Models for time-to-event data
From Cox’s proportional hazards model to deep learning

Sebastian Pölsterl
Artificial Intelligence in Medical Imaging | Ludwig Maximilian Universität Munich

October 2nd 2018
École Centrale de Nantes
1 What is Survival Analysis?

2 Parametric Survival Models

3 Semiparametric Survival Models

4 Non-Linear Survival Models

5 Survival Analysis with Deep Learning

6 Conclusion
Mild cognitive impairment (MCI) is a common precursor to dementia in Alzheimer’s disease and is associated with isolated memory loss.

Some patients with MCI remain stable, whereas others progress to Alzheimer’s disease.

For an effective therapy, we want to know the probability of conversion at any time point.
Most equipment, such as a pump, will experience failure eventually.

Failure is usually determined by threshold values on various censors: temperature cannot exceed 74°C and pressure must be under 10 bar.

We want to know the probability of failure at any time point such that replacing the equipment can be scheduled in advance to minimize downtime.
• All businesses will lose some of its customers (customer churn).
• For each customer, we have a record of purchases and previous interactions with the company.
• We want to know how likely it is for a customer to turn away (churn) at any given time point so we can provide targeted incentives to induce customers to stay.
1. What is Survival Analysis?

2. Parametric Survival Models

3. Semiparametric Survival Models

4. Non-Linear Survival Models

5. Survival Analysis with Deep Learning

6. Conclusion
A record is **uncensored** if an event was observed during the study period: the exact time of the event is known.

A record is **right censored** if a patient remained event-free: it is unknown whether an event occurred after the study ended.
• A record is **uncensored** if an event was observed during the study period: the exact time of the event is known.
A record is **uncensored** if an event was observed during the study period: the **exact time** of the event is known.

A record is **right censored** if a patient remained event-free: it is **unknown** whether an event occurred after the study ended.
Types of Censoring

Let $y_i$ denote the **observable** time, $t_i$ the **actual** time of an event, and $c_i$ the time of **censoring**.

- **Right censoring**
  
  $$y_i = \min(c_i^{\text{right}}, t_i)$$
Let $y_i$ denote the observable time, $t_i$ the actual time of an event, and $c_i$ the time of censoring.

- **Right censoring**
  \[ y_i = \min(c_i^{\text{right}}, t_i) \]

- **Left censoring**
  \[ y_i = \max(c_i^{\text{left}}, t_i) \]
Let $y_i$ denote the **observable** time, $t_i$ the **actual** time of an event, and $c_i$ the time of **censoring**.

- **Right censoring**
  $$y_i = \min(c_i^{\text{right}}, t_i)$$

- **Left censoring**
  $$y_i = \max(c_i^{\text{left}}, t_i)$$

- **Interval censoring**
  $$t_i \in (\tau_i^l; \tau_i^r]$$
Let $y_i$ denote the **observable** time, $t_i$ the **actual** time of an event, and $c_i$ the time of **censoring**.

- **Right censoring**
  \[ y_i = \min(c_i^{\text{right}}, t_i) \]

- **Left censoring**
  \[ y_i = \max(c_i^{\text{left}}, t_i) \]

- **Interval censoring**
  \[ t_i \in (\tau_i^l; \tau_i^r] \]

- Any combination of left, right, or interval censoring may occur in a study.
Let $T$ denote a **continuous** non-negative random variable corresponding to a patient’s survival time with probability density function $f(t)$.

**Survival function**

\[
S(t) = P(T > t) = 1 - P(T \leq t) = 1 - F(t) = \int_{t}^{\infty} f(u)\,du
\]
Let $T$ denote a **continuous** non-negative random variable corresponding to a patient’s survival time with probability density function $f(t)$.

