x})}$$

SAS Implementation:

Method 1: Default (at mean/reference):

```
proc phreg data=ds plots(cl)=s;
  model time*status(0) = x1 x2 /ties=EFRON;
  baseline out=baseout survival=s lower=lcl upper=ucl;
run;
```

Gives $\hat{S}(t \mid \bar{\mathbf{x}})$ where continuous vars at mean, categorical at reference.

Method 2: Specific covariate patterns:

```
/* Create dataset with desired covariate values */
data covpatterns;
  input id x1 x2;
  datalines;
1 25 1
2 50 0
;
run;

proc phreg data=ds plots(cl overlay)=survival;
  model time*status(0) = x1 x2 /ties=EFRON;
  baseline out=pred covariates=covpatterns
           survival=_all_ /rowid=id;
run;
proc print data=pred; run;
```

Gives $\hat{S}(t \mid \mathbf{x}_1)$ and $\hat{S}(t \mid \mathbf{x}_2)$ with CIs. Use `plots(overlay)` to compare curves.

VI. Model Building & Selection

A. Confounding Assessment

Definition: X_2 confounds effect of X_1 if including X_2 substantially changes $\hat{\beta}_1$.

Procedure:

1. Fit reduced: $h(t, x_1) = h_0(t)e^{\beta_1 x_1}$. Get $\hat{\beta}_1^{\text{crude}}$.
2. Fit full: $h(t, x_1, x_2) = h_0(t)e^{\beta_1 x_1 + \beta_2 x_2}$. Get $\hat{\beta}_1^{\text{adj}}$.
3. Compute **percent change**:

$$\Delta\% = 100 \times \frac{|\hat{\beta}_1^{\text{crude}} - \hat{\beta}_1^{\text{adj}}|}{|\hat{\beta}_1^{\text{adj}}|}$$

4. Threshold: $|\Delta\%| \ge 10$–20% (Hosmer et al.: 20%).

Clinical vs. statistical significance:

- Large $\Delta\%$ but $p > 0.05$ for β_2: may still keep X_2 if clinically important.
- Small $\Delta\%$ and $p > 0.05$: can remove.

B. Effect Modification (Interaction)

Definition: Effect of X_1 varies by levels of X_2.

Model with interaction:

$$h(t, x_1, x_2) = h_0(t)\exp(\beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2)$$

HR interpretation:

- At $x_2 = a$: $\text{HR}(x_1) = \exp(\beta_1 + \beta_3 a)$.
- Effect of x_1 changes with x_2.
- **Test interaction:** Wald test for $H_0 : \beta_3 = 0$.

Variance for $\beta_1 + \beta_3 a$:

$$\widehat{\text{Var}}(\hat{\beta}_1 + a\hat{\beta}_3) = \widehat{\text{Var}}(\hat{\beta}_1) + a^2\widehat{\text{Var}}(\hat{\beta}_3) + 2a\widehat{\text{Cov}}(\hat{\beta}_1, \hat{\beta}_3)$$

95% CI for HR at $x_2 = a$:

$$\exp\left[(\hat{\beta}_1 + a\hat{\beta}_3) \pm z_{\alpha/2}\sqrt{\widehat{\text{Var}}(\hat{\beta}_1 + a\hat{\beta}_3)}\right]$$

Key point: If interaction present, discussion of confounding becomes irrelevant. Focus on stratified or conditional effects.

C. Model Selection Strategy

Step 1: Univariable analysis

- Fit each covariate separately.
- Note significant variables ($p < 0.20$ or 0.25).

Step 2: Initial multivariable model

- Include variables from Step 1 plus clinically important variables.

Step 3: Backward elimination

- Remove non-significant variables one at a time (highest p-value first).
- Check confounding at each step.
- Stop when all remaining variables are significant or important confounders.

Step 4: Check continuous variable linearity

- Martingale residuals vs. covariate.
- Consider splines, polynomials, transformations if nonlinear.

Step 5: Add interactions

- Test clinically plausible interactions.
- Keep if significant.

Step 6: Check PH assumption (see Section VII).

