6.7930/HST.956 — Machine Learning for Healthcare
Spring 2025

Recitation 5

Missing Data
Survival Analysis

Ilker Demirel, 3/7/2025

Missing Data

- Very general concept
  - One example: Censoring
- Prevalent in ML
  - Some labs missing for some patients
- Not straightforward to handle
  - Ignore?
  - Impute?
  - Adjust for some other way?

Some Types of Missing Data

- Missing Completely at Random (MCAR)
    - Lab samples damaged during transport

Some Types of Missing Data

- Missing Completely at Random (MCAR)
  - Lab samples damaged during transport
  - Simply dropping —unbiased

Some Types of Missing Data

- Missing Completely at Random (MCAR)
  - Lab samples damaged during transport
  - Simply dropping —unbiased
- Missing at Random (MAR)
  - Elderly patients are more likely to miss appointments
  - No unobserved factors at play

Some Types of Missing Data

- Missing Completely at Random (MCAR)
  - Lab samples damaged during transport
  - Simply dropping —unbiased
- Missing at Random (MAR)
  - Elderly patients are more likely to miss appointments
  - No unobserved factors at play
  - Handling with right statistical approach should be unbiased

Some Types of Missing Data

- Missing Completely at Random (MCAR)
  - Lab samples damaged during transport
  - Simply dropping —unbiased
- Missing at Random (MAR)
  - Elderly patients are more likely to miss appointments
  - No unobserved factors at play
  - Handling with right statistical approach should be unbiased
- Missing Not at Random (MNAR)
  - Patients with worsening conditions can drop out

Some Types of Missing Data

- Missing Completely at Random (MCAR)
  - Lab samples damaged during transport
  - Simply dropping —unbiased
- Missing at Random (MAR)
  - Elderly patients are more likely to miss appointments
  - No unobserved factors at play
  - Handling with right statistical approach should be unbiased
- Missing Not at Random (MNAR)
  - Patients with worsening conditions can drop out
  - Most challenging

Handling Missing Data

- Think about medical studies you read:
  - What data were missing?
  - How did the authors handle it?
  - Were they explicit about their approach?
  - Why might the approach to missing data matter?

Handling Missing Data

- Think about medical studies you read:
  - What data were missing?
  - How did the authors handle it?
  - Were they explicit about their approach?
  - Why might the approach to missing data matter?
- Robustness / sensitivity analyses
  - Try out different approaches / compare to each other
  - Check (if possible) your modeling assumptions

Wrong Modeling Introduces Bias

- Troponin I: Protein found in heart muscle cells
- Leaks into blood in the event of heart attack
- Missing Troponin I measurement: which type?

Wrong Modeling Introduces Bias

- Troponin I: Protein found in heart muscle cells
- Leaks into blood in the event of heart attack
- Missing Troponin I measurement: which type?
  - missing not at random
  - the missingness itself indicates lower risk of heart attack

Wrong Modeling Introduces Bias

- Troponin I: Protein found in heart muscle cells
- Leaks into blood in the event of heart attack
- Missing Troponin I measurement: which type?
  - missing not at random
    - the missingness itself indicates lower risk of heart attack
- Consider a predictive model for heart attack, which uses as a feature troponin I measurements
  - What happens if we impute assuming its missing at random?

Wrong Modeling Introduces Bias

- Troponin I: Protein found in heart muscle cells
- Leaks into blood in the event of heart attack
- Missing Troponin I measurement: which type?
  - missing not at random
    - the missingness itself indicates lower risk of heart attack
- Consider a predictive model for heart attack, which uses as a feature troponin I measurements
  - What happens if we impute assuming its missing at random?
  - Overestimate the risk of heart attack