**Survival function**

$$S(t) = P(T > t) = 1 - P(T \leq t) = 1 - F(t) = \int_t^\infty f(u)\,du$$

**Hazard function**

$$h(t) = \lim_{\Delta t \to 0} \frac{P(t \leq T < t + \Delta t \mid T \geq t)}{\Delta t} \geq 0$$
Let $T$ denote a \textbf{continuous} non-negative random variable corresponding to a patient’s survival time with probability density function $f(t)$.

\textbf{Survival function}

\[
S(t) = P(T > t) = 1 - P(T \leq t) = 1 - F(t) = \int_t^{\infty} f(u) du
\]

\textbf{Hazard function}

\[
h(t) = \lim_{\Delta t \to 0} \frac{P(t \leq T < t + \Delta t \mid T \geq t)}{\Delta t} \geq 0
\]

\textbf{Cumulative hazard function}

\[
H(t) = \int_0^t h(u) du
\]
Survival and Hazard Function

\[ h(t) = \frac{f(t)}{S(t)}; \quad H(t) = -\log S(t) \]
Let $T$ be a **discrete** random variable, which can take on values $t_i$ ($i \in \mathbb{N}$) with probability mass function $P(T = t_i)$ and $t_i < t_j$ if and only if $i < j$.

**Survival function**

$$
S(t) = \sum_{\{i | t_i > t\}} P(T = t_i) \iff P(T = t_i) = S(t_{i-1}) - S(t_i)
$$
Let $T$ be a **discrete** random variable, which can take on values $t_i \ (i \in \mathbb{N})$ with probability mass function $P(T = t_i)$ and $t_i < t_j$ if and only if $i < j$.

**Survival function**

$$S(t) = \sum_{\{i \mid t_i > t\}} P(T = t_i) \iff P(T = t_i) = S(t_{i-1}) - S(t_i)$$

**Hazard function**

$$h(t) = P(T = t_i \mid T \geq t_i)$$
Let $T$ be a \textit{discrete} random variable, which can take on values $t_i$ ($i \in \mathbb{N}$) with probability mass function $P(T = t_i)$ and $t_i < t_j$ if and only if $i < j$.

\textbf{Survival function}

$$S(t) = \sum_{\{i \mid t_i > t\}} P(T = t_i) \Leftrightarrow P(T = t_i) = S(t_{i-1}) - S(t_i)$$

\textbf{Hazard function}

$$h(t) = P(T = t_i \mid T \geq t_i)$$

\textbf{Cumulative hazard function}

$$H(t) = \sum_{\{i \mid t_i \leq t\}} h(t_i)$$
1. What is Survival Analysis?

2. Parametric Survival Models

3. Semiparametric Survival Models

4. Non-Linear Survival Models

5. Survival Analysis with Deep Learning

6. Conclusion
• Assume we have a dataset of \( d \) covariates for each of \( n \) observations:

\[
D = \{(y_i, x_i)\}_{i=1}^{n}
\]

• We want to fit a model with parameters \( \Theta \) to estimate \( S(t) \) – the probability of survival beyond time \( t \) – via maximum likelihood optimization.

• Observed times \( y_i \) can be
  1. uncensored
  2. right-censored
  3. left-censored
  4. interval-censored

• We need to consider carefully what information each observation gives us.
Noninformative Censoring

Definition (Noninformative Censoring)

Usually, we assume that the distribution of survival times $T$ is independent of the distribution of censoring times $C$:

$$T \perp C \mid x$$

This assumption would be violated if the prognosis of individuals who get censored is worse compared to those who are not censored.
Exact time of event is known

\[ \arg\max_{\Theta} P(T = y_i; \Theta | x_i) = f(y_i; \Theta | x_i) \]
Time of event is right-censored

\[
\argmax_\Theta P(T > c_i; \Theta \mid x_i) = S(c_i; \Theta \mid x_i)
\]
Time of event is left-censored

\[
\arg\max_{\Theta} P(T \leq c_i; \Theta \mid \mathbf{x}_i) = 1 - S(c_i; \Theta \mid \mathbf{x}_i)
\]
Time of event is interval-censored