Step 7: Assess fit & influence (see Section VIII).

Model comparison (nested):

- LR test: $G = -2(\ell_{\text{small}} - \ell_{\text{large}}) \sim \chi^2_{\Delta\text{df}}$.
- AIC: $-2\ell_p + 2p$ (lower better).
- BIC: $-2\ell_p + p\log n$ (lower better, penalizes complexity more).

VII. Checking PH Assumption

A. Graphical Methods

1. **Log-log survival plot:**
   - Plot $\log[-\log \hat{S}(t)]$ vs. t (or $\log t$) for each group.
   - Under PH: curves should be roughly parallel.
   - Works for categorical covariates.
2. **Observed vs. expected plots:**
   - Compare KM curve to predicted $\hat{S}(t, \mathbf{x})$ from Cox model.
   - Large deviations suggest PH violation.

B. Schoenfeld Residuals

Definition: For event at t_j, covariate k:

$$r_{jk} = x_{jk} - \frac{\sum_{l \in R(t_j)} x_{lk}\exp(\hat{\boldsymbol{\beta}}^\top \mathbf{x}_l)}{\sum_{l \in R(t_j)} \exp(\hat{\boldsymbol{\beta}}^\top \mathbf{x}_l)}$$

Property: Under PH, $E(r_{jk}) = 0$ for all t_j. If PH violated, r_{jk} trends with time.

Test:

- Regress scaled Schoenfeld residuals on time (or rank of time).
- Slope $\neq 0$ suggests non-PH.
- Global test: combine across covariates.

SAS: `assess ph / resample;`

- Produces supremum test (Kolmogorov-type).
- Martingale-based empirical score process.
- Simulated paths under H_0 vs. observed path.
- p-value from resampling (e.g., 1000 reps).

C. Time-Dependent Coefficients

Fit model: $h(t, \mathbf{x}) = h_0(t)\exp[\boldsymbol{\beta}(t)^\top \mathbf{x}]$. If $\beta_j(t)$ non-constant, PH violated for x_j.

D. Remedies for Non-PH

1. Stratification

Model: $h_s(t, \mathbf{x}) = h_{0s}(t)\exp(\boldsymbol{\beta}^\top \mathbf{x})$ for stratum s.

- Separate baseline hazards per stratum.
- $\boldsymbol{\beta}$ common across strata.
- No estimate for stratifying variable (nuisance).

- SAS: `strata age_group;`
- **Pros:** Robust (no parametric form for time-dependency).
- **Cons:** Loss of information on stratified variable; less efficient.

2. Time Interactions

Linear in time:

$$h(t, x_j) = h_0(t) \exp[(\beta_j + \beta_{jt} \cdot t)x_j]$$

At $t = 0$: HR$= e^{\beta_j}$. HR changes linearly with t.

Log-time:

$$h(t, x_j) = h_0(t) \exp[(\beta_j + \beta_{jt} \log t)x_j]$$

HR changes with $\log t$.

Step function (piecewise):

$$h(t, x_j) = h_0(t) \exp[(\beta_j + \beta_{jt} I(t > \tau))x_j]$$

Different HR before/after τ.

Implementation:

- Create interaction variable (e.g., `x_time = x*time;`).
- Include in model.
- Test $H_0 : \beta_{jt} = 0$.

3. Time-Dependent Covariates (see Section IX)

VIII. Diagnostics & Influence

A. Residuals

1. Martingale residuals:

$$M_i = \delta_i - \hat{H}(Y_i, \mathbf{x}_i)$$

where $\hat{H}(Y_i, \mathbf{x}_i) = \hat{H}_0(Y_i) \exp(\hat{\boldsymbol{\beta}}^\top \mathbf{x}_i)$.

- Range: $(-\infty, 1]$.
- Sum to ≈ 0.
- **Use:** Check functional form. Plot M_i vs. x_j. Lowess smooth should be near 0.
- Nonlinear pattern $\Rightarrow$ consider transformation/spline.