Survival Analysis

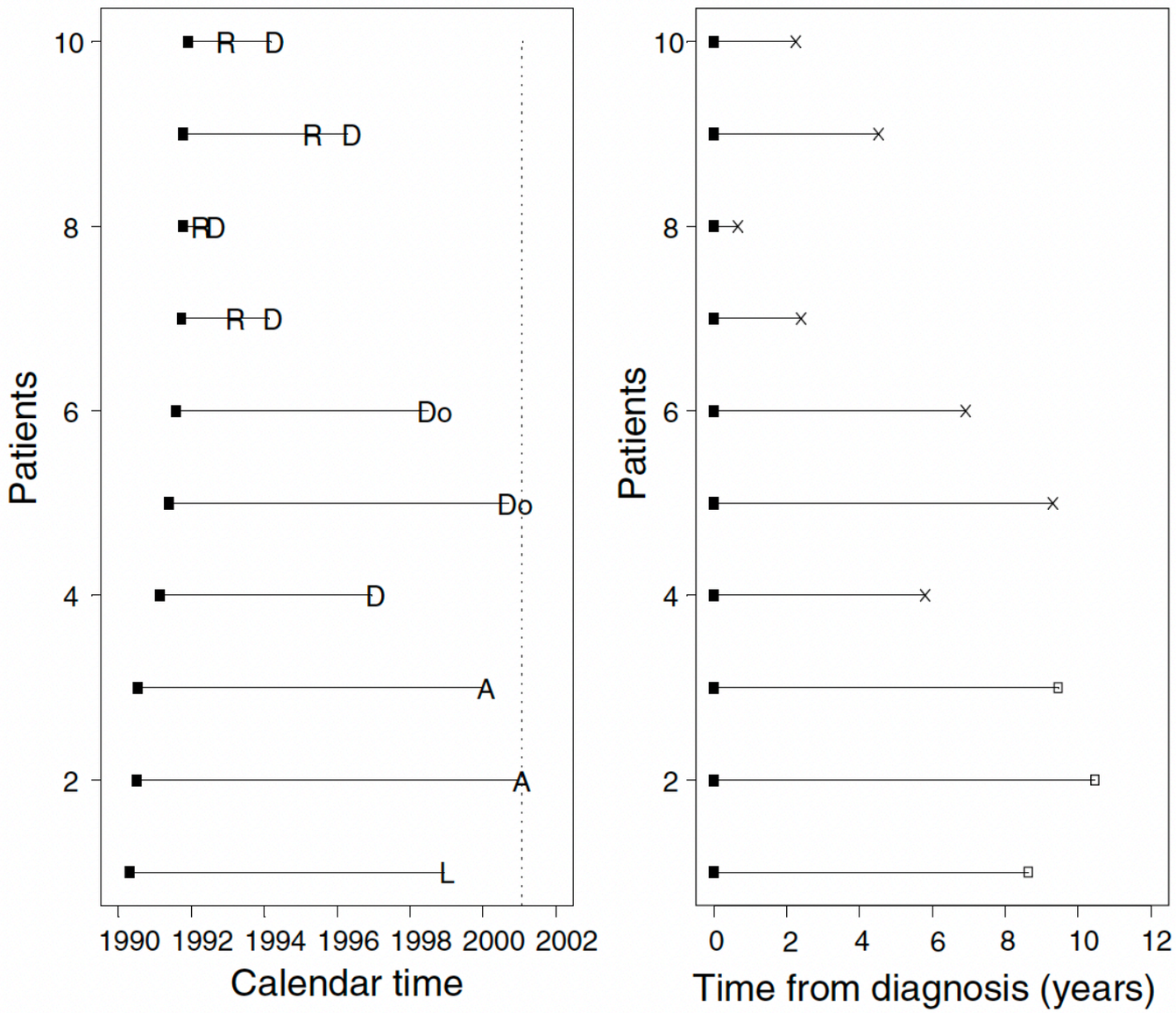


Figure 1 Converting calendar time in the ovarian cancer study to a survival analysis format. Dashed vertical line is the date of the last follow-up, R = relapse, D = death from ovarian cancer, Do = death from other cause, A = attended last clinic visit (alive), L = loss to follow-up, X = death, □ = censored.

- Different censoring reasons
  - Loss-to-follow-up
  - End-of-study
  - Ongoing recruitment
- Uninformative censoring?
  - $T \perp\!\!\!\perp C \mid X$
  - T: time-to-event
  - C: time-to-censoring
  - X: Baseline patient features

See optional readings 1 & 2 on mlhcmit.github.io

Kaplan-Meier (KM) Curve

Table 2 Calculation of the relapse-free survival probability for patients in the lung cancer trial

Radiotherapy ($n = 86$)		Radiotherapy+CAP ($n = 78$)	
Survival times (days)	**Kaplan–Meier survivor function $S(t)$**	**Survival times (days)**	**Kaplan–Meier survivor function $S(t)$**
18	$1 \times (1\text{-}1/86) = 0.988$	9	$1 \times (1\text{-}1/78) = 0.987$
23[a]	$S(18) \times (1\text{-}0/85) = 0.988$	22	$S(18) \times (1\text{-}1/77) = 0.974$
25	$S(23) \times (1\text{-}1/84) = 0.977$	35	$S(22) \times (1\text{-}1/76) = 0.962$
27	$S(25) \times (1\text{-}1/83) = 0.965$	53	$S(35) \times (1\text{-}1/75) = 0.949$
28	$S(27) \times (1\text{-}1/82) = 0.953$	76	$S(53) \times (1\text{-}1/74) = 0.936$
30	$S(28) \times (1\text{-}1/81) = 0.941$	81	$S(76) \times (1\text{-}1/73) = 0.923$
36	$S(30) \times (1\text{-}1/80) = 0.930$	94	$S(81) \times (1\text{-}1/72) = 0.910$
45	$S(36) \times (1\text{-}1/79) = 0.918$	97	$S(94) \times (1\text{-}1/71) = 0.897$
55	$S(45) \times (1\text{-}1/78) = 0.906$	103	$S(97) \times (1\text{-}1/70) = 0.885$
56	$S(55) \times (1\text{-}1/77) = 0.894$	114	$S(103) \times (1\text{-}1/69) = 0.872$
57	$S(56) \times (1\text{-}3/76) = 0.859$	115	$S(114) \times (1\text{-}1/68) = 0.859$
57	$S(56) \times (1\text{-}3/76) = 0.859$	121[a]	$S(115) \times (1\text{-}0/67) = 0.859$
57	$S(56) \times (1\text{-}3/76) = 0.859$	126	$S(121) \times (1\text{-}1/66) = 0.846$
59	$S(57) \times (1\text{-}1/73) = 0.847$	147	$S(126) \times (1\text{-}1/65) = 0.833$
62	$S(59) \times (1\text{-}1/72) = 0.835$	154	$S(147) \times (1\text{-}1/64) = 0.820$
⋮	⋮	⋮	⋮
2252[a]	$S(2209) \times (1\text{-}0/5) = 0.115$	2220[a]	$S(2218) \times (1\text{-}0/5) = 0.273$
2286[a]	$S(2286) \times (1\text{-}0/4) = 0.115$	2375	$S(2220) \times (1\text{-}0/4) = 0.205$
2305[a]	$S(2305) \times (1\text{-}0/3) = 0.115$	2566	$S(2375) \times (1\text{-}0/3) = 0.137$
2318[a]	$S(2318) \times (1\text{-}0/2) = 0.115$	2875[b]	$S(2566) \times (1\text{-}0/2) = 0.137$
2940[a]	$S(2940) \times (1\text{-}0/1) = 0.115$	3067[b]	$S(2875) \times (1\text{-}0/1) = 0.137$