\[
\arg\max_{\Theta} P(\tau^l_i < T \leq \tau^r_i; \Theta \mid x_i) = \int_{\tau^l_i}^{\tau^r_i} f(u; \Theta \mid x_i) \, du \\
= S(\tau^l_i; \Theta \mid x_i) - S(\tau^r_i; \Theta \mid x_i)
\]
For training, we need to solve the optimization problem

$$ \arg\max_{\Theta} \quad LL(\Theta) $$

where the likelihood function comprises all of the components

$$ LL(\Theta) = \prod_{i \in \text{uncensored}} f(y_i; \Theta | x_i) \cdot \prod_{i \in \text{right-censored}} S(y_i; \Theta | x_i) \cdot \prod_{i \in \text{left-censored}} (1 - S(y_i; \Theta | x_i)) \cdot \prod_{i \in \text{interval-censored}} \left( S(\tau_i^l; \Theta | x_i) - S(\tau_i^r; \Theta | x_i) \right) $$
Common Parametric Distributions

Exponential
Weibull
Log logistic
Gamma
Gompertz

Time $t$
Hazard $h(t)$

0.0 0.2 0.4 0.6 0.8 1.0 1.2 1.4

Exponential
Weibull
Log logistic
Gamma
Gompertz
1 What is Survival Analysis?

2 Parametric Survival Models

3 Semiparametric Survival Models

4 Non-Linear Survival Models

5 Survival Analysis with Deep Learning

6 Conclusion
Parametric Models

- Distribution’s parameters are data-dependent based on covariates.
- Work extremely well when survival times follow the chosen distribution.
- Can easily account for various censoring schemes.
- Inference is easy.
Parametric Models

- Distribution’s parameters are data-dependent based on covariates.
- Work extremely well when survival times follow the chosen distribution.
- Can easily account for various censoring schemes.
- Inference is easy.

Semiparametric Models

- Often, we do not know what distribution we should choose.
- Split the model into 2 parts:
  1. part that models influence of covariates.
  2. part that models time.
- Usually only account for right-censoring.
• Cox’s Proportional Hazards model (Cox PH)

\[ h(t \mid x) = h_0(t) \exp \left( x^\top \beta \right) \iff \frac{h(t \mid x)}{h_0(t)} = \exp \left( x^\top \beta \right) \]
Common Semiparametric Linear Models

- Cox’s Proportional Hazards model (Cox PH)

\[ h(t \mid \mathbf{x}) = h_0(t) \exp \left( \mathbf{x}^\top \beta \right) \iff \frac{h(t \mid \mathbf{x})}{h_0(t)} = \exp \left( \mathbf{x}^\top \beta \right) \]

- Accelerated Failure Time model (AFT)

\[ h(t \mid \mathbf{x}) = h_0(t \exp(-\mathbf{x}^\top \beta)) \exp(-\mathbf{x}^\top \beta) \]
• Cox’s Proportional Hazards model (Cox PH)
  \[ h(t \mid \mathbf{x}) = h_0(t) \exp \left( \mathbf{x}^\top \beta \right) \iff \frac{h(t \mid \mathbf{x})}{h_0(t)} = \exp \left( \mathbf{x}^\top \beta \right) \]

• Accelerated Failure Time model (AFT)
  \[ h(t \mid \mathbf{x}) = h_0(t \exp(-\mathbf{x}^\top \beta)) \exp(-\mathbf{x}^\top \beta) \]

• Proportional Odds model
  \[
  \frac{P(T > t \mid \mathbf{x})}{P(T \leq t \mid \mathbf{x})} = \frac{1 - S(t \mid \mathbf{x})}{S(t \mid \mathbf{x})} = \frac{1 - S_0(t)}{S_0(t)} \exp \left( \mathbf{x}^\top \beta \right)
  \]
• Cox’s Proportional Hazards model (Cox PH)

\[ h(t \mid \mathbf{x}) = h_0(t) \exp \left( \mathbf{x}^\top \beta \right) \iff \frac{h(t \mid \mathbf{x})}{h_0(t)} = \exp \left( \mathbf{x}^\top \beta \right) \]