2. Deviance residuals:

$$D_i = \text{sign}(M_i)\sqrt{-2[M_i + \delta_i \log(\delta_i - M_i)]}$$

- More symmetric than martingale.
- **Use:** Identify outliers. $|D_i| > 3$ suspicious.

3. Schoenfeld residuals: See Section VII.B.

4. Score residuals: Contribution to score function. For influence analysis.

B. Influence Measures

1. dfbeta:

$$\text{dfbeta}_{ij} = \hat{\beta}_j - \hat{\beta}_{j(-i)}$$

Change in $\hat{\beta}_j$ when subject i removed.

- Large $|\text{dfbeta}_{ij}| \Rightarrow$ influential.
- Cutoff: $> 2/\sqrt{n}$ or visual inspection.

2. Likelihood displacement (LD):

$$LD_i = 2[\ell_p(\hat{\boldsymbol{\beta}}) - \ell_p(\hat{\boldsymbol{\beta}}_{(-i)})]$$

Overall influence on likelihood.

3. LMAX:

$$\text{LMAX}_i = \max_j \left| \frac{\text{dfbeta}_{ij}}{\widehat{\text{SE}}(\hat{\beta}_j)} \right|$$

Maximum standardized change.

C. Overall Fit

Cox-Snell residuals:

$$r_i^C = \hat{H}(Y_i, \mathbf{x}_i)$$

Under correct model, $r_i^C \sim \text{Exp}(1)$. Check: KM plot of r_i^C should match $S(r) = e^{-r}$.

IX. Time-Dependent Covariates

A. Types

1. **External:** Defined independently of subject (e.g., policy change, calendar time).
2. **Internal:** Defined by subject's history (e.g., biomarker levels, disease progression, treatment received).
3. **Time interactions:** Test PH (covariate $\times$ time).

B. Counting Process Data Format

Each subject contributes multiple rows: $(t_{\text{start}}, t_{\text{stop}}]$ intervals.

- Covariate values constant within each interval.
- Update at change points.
- Event indicator: 1 only in interval containing event.

Example:

ID	start	stop	event	trt
1	0	5	0	0
1	5	10	1	1
2	0	8	0	0

Subject 1: untreated 0–5, treated 5–10, event at 10.

C. Cox Model with TDC

$$h(t, \mathbf{x}(t)) = h_0(t) \exp[\boldsymbol{\beta}^\top \mathbf{x}(t)]$$

Partial likelihood: Risk set $R(t_j)$ includes all with $t_{\text{start}} < t_j \leq t_{\text{stop}}$, using covariate values at t_j.

SAS — Method 1 (Counting Process):

```
proc phreg data=ds_counting;
  model (tstart, tstop)*event(0) = trt age /ties=EFRON;
run;
```

Data must have multiple rows per subject with $(t_{\text{start}}, t_{\text{stop}}]$ intervals.

SAS — Method 2 (Programming Statement):

```
proc phreg data=ds;
  model time*event(0) = trt_td age /ties=EFRON;
  if wait_time >= time or wait_time=. then trt_td = 0;
  else trt_td = 1;
run;
```

Key: Programming statements after MODEL. Compares each event time with all at-risk subjects' waiting times.

D. Immortal Time Bias

Problem: Coding treatment as baseline fixed when it occurs during follow-up.

Example: "Ever treated" vs. "never treated" comparison. Subjects must survive to treatment to be in "ever" group $\Rightarrow$ survival advantage artifactually assigned to treatment.

Solution: Use counting process format. Code treatment as time-varying: 0 until treatment, 1 after.

E. Cautions

- Internal TDC can induce bias if not carefully modeled (e.g., CD4 count as TDC in AIDS studies).
- Avoid "future information": covariate at t shouldn't depend on events after t.