$S(0) = 1$, (CAP = cytoxan, doxorubicin and platinum-based chemotherapy.) [a]Lost to follow-up and considered censored. [b]Relapse-free at time of analysis and considered censored.

See optional readings 1 & 2 on mlhcmit.github.io

Kaplan-Meier (KM) Curve

Table 2 Calculation of the relapse-free survival probability for patients in the lung cancer trial

Radiotherapy ($n = 86$)		Radiotherapy+CAP ($n = 78$)	
Survival times (days)	**Kaplan–Meier survivor function $S(t)$**	**Survival times (days)**	**Kaplan–Meier survivor function $S(t)$**
18	$1 \times (1\text{-}1/86) = 0.988$		.987
23[a]	$S(18) \times (1\text{-}0/85) = 0.988$		= 0.974
25	$S(23) \times (1\text{-}1/84) = 0.977$		= 0.962
27	$S(25) \times (1\text{-}1/83) = 0.965$		= 0.949
28	$S(27) \times (1\text{-}1/82) = 0.953$		= 0.936
30	$S(28) \times (1\text{-}1/81) = 0.941$		= 0.923
36	$S(30) \times (1\text{-}1/80) = 0.930$		= 0.910
45	$S(36) \times (1\text{-}1/79) = 0.918$		= 0.897
55	$S(45) \times (1\text{-}1/78) = 0.906$		= 0.885
56	$S(55) \times (1\text{-}1/77) = 0.894$		= 0.872
57	$S(56) \times (1\text{-}3/76) = 0.859$		= 0.859
57	$S(56) \times (1\text{-}3/76) = 0.859$		= 0.859
57	$S(56) \times (1\text{-}3/76) = 0.859$		= 0.846
59	$S(57) \times (1\text{-}1/73) = 0.847$		= 0.833
62	$S(59) \times (1\text{-}1/72) = 0.835$		= 0.820
	⋮		
2252[a]	$S(2209) \times (1\text{-}0/5) = 0.115$		= 0.273
2286[a]	$S(2286) \times (1\text{-}0/4) = 0.115$		= 0.205
2305[a]	$S(2305) \times (1\text{-}0/3) = 0.115$		= 0.137
2318[a]	$S(2318) \times (1\text{-}0/2) = 0.115$		= 0.137
2940[a]	$S(2940) \times (1\text{-}0/1) = 0.115$		= 0.137