• Accelerated Failure Time model (AFT)

\[ h(t \mid \mathbf{x}) = h_0(t \exp(-\mathbf{x}^\top \beta)) \exp(-\mathbf{x}^\top \beta) \]

• Proportional Odds model

\[ \frac{P(T > t \mid \mathbf{x})}{P(T \leq t \mid \mathbf{x})} = \frac{1 - S(t \mid \mathbf{x})}{S(t \mid \mathbf{x})} = \frac{1 - S_0(t)}{S_0(t)} \exp \left( \mathbf{x}^\top \beta \right) \]

• All models are multiplicative.
**Definition (Survival data)**

*Right-censored survival data* consists of $n$ triplets:

- $x_i \in \mathbb{R}^d$: a $d$-dimensional feature vector.
- $y_i > 0$: observed time (time of event or time of censoring).
- $\delta_i \in \{0; 1\}$: a boolean event indicator (right censoring).
Cox’s Proportional Hazards model

- Cox PH is by far the most popular survival model.
- Coefficients can be interpreted in terms of hazard ratio:
  \[
  \frac{h(t \mid x_1, \ldots, x_j, \ldots, x_p)}{h(t \mid x_1, \ldots, x_j + 1, \ldots, x_p)} = \exp(\beta_j).
  \]
- The hazard ratio is a constant independent of time (proportional hazards assumption).
- Optimization is easy: baseline hazard function \( h_0(t) \) can be ignored until \( \beta \) has been estimated (partial likelihood optimization):
  \[
  \arg\max_{\beta} \sum_{i=1}^{n} \delta_i \left[ \mathbf{x}_i^\top \beta - \log \left( \sum_{j \in \mathcal{R}_i} \exp(\mathbf{x}_j^\top \beta) \right) \right],
  \]
  where \( \mathcal{R}_i = \{ j \mid y_j \geq t_i \} \) denotes the risk set.
**Comparable Pairs**

**Definition (Set of comparable pairs)**

\[ \mathcal{P} = \{(i, j) \mid y_i > y_j \land \delta_j = 1\}_{i,j=1,...,n} \]

---

**Diagram**

- **A**: ? (\(\delta_A = 0\))
- **B**: † (\(\delta_B = 1\))
- **C**: ? (\(\delta_C = 0\))
- **D**: † (\(\delta_D = 1\))
- **E**: ? (\(\delta_E = 0\))

**Time since enrollment in months**

- 1
- 2
- 3
- 4
- 5
- 6

\[ \mathcal{P} = \{\} \]
Definition (Set of comparable pairs)

\[ \mathcal{P} = \{(i, j) \mid y_i > y_j \land \delta_j = 1\}_{i,j=1,\ldots,n} \]

\[ \mathcal{P} = \{(\text{B}, \text{D})\} \]
Comparable Pairs

**Definition (Set of comparable pairs)**

\[ \mathcal{P} = \{(i, j) \mid y_i > y_j \land \delta_j = 1\}_{i,j=1,...,n} \]

**Diagram:**

- A: ? (\(\delta_A = 0\))
- B: ↑ (\(\delta_B = 1\))
- C: ? (\(\delta_C = 0\))
- D: ↑ (\(\delta_D = 1\))
- E: ? (\(\delta_E = 0\))

Incomparable (\(t_A > t_C\) or \(t_C > t_A\)?)

\[ \mathcal{P} = \{(B, D)\} \]
Comparable Pairs

Definition (Set of comparable pairs)

\[ \mathcal{P} = \{(i, j) \mid y_i > y_j \land \delta_j = 1\}_{i,j=1,...,n} \]

\( \mathcal{P} = \{(B, D), (C, D)\} \)
Comparable Pairs

Definition (Set of comparable pairs)

\[ \mathcal{P} = \{(i, j) \mid y_i > y_j \land \delta_j = 1\}_{i,j=1,...,n} \]

\[ \mathcal{P} = \{(B, D), (C, D)\} \]