X. Data Preparation & SAS Basics

A. Reading Data

```
/* From external file */
data mydata;
  infile 'C:\data.csv' delimiter=','
         MISSOVER DSD firstobs=2;
  input id time status age trt;
run;

/* Inline data */
data mydata;
  input id time status age trt;
  datalines;
1 10 1 55 0
2 15 0 62 1
;
run;
```

B. Data Manipulation

```
/* Create categorical from continuous */
data mydata;
  set mydata;
  if age < 60 then age_grp = 1;
  else if age < 70 then age_grp = 2;
  else age_grp = 3;

  /* Create interaction */
```

```
  trt_age = trt * age;

  /* Log transformation */
  log_time = log(time);
run;

/* Sort data */
proc sort data=mydata;
  by trt age;
run;
```

C. Descriptive Statistics

```
/* Summary by group */
proc means data=mydata n mean std median min max;
  class trt;
  var age time;
run;

/* Frequency tables */
proc freq data=mydata;
  tables trt*status / chisq;
run;
```

XI. SAS Implementation: Survival Analysis

A. PROC LIFETEST (Kaplan-Meier, Tests)

```
ods graphics on;
proc lifetest data=ds method=KM alpha=0.05
              plots=survival(cl test) conftype=loglog
              outsurv=km_out;
  time time*status(0);           /* 0=censored */
  strata group;                  /* optional: compare groups */
  strata group / test=all;       /* all tests */
  strata group / adjust=bon;     /* Bonferroni */
  strata age(60 70 80);          /* on-the-fly grouping */
run;
ods graphics off;
```

Options:

- `method=KM` (default) or `LT` (life-table).
- `alpha=`: significance level for CI.
- `conftype=`: `loglog` (preferred), `linear`, `log`, `asinsqrt`.
- `plots=survival(cl)` adds confidence bands; `(test)` adds test result on plot.
- `test=`: choose weights for log-rank test. Options: `logrank`, `wilcoxon`, `tarone`, `peto`, `modpeto`, `fleming`. `test=all` shows all.
- `adjust=bon`: Bonferroni adjustment for pairwise comparisons.
- `outsurv=` outputs KM estimates.

B. PROC PHREG (Cox PH)

Basic syntax:

```
proc phreg data=ds;
  class catvar(ref=first)/param=ref;
  model time*status(0) = x1 x2 catvar
        /ties=EFRON risklimits covb;
  /* Baseline survival - default (at mean/reference) */
  baseline out=baseout survival=s lower=lcl upper=ucl;
  /* Conditional survival - specific covariate values */
  baseline out=pred covariates=covar_ds survival=_all_
           /rowid=id;
  /* Assess PH */
  assess ph / resample crpanel;
  /* All pairwise HRs */
  hazardratio 'label' catvar / diff=all cl=pl;
  /* Estimate specific HR with CI */
  estimate 'label' catvar 1 -1 / exp cl;
  /* Store parameter estimates */
  ods output ParameterEstimates=pe
             CovB=covmat;
run;
```

Options:

- `class`: declare categorical variables.
  - `ref=first` (default) or `ref=last` or `ref='level'`.
  - `param=ref` (reference cell), `glm` (GLM coding), `effect`.
- `model` options:
  - `ties=EFRON` (recommended), `BRESLOW`, `EXACT`, `DISCRETE`.
  - `risklimits`: HR CIs for main effects.
  - `covb`: covariance matrix (for linear combinations).
  - `selection=`: stepwise, forward, backward.
  - `slentry=, slstay=`: significance levels for selection.
- `baseline`: compute baseline and conditional survival curves.
  - **Without covariates**=: survival at mean (continuous) / reference (categorical).
  - **With covariates**=: survival for specific covariate patterns in dataset.
  - `out=`: output dataset name.
  - `survival=s`: variable name for $\hat{S}(t)$; use `survival=_all_` for $\hat{S}(t)$, SE, lower, upper.
  - `cumhaz=h`: cumulative hazard $\hat{H}(t)$.
  - `lower=lcl, upper=ucl`: 95% CI bounds.
  - `rowid=id`: identifier for covariate patterns (required with `covariates=` for plotting).
- `assess ph`: test PH assumption.
  - `resample`: resampling-based p-value.
  - `crpanel`: cumulative residual panel plots.
- `hazardratio var / diff=all`: compute all pairwise HRs for categorical var.
  - `cl=pl` (profile likelihood CI, default), `wald`.
- `estimate`: custom linear combinations.
  - `exp`: exponentiate (for HR).
  - `cl`: confidence limits.