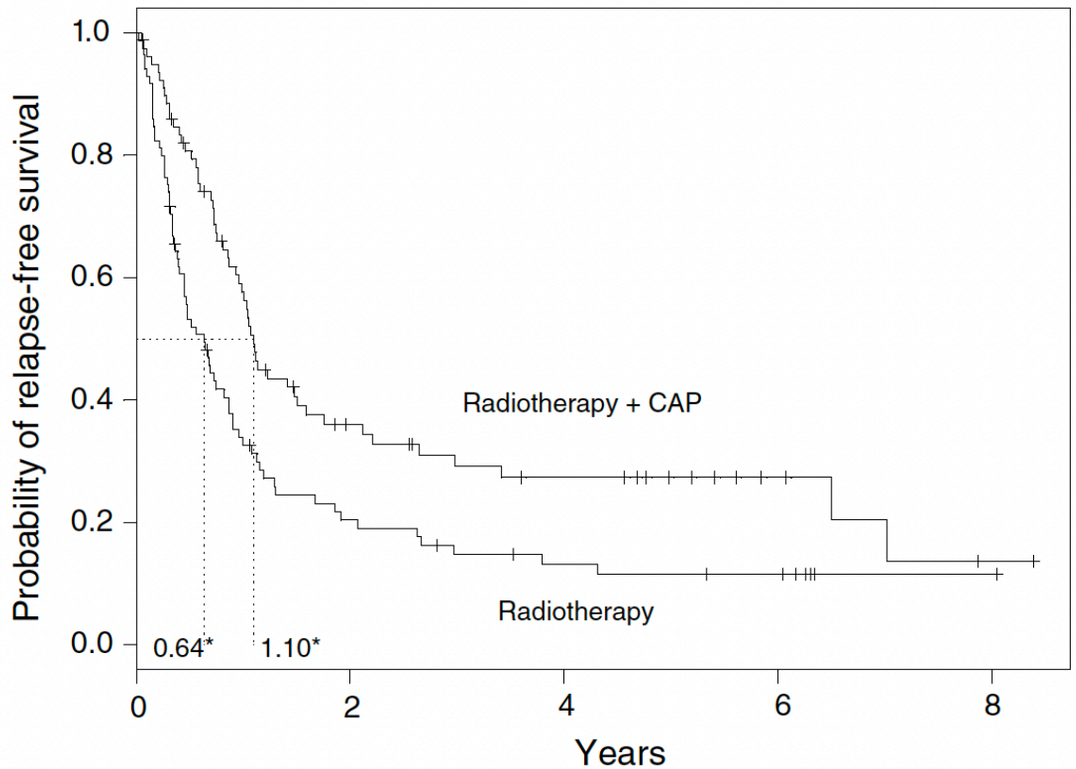


Figure 2 Relapse-free survival curves for the lung cancer trial. * Median relapse-free survival time for each arm, + censoring times, CAP = cytoxan, doxorubicin and platinum-based chemotherapy.

$S(0) = 1$, (CAP = cytoxan, doxorubicin and platinum-based chemotherapy.) [a]Lost to follow-up and considered censored. [b]Relapse-free at time of analysis and considered censored.

Nonparametric

Not personalized

See optional readings 1 & 2 on mlhcmit.github.io

CoxPH Models (Semi-parametric)

https://colab.research.google.com/drive/1Mg_vX-IjvxHSyC7FObrjcYHY2guhoR0j?usp=sharing

Hazard Function

$$\lambda(t \mid X_i) = \lambda_0(t)\exp(\beta X_i)$$

- Personalized (depends on X)
- Proportional Hazards (PH) assumption
- $\exp(\beta(j))$: Hazard Ratio (HR) of X(j)

CoxPH Models (Semi-parametric)

https://colab.research.google.com/drive/1Mg_vX-IjvxHSyC7FObrjcYHY2guhoR0j?usp=sharing

Hazard Function

$$\lambda(t \mid X_i) = \lambda_0(t)\exp(\beta X_i)$$

—> 1:1

$$S(t \mid X) = \exp\left(-\int_0^t \lambda(s \mid X)\, ds\right)$$

- Personalized (depends on X)
- Proportional Hazards (PH) assumption
- $\exp(\beta(j))$: Hazard Ratio (HR) of X(j)

https://www.nature.com/articles/s41467-023-43363-3.pdf

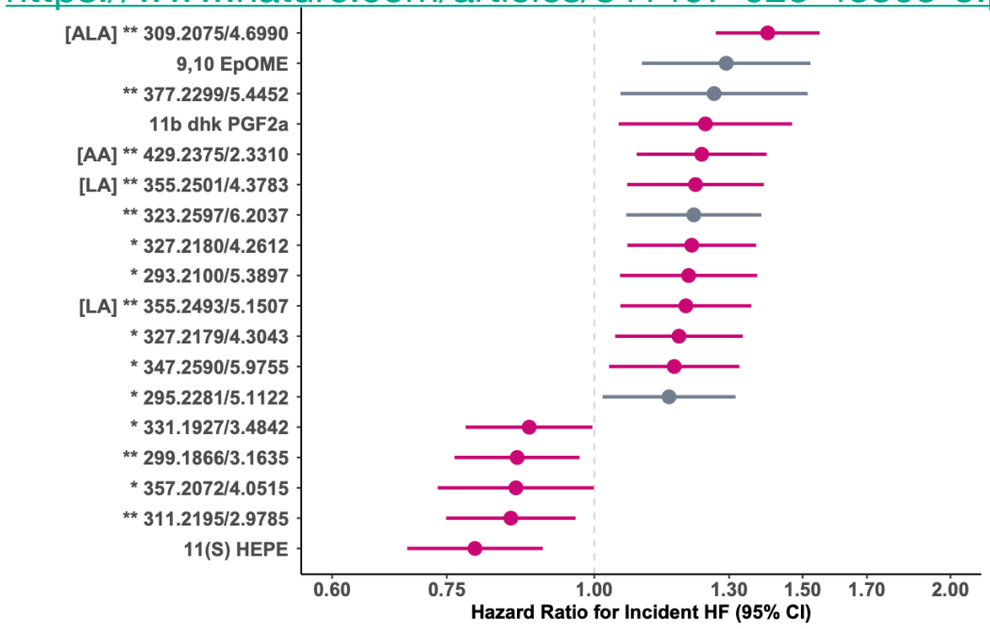


Fig. 4 | Association of HFpEF-related eicosanoids and eicosanoid-related metabolites (from MGH CPET) with incident HF in MESA. Eicosanoids and eicosanoid-related metabolites chosen for analysis in MESA were significantly associated with HFpEF status in the MGH CPET cohort. Hazards ratios displayed