Time since enrollment in months

A → ? (\(\delta_A = 0\))
B → † (\(\delta_B = 1\))
C → ? (\(\delta_C = 0\))
D → † (\(\delta_D = 1\))
E → ? (\(\delta_E = 0\))

Incomparable (\(t_B > t_C\) or \(t_C > t_B\)?)
Comparable Pairs

Definition (Set of comparable pairs)

\[ \mathcal{P} = \{(i, j) \mid y_i > y_j \land \delta_j = 1\}_{i,j=1,...,n} \]

\[ \mathcal{P} = \{(B, D), (C, D), (A, D), (E, D), (E, B)\} \]
• The concordance index (c index) is a measure of rank correlation between predicted risk scores \( \hat{f}(x) \) and observed time points \( y \).

• It is the ratio of correctly ordered (concordant) pairs to comparable pairs:

\[
\hat{c}_{\text{Harrell}} = \frac{1}{|\mathcal{P}|} \sum_{(i,j) \in \mathcal{P}} I(\hat{f}(x_i) < \hat{f}(x_j)).
\]

• A random model has c index 0.5, a perfect model 1.0

• Risk scores can be on any scale, only their relative ordering matters.

• c index is independent of time.
**Concordance Index**

**Example**

**Definition (Concordance index)**

\[ \frac{1}{|\mathcal{P}|} \sum_{(i,j) \in \mathcal{P}} I(\hat{f}(x_i) < \hat{f}(x_j)) \]

\[ \mathcal{P} = \{(B, D), (C, D), (A, D), (E, D), (E, B)\} \Rightarrow \hat{c} = ? \]
Definition (Concordance index)

\[
\frac{1}{|\mathcal{P}|} \sum_{(i,j) \in \mathcal{P}} I(\hat{f}(x_i) < \hat{f}(x_j))
\]

\[
P = \{(B, D), (C, D), (A, D), (E, D), (E, B)\} \Rightarrow \hat{c} = ?
\]
Definition (Concordance index)

\[
\frac{1}{|\mathcal{P}|} \sum_{(i,j) \in \mathcal{P}} I(\hat{f}(x_i) < \hat{f}(x_j))
\]

\[\mathcal{P} = \{(B, D), (C, D), (A, D), (E, D), (E, B)\} \Rightarrow \hat{c} = ?\]
Definition (Concordance index)

\[
\frac{1}{|\mathcal{P}|} \sum_{(i, j) \in \mathcal{P}} I(\hat{f}(x_i) < \hat{f}(x_j))
\]

\[\mathcal{P} = \{(B, D), (C, D), (A, D), (E, D), (E, B)\} \Rightarrow \hat{c} = ?\]
**Concordance Index**

**Example**

**Definition (Concordance index)**

\[
\frac{1}{|\mathcal{P}|} \sum_{(i,j) \in \mathcal{P}} I(\hat{f}(x_i) < \hat{f}(x_j))
\]

\[\mathcal{P} = \{(B, D), (C, D), (A, D), (E, D), (E, B)\} \Rightarrow \hat{c} = ?\]
Concordance Index

Example

Definition (Concordance index)

\[
\frac{1}{|\mathcal{P}|} \sum_{(i,j) \in \mathcal{P}} I(\hat{f}(x_i) < \hat{f}(x_j))
\]

\[\mathcal{P} = \{ (B, D), (C, D), (A, D), (E, D), (E, B) \} \Rightarrow \hat{c} = ?\]
**Concordance Index**

**Example**

**Definition (Concordance index)**

\[
\frac{1}{|\mathcal{P}|} \sum_{(i,j) \in \mathcal{P}} I(\hat{f}(x_i) < \hat{f}(x_j))
\]

\[\mathcal{P} = \{(B, D), (C, D), (A, D), (E, D), (E, B)\} \Rightarrow \hat{c} = 3/5\]
1. What is Survival Analysis?