Time-dependent covariates (counting process):

```
proc phreg data=ds_cp;
  model (tstart, tstop)*status(0) = trt_tdc age
        /ties=EFRON;
run;
```

Data must have $(t_{\text{start}}, t_{\text{stop}}]$ format.

Stratified Cox:

```
proc phreg data=ds;
  model time*status(0) = x1 x2 /ties=EFRON;
  strata stratum_var;  /* no coef for stratum_var */
run;
```

Time interaction:

```
proc phreg data=ds;
  model time*status(0) = x x_time /ties=EFRON;
  x_time = x * time;  /* or x*log(time) */
run;
```

XII. Complete Analysis Workflow

Step-by-Step Survival Data Analysis

1. Exploratory Data Analysis

- Check data structure: `proc contents`, `proc print` (first 10 obs).
- Descriptive statistics: `proc means`, `proc freq`.
- Check missing data, outliers.
- Calculate follow-up time if needed: `time = end_date - start_date;`

2. Univariate Survival Analysis

- Overall KM curve: `proc lifetest` without strata.
- Report median survival with CI.
- Check if survival curve reaches 0.

3. Bivariate Analysis

- KM curves by exposure/treatment: `strata group;`
- Log-rank test for group differences.
- Assess graphically: do curves cross? (suggests non-PH).

4. Cox Regression — Univariable

- Fit model for each covariate separately.
- Identify candidates ($p < 0.20$ or clinically important).
- Check HR direction and magnitude.

5. Cox Regression — Multivariable

- Include significant + clinically important variables.
- Assess confounding (percent change in exposure effect).
- Test for interactions (exposure × potential effect modifiers).
- Model selection: backward elimination, LR tests.

6. Model Diagnostics

- PH assumption: `assess ph / resample;`
- Functional form: martingale residuals vs. continuous covariates.
- Outliers/influence: deviance residuals, dfbeta.
- Overall fit: Cox-Snell residuals.

7. Address Violations

- Non-PH: stratify, add time interactions, or use TDC.
- Non-linearity: transform variable, use splines, categorize.
- Outliers: sensitivity analysis (refit without outliers).

8. Final Model & Interpretation

- Report adjusted HRs with 95% CIs and p-values.
- Interpret in context (clinical significance).
- Predicted survival curves for key covariate patterns.

9. Reporting

- Table 1: Baseline characteristics by group.
- Figure 1: KM curves with log-rank p.
- Table 2: Univariable and multivariable Cox models.
- Text: Describe methods, results, limitations.

XIII. Interpretation & Reporting

A. HR Reporting Templates

1. **Binary exposure:**
   "Adjusted for [covariates], the hazard of [event] for [exposed] was $\widehat{\text{HR}}$ times that of [reference] (95% CI: [lower, upper], p=[value], Wald test)."
   Example: "Adjusted for age and BMI, the hazard of death for smokers was 1.85 times that of non-smokers (95% CI: 1.23, 2.79, $p = 0.003$)."
2. **Continuous exposure:**
   "Adjusted for [covariates], each [k]-unit increase in [var] was associated with a [HR] times hazard of [event] (95% CI: [lower, upper], p=[value])."
   Example: "Adjusted for sex, each 10 mmHg increase in systolic BP was associated with a 1.25 times hazard of stroke (95% CI: 1.12, 1.39, $p < 0.001$)."
3. **Categorical exposure:**
   "Compared to [reference], the hazard of [event] for [level] was $\widehat{\text{HR}}$ (95% CI: [lower, upper], p=[value])."
   List all levels vs. reference or provide pairwise comparisons.
4. **Interaction:**
   "The effect of [var1] on [event] differed by [var2] ($p_{\text{interaction}}$=[value]). Among [var2=level1], the HR for [var1] was [HR] (95% CI: [lower, upper]). Among [var2=level2], the HR was [HR] (95% CI: [lower, upper])."