CoxPH Models (Semi-parametric)

https://colab.research.google.com/drive/1Mg_vX-IjvxHSyC7FObrjcYHY2guhoR0j?usp=sharing

Hazard Function

$$\lambda(t \mid X_i) = \lambda_0(t)\exp(\beta X_i)$$

—> 1:1

$$S(t \mid X) = \exp\left(-\int_0^t \lambda(s \mid X)\, ds\right)$$

- Personalized (depends on X)
- Proportional Hazards (PH) assumption
- $\exp(\beta(j))$: Hazard Ratio (HR) of X(j)

How to estimate β?

https://www.nature.com/articles/s41467-023-43363-3.pdf

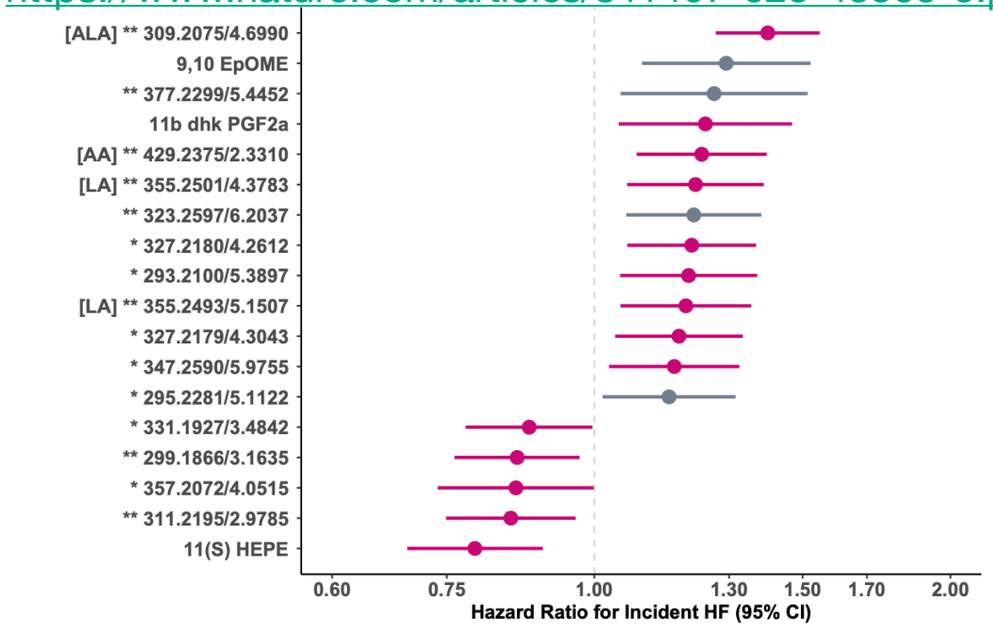


Fig. 4 | Association of HFpEF-related eicosanoids and eicosanoid-related metabolites (from MGH CPET) with incident HF in MESA. Eicosanoids and eicosanoid-related metabolites chosen for analysis in MESA were significantly associated with HFpEF status in the MGH CPET cohort. Hazards ratios displayed

CoxPH Models (Semi-parametric)

https://colab.research.google.com/drive/1Mg_vX-IjvxHSyC7FObrjcYHY2guhoR0j?usp=sharing

Hazard Function

$$\lambda(t \mid X_i) = \lambda_0(t)\exp(\beta X_i)$$

—> 1:1

$$S(t \mid X) = \exp\left(-\int_0^t \lambda(s \mid X)\, ds\right)$$

- Personalized (depends on X)
- Proportional Hazards (PH) assumption

- $\exp(\beta(j))$: Hazard Ratio (HR) of X(j)

How to estimate β?