2. Parametric Survival Models

3. Semiparametric Survival Models

4. Non-Linear Survival Models

5. Survival Analysis with Deep Learning

6. Conclusion
• Take a linear model and replace the linear predictor $x_i^T \beta$ with an unknown, more complex function $f(x)$.

• We can model $f(x)$ as an additive model by performing gradient descent in function space (gradient boosting).

• Loss function:
  ◦ Cox PH (Binder and Schumacher, 2008; Li and Luan, 2005; Ridgeway, 1999)
  ◦ AFT (Hothorn et al., 2006; Schmid and Hothorn, 2008; Wang and Wang, 2010)
  ◦ $c$ index (Benner, 2002; Mayr and Schmid, 2014)

• Base learner:
  ◦ regression tree (Breiman et al., 1984)
  ◦ componentwise least squares (Bühlmann and Yu, 2003)
• We can treat survival analysis as ranking problem (Van Belle et al., 2008).

• We want to optimize a smooth approximation of the $c$ index:

$$\min_{\mathbf{w}} \frac{1}{2} \left\| \mathbf{w} \right\|_2^2 + \gamma \sum_{(i,j) \in \mathcal{P}} \xi_{ij}$$

subject to

$$\mathbf{w}^\top \mathbf{x}_i - \mathbf{w}^\top \mathbf{x}_j \geq 1 - \xi_{ij}, \quad \forall (i, j) \in \mathcal{P},$$

$$\xi_{ij} \geq 0, \quad \forall (i, j) \in \mathcal{P}$$

• Optimization algorithm needs to be clever to avoid dependency on kernel matrix of size $O(|\mathcal{P}|^2) = O(n^4)$ (Pölsterl et al., 2015, 2016).

• Alternative models: regression with non-symmetric loss (Khan and Zubek, 2008; Shivaswamy et al., 2007), quantile regression (Eleuteri, 2008; Eleuteri and Taktak, 2012).
Faraggi and Simon (1995) proposes a multi-layer perceptron that extends the Cox PH model.

Biganzoli et al. (1998) and Liestøl et al. (1994) propose the *Partial Logistic Artificial Neural Network* that considers survival times grouped into mutually exclusive intervals and a loss based on a piecewise exponential model.
\[
\text{Cox PH loss} = \arg\max_{\beta} \sum_{i=1}^{n} \delta_i \left[ x_i^\top \beta \right] \\
- \log \left( \sum_{j \in R_i} \exp(x_j^\top \beta) \right),
\]
Loss by Faraggi and Simon

\[
\text{argmin}_\Theta \sum_{i=1}^{n} \delta_i \left[ o(x_i | \Theta) - \log \left( \sum_{j \in R_i} \exp(o(x_j | \Theta)) \right) \right]
\]
• Samples need to be sorted by observed time $y_i$ due to sum over $\mathcal{R}_i = \{j \mid y_j \geq t_i\}$.

• Batch size needs to be large, otherwise gradient is very noisy.

• Only considers \textbf{time-invariant features} (proportional hazards assumption).
The *Partial Logistic Artificial Neural Network* considers survival times grouped into mutually exclusive intervals.
The **Partial Logistic Artificial Neural Network** considers survival times grouped into mutually exclusive intervals.

\[
\begin{align*}
\tau_0 & \quad \tau_1 & \quad \tau_2 & \quad \tau_3 \\
A & \quad B & \quad C & \quad D & \quad E
\end{align*}
\]

\[
\begin{align*}
\text{Event in } k\text{-th interval?} \\
\delta_A &= 0, \quad \delta_B = 0, \quad \delta_C = 0
\end{align*}
\]

\[
\begin{align*}
\text{Time spent in } k\text{-th interval:} \\
\tilde{y}_A &= 2, \quad \tilde{y}_B = 1, \quad \tilde{y}_C = 0
\end{align*}
\]
The *Partial Logistic Artificial Neural Network* considers survival times grouped into mutually exclusive intervals.