B. Model Comparison

"Model 2 (including [additional vars]) provided significantly better fit than Model 1 (LR χ^2=[value], df=[df], p=[value]). The AIC decreased from [AIC1] to [AIC2]."

C. PH Check

"The proportional hazards assumption was assessed using Schoenfeld residuals. The global test was non-significant (p=[value]), and covariate-specific tests showed no evidence of non-proportionality (all $p > 0.05$)."

If violated: "The PH assumption was violated for [var] (p=[value]). We addressed this by [stratifying/time-interaction/TDC]."

D. Confounding

"Including [var2] changed the coefficient for [var1] by [X]%, suggesting [var2] confounds the relationship between [var1] and [outcome]."

XIV. Key Formulas Summary

Survival relationships:

$$S(t) = P(T > t) = \exp[-H(t)]$$

$$h(t) = \lim_{\Delta t \to 0} \frac{P(t \le T < t + \Delta t \mid T \ge t)}{\Delta t} = \frac{f(t)}{S(t)}$$

$$H(t) = \int_0^t h(u)\, du = -\log S(t)$$

KM & variance:

$$\hat{S}(t) = \prod_{t_i \le t} \left(1 - \frac{d_i}{n_i}\right), \quad \widehat{\text{Var}}[\hat{S}(t)] = \hat{S}(t)^2 \sum_{t_i \le t} \frac{d_i}{n_i(n_i - d_i)}$$

Nelson-Aalen:

$$\hat{H}(t) = \sum_{t_i \le t} \frac{d_i}{n_i}, \quad \tilde{S}(t) = e^{-\hat{H}(t)}$$

Log-rank test:

$$\chi^2 = \frac{\left[\sum_j (d_{1j} - e_{1j})\right]^2}{\sum_j v_{1j}}, \quad e_{1j} = \frac{n_{1j} d_j}{n_j}, \quad v_{1j} = \frac{n_{1j} n_{0j} d_j (n_j - d_j)}{n_j^2 (n_j - 1)}$$

Cox model:

$$h(t, \mathbf{x}) = h_0(t) \exp(\boldsymbol{\beta}^\top \mathbf{x})$$

$$\text{HR}(\mathbf{x}_1 : \mathbf{x}_0) = \exp[(\mathbf{x}_1 - \mathbf{x}_0)^\top \boldsymbol{\beta}]$$

$$S(t, \mathbf{x}) = [S_0(t)]^{\exp(\boldsymbol{\beta}^\top \mathbf{x})}$$

Variance of linear combination:

$$\widehat{\text{Var}}(c_1\hat{\beta}_1 + c_2\hat{\beta}_2) = c_1^2 \widehat{\text{Var}}(\hat{\beta}_1) + c_2^2 \widehat{\text{Var}}(\hat{\beta}_2) + 2c_1c_2 \widehat{\text{Cov}}(\hat{\beta}_1, \hat{\beta}_2)$$

Confounding percent change:

$$\Delta\% = 100 \times \frac{|\hat{\beta}_{\text{crude}} - \hat{\beta}_{\text{adj}}|}{|\hat{\beta}_{\text{adj}}|}$$

XV. Common Pitfalls & Tips

1. **Immortal time bias:** Never code time-varying treatment as baseline fixed. Use counting process format.
2. **Ignoring PH:** Always test with `assess ph`. If violated, use stratification, time interactions, or TDC.
3. **Censoring after last event:** If last time is censored, $\hat{S}(t)$ doesn't reach 0. Mean is underestimated. Use RMST or report restriction.
4. **Ties:** With many ties (discrete time), use EFRON or EXACT. BRESLOW can be biased. Default in PROC PHREG is BRESLOW, but EFRON recommended.
5. **Categorical reference:** State reference level explicitly. Check `class` statement output.
6. **Continuous linearity:** Don't assume. Check with martingale residuals. Use splines or transformations if needed.
7. **Multiple comparisons:** Adjust p-values (Bonferroni: α/m) when testing multiple pairwise differences.
8. **Extrapolation:** Don't predict survival far beyond observed follow-up.
9. **Small sample:** Exact tie handling may be needed. PH tests have low power.
10. **Influential observations:** Check dfbeta and deviance residuals. One outlier can distort $\hat{\beta}$.
11. **Reporting:** Always give HRs with CIs and p-values. State adjustments. Describe tie handling and PH checks.