The **partial likelihood function** for β is:

$$L(\beta) = \prod_{i=1}^{D} \frac{\exp(\beta^T X_i)}{\sum_{j \in R(t_i)} \exp(\beta^T X_j)}$$

where:

- D is the number of observed failure (event) times.
- X_i is the covariate vector for the individual who experienced an event at time t_i.
- $R(t_i)$ is the **risk set**, which consists of all individuals who were still at risk just before time t_i
- The **numerator** represents the contribution of the individual who failed at t_i.
- The **denominator** is the sum of risk scores for all individuals who were still at risk at t_i.

https://www.nature.com/articles/s41467-023-43363-3.pdf

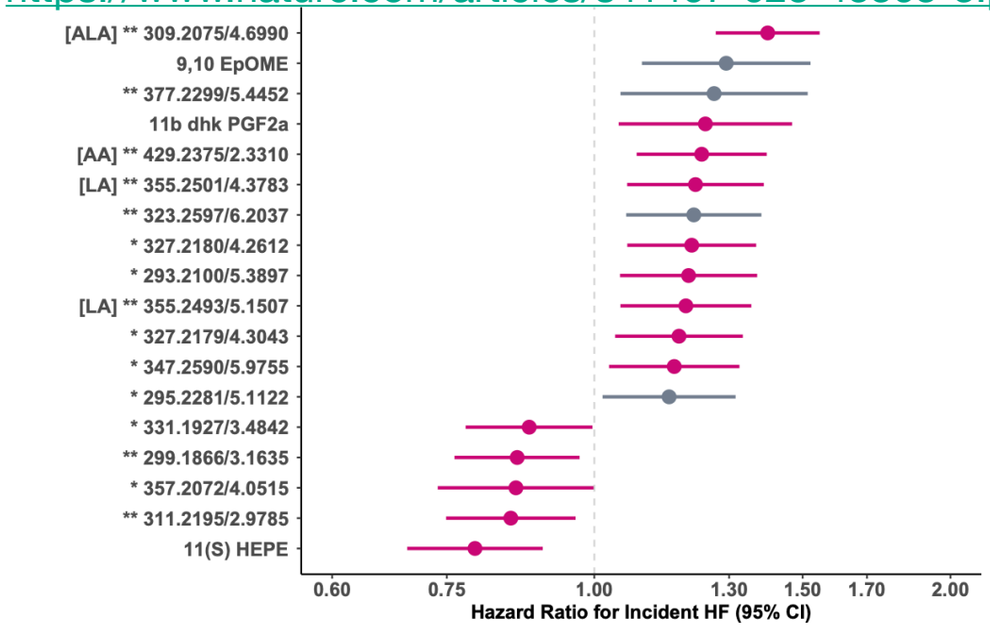


Fig. 4 | Association of HFpEF-related eicosanoids and eicosanoid-related metabolites (from MGH CPET) with incident HF in MESA. Eicosanoids and eicosanoid-related metabolites chosen for analysis in MESA were significantly associated with HFpEF status in the MGH CPET cohort. Hazards ratios displayed

Taking the natural logarithm, we get the **log-partial likelihood**:

$$\ell(\beta) = \sum_{i=1}^{D} \left[\beta^T X_i - \log \sum_{j \in R(t_i)} \exp(\beta^T X_j) \right]$$

This function is maximized to estimate β.

CoxPH Models (Semi-parametric)

https://colab.research.google.com/drive/1Mg_vX-IjvxHSyC7FObrjcYHY2guhoR0j?usp=sharing

Hazard Function

$$\lambda(t \mid X_i) = \lambda_0(t)\exp(\beta X_i)$$

- Proportional Hazards (PH) assumption:
- Can we check this assumption?

CoxPH Models (Semi-parametric)

https://colab.research.google.com/drive/1Mg_vX-IjvxHSyC7FObrjcYHY2guhoR0j?usp=sharing

Hazard Function

$$\lambda(t \mid X_i) = \lambda_0(t)\exp(\beta X_i)$$

- Proportional Hazards (PH) assumption:
  - Can we check this assumption?
  - This implies that the hazard curves across subgroups cannot cross