$$\tau_0 \quad \tau_1 \quad \tau_2 \quad \tau_3$$

Event in $k$-th interval?
$$\delta_{B1} = 0, \quad \delta_{B2} = 0, \quad \delta_{B3} = 1$$

Time spent in $k$-th interval:
$$\tilde{y}_{B1} = 2, \quad \tilde{y}_{B2} = 2, \quad \tilde{y}_{B3} = 0.5$$
The *Partial Logistic Artificial Neural Network* considers survival times grouped into mutually exclusive intervals.

\[
\begin{align*}
\tau_0 \quad \tau_1 \quad \tau_2 \quad \tau_3
\end{align*}
\]

A | B | C | D | E
---|---|---|---|---
1 | 2 | 3 | 4 | 5 | 6

Time since enrollment in months

**Event in \( k \)-th interval?**

\( \delta_{C_1} = 0, \quad \delta_{C_2} = 0, \quad \delta_{C_3} = 0, \)

**Time spent in \( k \)-th interval:**

\( \tilde{y}_{C_1} = 2, \quad \tilde{y}_{C_2} = 1.5, \quad \tilde{y}_{C_3} = 0 \)
The *Partial Logistic Artificial Neural Network* considers survival times grouped into mutually exclusive intervals.

\[
\begin{align*}
\tau_0 & \quad \tau_1 & \quad \tau_2 & \quad \tau_3 \\
A & \quad \text{?} & \quad \text{?} & \quad \text{\textdagger} \\
B & \quad \text{\textdagger} & \quad \text{?} & \quad \text{?} \\
C & \quad \text{?} & \quad \text{?} & \quad \text{?} \\
D & \quad \text{\textdagger} & \quad \text{?} & \quad \text{?} \\
E & \quad \text{?} & \quad \text{?} & \quad \text{?} \\
\end{align*}
\]

Event in \( k \)-th interval?
\[
\delta_{D1} = 1, \quad \delta_{D2} = 0, \quad \delta_{D3} = 0,
\]

Time spent in \( k \)-th interval:
\[
\tilde{y}_{D1} = 2, \quad \tilde{y}_{D2} = 0, \quad \tilde{y}_{D3} = 0
\]
• A piecewise exponential model has a constant hazard rate $\lambda_l > 0$ in the $l$-th interval and has survival function

$$S(t) = \exp(-\lambda_l(t - \tau_{l-1})) \prod_{k=1}^{l-1} \exp(-\lambda_k(\tau_k - \tau_{k-1}))$$
• A **piecewise exponential model** has a constant hazard rate $\lambda_l > 0$ in the $l$-th interval and has survival function

$$S(t) = \exp(-\lambda_l(t - \tau_{l-1})) \prod_{k=1}^{l-1} \exp(-\lambda_k(\tau_k - \tau_{k-1}))$$

• Substituting the definition into the log-likelihood function of a parametric model, we obtain

$$\arg\max_{\{\lambda_1, \ldots, \lambda_L\}} \sum_{i=1}^{n} \sum_{k=1}^{L} [\delta_{ik} \log(\lambda_k) - \lambda_k \tilde{y}_{ik}]$$
• A *piecewise exponential model* has a constant hazard rate $\lambda_l > 0$ in the $l$-th interval and has survival function

$$S(t) = \exp(-\lambda_l(t - \tau_{l-1})) \prod_{k=1}^{l-1} \exp(-\lambda_k(\tau_k - \tau_{k-1}))$$

• Substituting the definition into the log-likelihood function of a parametric model, we obtain

$$\arg\max_{\{\lambda_1, \ldots, \lambda_L\}} \sum_{i=1}^{n} \sum_{k=1}^{L} [\delta_{ik} \log(\lambda_k) - \lambda_k \tilde{y}_{ik}]$$

• Finally, the parameters $\lambda_k$ are modeled by a neural network $o(x_i | \Theta)$ conditional on feature vectors $x_i$ as

$$\lambda_k(x_i) = \exp\left( \log \lambda_{0k} + w^\top o(x_i | \Theta) \right)$$
1. What is Survival Analysis?