XVI. Quick Reference: Common Questions

Q: How to choose between log-rank and Wilcoxon?
A: Log-rank if assume PH; Wilcoxon if early differences more important or curves cross.

Q: When is median survival undefined?
A: When last observation is censored and $\hat{S}(t_{\max}) > 0.5$. Report RMST instead.

Q: How to interpret HR < 1 vs HR > 1?
A: HR<1: exposure protective (lower hazard); HR>1: exposure harmful (higher hazard). HR=1: no effect.

Q: What if PH assumption violated?
A: (1) Stratify on violating variable, (2) Add time interaction $x \times g(t)$, (3) Use TDC, (4) Use alternative models (parametric, AFT).

Q: Difference between counting process vs programming statement for TDC?
A: Counting process: more flexible, handles complex scenarios. Programming statement: simpler for basic TDC, but limited to simple time-dependencies.

Q: How many events needed for Cox regression?
A: Rule of thumb: ≥ 10 events per covariate. Fewer events $\Rightarrow$ unstable estimates, wide CIs.

Q: Can I use Cox model with small sample?
A: Yes, but use Exact tie handling, check residuals carefully, avoid overfitting (limit covariates).

Q: How to handle tied survival times?
A: EFRON (recommended) for moderate ties; EXACT for many ties or small samples; BRESLOW fast but less accurate. Always specify `ties=EFRON` in PROC PHREG.

Q: What if covariate has missing values?
A: (1) Complete case analysis (if missing completely at random), (2) Multiple imputation, (3) Sensitivity analysis. Never ignore missingness.

Q: How to report censored observations?
A: State # events, # censored, % censored, median follow-up time. Indicate censoring on KM plots (e.g., tick marks).

XVII. Analysis Checklist

Before Analysis:

- ☐ Data cleaned, variables coded correctly (event=1, censored=0).
- ☐ Follow-up time calculated (non-negative).
- ☐ Check for left truncation, interval censoring (special methods needed).
- ☐ Identify time-varying covariates.

During Analysis:

- ☐ KM curves for exposure and key covariates.
- ☐ Test group differences (log-rank or Wilcoxon).
- ☐ Fit univariable Cox models (screening).
- ☐ Fit multivariable model (main effects).
- ☐ Test for confounding (percent change).
- ☐ Test for interactions (if clinically relevant).
- ☐ Check PH assumption (`assess ph`).
- ☐ Check functional form (martingale residuals).
- ☐ Check influence/outliers (dfbeta, deviance residuals).
- ☐ Model comparison (LR tests, AIC).

Reporting:

- ☐ Sample size, # events, % censored, median follow-up.
- ☐ Baseline table (by exposure group).
- ☐ KM curves with CI, log-rank p-value.
- ☐ Cox model table: HR, 95% CI, p-value (univariable + multivariable).
- ☐ State tie handling method (EFRON/BRESLOW/EXACT). Use EFRON in assignments.
- ☐ Report PH assumption check results.
- ☐ Describe how violations addressed.
- ☐ Clinical interpretation of HRs.

XVIII. Software Quick Commands

Read CSV:

```
proc import datafile='data.csv' out=ds dbms=csv replace;
  getnames=yes;
run;
```

Check data:

```
proc contents data=ds; run;
proc print data=ds(obs=10); run;
```

Univariate Cox for all vars:

```
%macro univar(varlist);
  %let n=%sysfunc(countw(&varlist));
  %do i=1 %to &n;
    %let var=%scan(&varlist,&i);
    proc phreg data=ds;
      model time*status(0)=&var;
      title "Univariate: &var";
    run;
  %end;
%mend;
%univar(age sex bmi);
```

Export results:

```
ods rtf file='results.rtf';
proc lifetest data=ds plots=survival;
  time time*status(0);
  strata trt;
run;
ods rtf close;
```