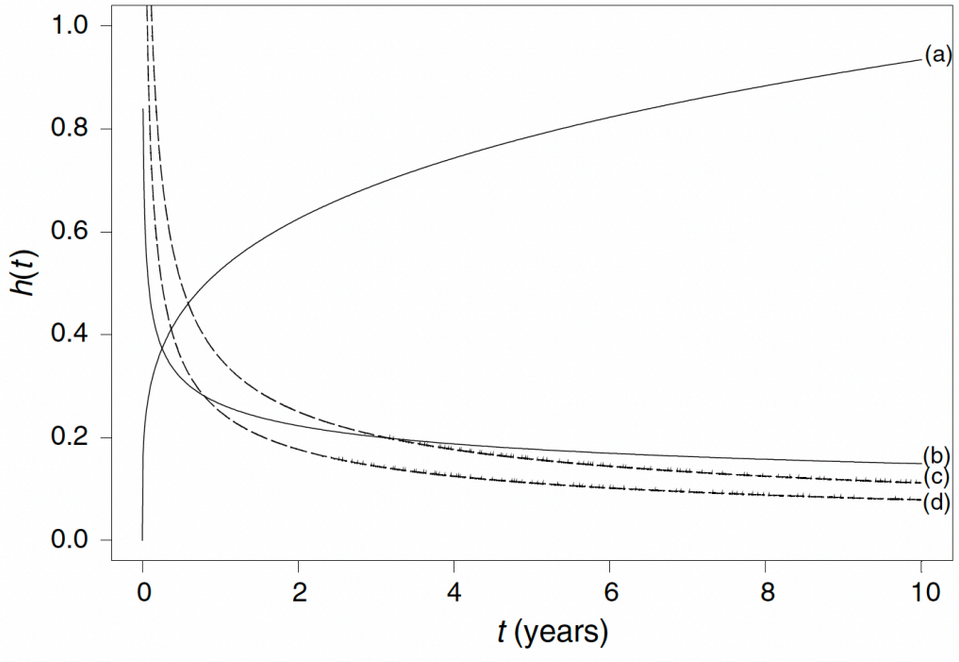


Figure 1 Example of (non-) proportional hazards (groups (c) and (d) only have proportional hazards) using the Weibull distribution. For the Weibull survival model, the hazard function $h(t) = \lambda s(\lambda t)^{s-1}$ for λ, $s > 0$: (a) increasing hazard ($\lambda = 0.5$, $s = 1.25$); (b) decreasing hazard ($\lambda = 0.25$, $s = 0.75$); (c) decreasing hazard ($\lambda = 0.5$, $s = 0.5$); (d) decreasing hazard ($\lambda = 0.25$, $s = 0.5$).

See optional readings 1 & 2 on mlhcmit.github.io

CoxPH Models (Semi-parametric)

https://colab.research.google.com/drive/1Mg_vX-IjvxHSyC7FObrjcYHY2guhoR0j?usp=sharing

Hazard Function

$$\lambda(t \mid X_i) = \lambda_0(t)\exp(\beta X_i)$$

- Proportional Hazards (PH) assumption:
  - Can we check this assumption?
  - This implies that the hazard curves across subgroups cannot cross

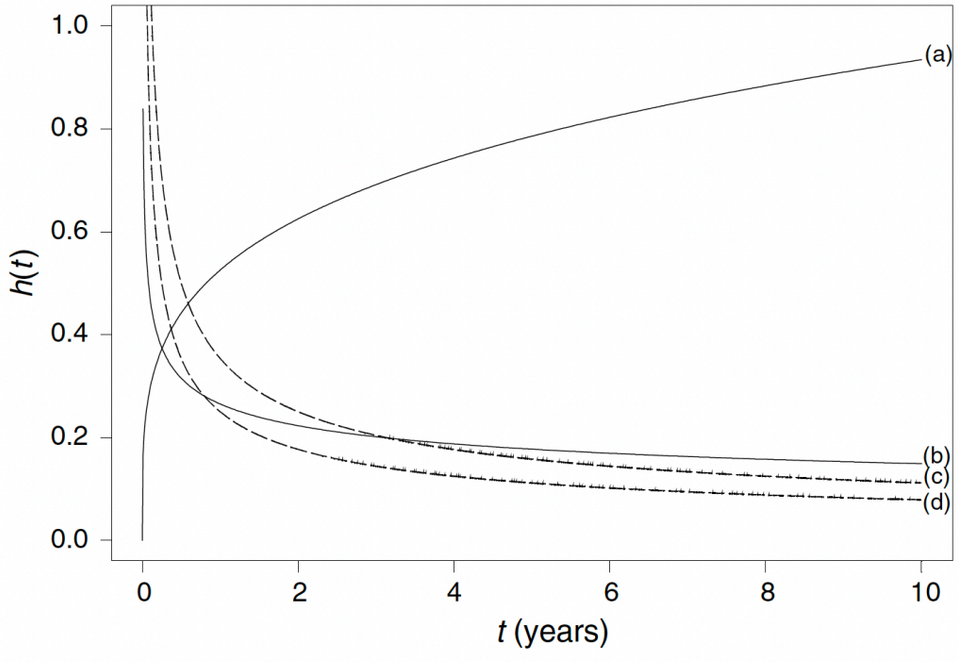


Figure 1 Example of (non-) proportional hazards (groups (c) and (d) only have proportional hazards) using the Weibull distribution. For the Weibull survival model, the hazard function $h(t) = \lambda s(\lambda t)^{s-1}$ for $\lambda, s > 0$: (a) increasing hazard ($\lambda = 0.5$, $s = 1.25$); (b) decreasing hazard ($\lambda = 0.25$, $s = 0.75$); (c) decreasing hazard ($\lambda = 0.5$, $s = 0.5$); (d) decreasing hazard ($\lambda = 0.25$, $s = 0.5$).

How to estimate $\lambda_0(t)$ non-parametrically?