2. Parametric Survival Models

3. Semiparametric Survival Models

4. Non-Linear Survival Models

5. Survival Analysis with Deep Learning

6. Conclusion
• I could find 24 papers using deep learning\textsuperscript{1} techniques with a loss accounting for censored event times.
• 10 use the Cox PH loss of Faraggi and Simon (1995).
• 18 have been applied to medical data.
  ○ 8 to medical images (6 of which are on histopathology images).
  ○ 4 to genomic data.
  ○ The remaining use tabular clinical data or EHR.

\textsuperscript{1}excluding work using Deep Gaussian Processes
Mobadersany et al. (2018), “Predicting cancer outcomes from histology and genomics using convolutional networks”, PNAS.

- **Objective**: Survival prediction of patients with diffuse gliomas.
- **Network** integrates information from both histology images and genomic biomarkers.
- **Uses** a modified VGG-19 architecture with loss of Faraggi and Simon.
- **Training and testing** use random sampling of patches from region of interest.
- **Genomic markers** (IDH mutation status and 1p/19q co-deletion) are integrated as input to shared FC layer.
Example 1: Histology + Genomics
Mobadersany et al. (2018)

SCNN prediction accuracy

GSCNN prediction accuracy

Concordance index

Grade
Subtype
Subtype + grade
SCNN
SCNN w/o resampling

Concordance index

Subtype + grade
GSCNN
SCNN + subtype
SCNN

p=0.307
p=4.68e-2
p=2.61e-3
p=6.55e-4

p=1.06e-2
p=4.68e-2
Example 2: Web User Return Time
Grob et al. (2018)


- **Objective**: Predict the return times of users to a website.
- Each user has a sequence of previous sessions.
- Each session has a start time and a set of features.
- Time $T$ is defined as the period between the end of a session and the beginning of the succeeding session.
- The hazard function up to the $j$-th session $h_j(t)$ is modeled as a recurrent marked temporal point process:

$$h_j(t) = \exp \left( \underbrace{v(t) h_j}_{\text{past}} + \underbrace{w(t - t_j)}_{\text{temporal}} + \underbrace{b(t)}_{\text{bias}} \right)$$
### Example 2: Web User Return Time

Grob et al. (2018)

<table>
<thead>
<tr>
<th></th>
<th>Baseline</th>
<th>Cox PH</th>
<th>RNN-MSE</th>
<th>RNN-SM</th>
</tr>
</thead>
<tbody>
<tr>
<td>RMSE (days)</td>
<td>43.25</td>
<td>49.99</td>
<td><strong>28.69</strong></td>
<td>59.99</td>
</tr>
<tr>
<td>Concordance</td>
<td>0.500</td>
<td><strong>0.816</strong></td>
<td>0.706</td>
<td>0.739</td>
</tr>
<tr>
<td>Non-returning AUC</td>
<td>0.743</td>
<td>0.793</td>
<td>0.763</td>
<td><strong>0.796</strong></td>
</tr>
<tr>
<td>Non-returning recall</td>
<td>0.000</td>
<td>0.246</td>
<td>0.000</td>
<td><strong>0.538</strong></td>
</tr>
</tbody>
</table>
1. What is Survival Analysis?

2. Parametric Survival Models

3. Semiparametric Survival Models

4. Non-Linear Survival Models

5. Survival Analysis with Deep Learning

6. Conclusion
• Time-to-event analysis is applicable across a wide range of domains.
• It is a well studied topic in statistics.
• Most classical machine learning models have been modified for time-to-event data.
• It is slowly being adapted by the deep learning community, although most of the approaches are rather naive.
• Cox PH model is surprisingly hard to beat.