Most common: Breslow estimator

Alternative: Parametric CoxPH model

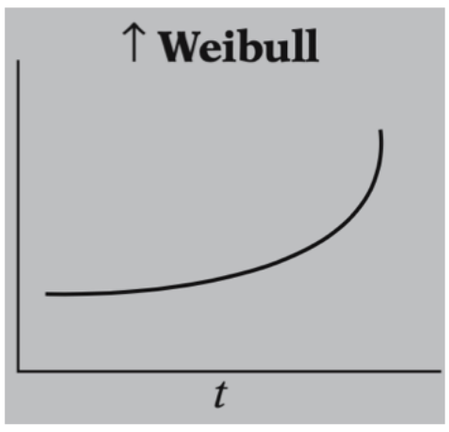
See optional readings 1 & 2 on mlhcmit.github.io

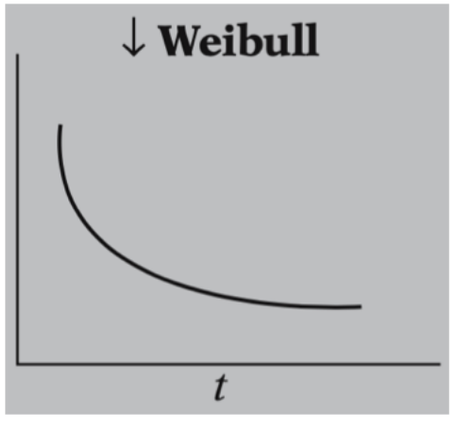
Parametric CoxPH

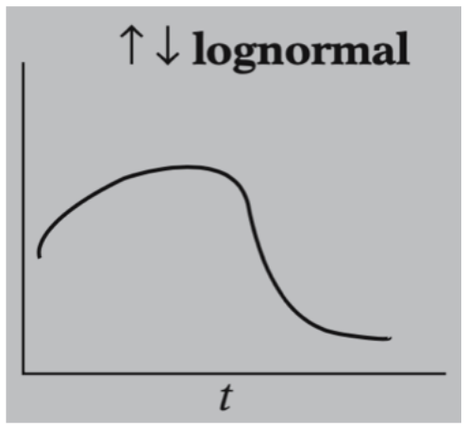
Table IV: Density, Survival and Hazard functions for the distributions commonly used in the parametric methods in survival analysis.

Distribution	**PDF** $f(t)$	**Survival** $S(t)$	**Hazard** $h(t)$
Exponential	$\lambda exp(-\lambda t)$	$exp(-\lambda t)$	λ
Weibull	$\lambda k t^{k-1} exp(-\lambda t^k)$	$exp(-\lambda t^k)$	$\lambda k t^{k-1}$
Logistic	$\frac{e^{-(t-\mu)/\sigma}}{\sigma(1+e^{-(t-\mu)/\sigma})^2}$	$\frac{e^{-(t-\mu)/\sigma}}{1+e^{-(t-\mu)/\sigma}}$	$\frac{1}{\sigma(1+e^{-(t-\mu)/\sigma})}$
Log-logistic	$\frac{\lambda k t^{k-1}}{(1+\lambda t^k)^2}$	$\frac{1}{1+\lambda t^k}$	$\frac{\lambda k t^{k-1}}{1+\lambda t^k}$
Normal	$\frac{1}{\sqrt{2\pi}\sigma} exp(-\frac{(t-\mu)^2}{2\sigma^2})$	$1-\Phi(\frac{t-\mu}{\sigma})$	$\frac{1}{\sqrt{2\pi}\sigma(1-\Phi((t-\mu)/\sigma))} exp(-\frac{(t-\mu)^2}{2\sigma^2})$
Log-normal	$\frac{1}{\sqrt{2\pi}\sigma t} exp(-\frac{(log(t)-\mu)^2}{2\sigma^2})$	$1-\Phi(\frac{log(t)-\mu}{\sigma})$	$\frac{\frac{1}{\sqrt{2\pi}\sigma t} exp(-(log(t)-\mu)^2/2\sigma^2)}{1-\Phi(\frac{log(t)-\mu}{\sigma})}$

We obtain **conditional** models $f(t \mid \boldsymbol{x}; \beta)$ by letting, e.g., $\lambda = \exp(\boldsymbol{\beta} \cdot \boldsymbol{x})$







Slide from https://mlhcmit.github.io/slides/lecture9.pdf (pages 4 and 5)

More parametric options

Accelerated Failure Time Models (AFT)

Effect of covariates:

Delay/accelerate time-to-event

Model time-to-event directly

$$\log(T) = \beta(1)X(1) + \ldots + \beta(p)X(p) + \sigma\epsilon$$

Can do the parametrization with

neural nets too (see notebook)

More parametric options

Accelerated Failure Time Models

Effect of covariates:

Delay/accelerate time-to-event

Model time-to-event directly

$$\log(T) = \beta(1)X(1) + \ldots + \beta(p)X(p) + \sigma\epsilon$$

Can do the parametrization with

neural nets too (see notebook)

Interpreting AFT Results

- Time Ratios (TR) = exp(β)
- TR > 1: Longer survival time
- TR < 1: Shorter survival time
- Example: TR = 2 means survival time is doubled

Advantages of AFT Models

- Direct interpretation of survival time
- Can extrapolate beyond observed data
- No proportional hazards assumption
- More efficient when distribution is correct

Limitations of AFT Models

- Requires correct specification of distribution
- Less widely used in medical literature
- Software implementation varies
- More complex for time-varying covariates