Fred Hutch Cancer Center

A Brief Introduction to Survival Analysis & (Semi-) Competing Risks

Stephen Salerno
Postdoctoral Fellow, Fred Hutch Data Science Lab

Etzioni Lab Meeting
September 26, 2023

Land Acknowledgement

Fred Hutchinson Cancer Center acknowledges the Coast Salish peoples of this land, the land which touches the shared waters of all tribes and bands within the Duwamish, Puyallup, Suquamish, Tulalip and Muckleshoot nations.

1 Survival Analysis Preliminaries

2 High-Dimensional Survival Data

3 Machine Learning & Deep Learning

4 (Semi-) Competing Risks

5 Time Permitting, Causal Inference & Beyond

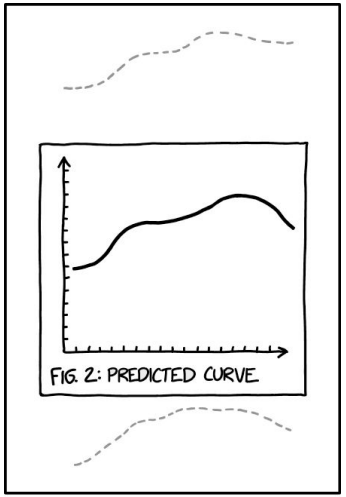
Survival Analysis Preliminaries

Background, Motivation, and Notation

Survival Analysis Preliminaries

Background and Motivation

- ***Survival Analysis:*** Area of statistics where r.v. is ***time until a specific event:***
  - ***Dependent Variable***: Time to event (i.e., failure time or survival time)
  - ***Event***: Qualitative transition from one state to another
    - (e.g., disease-free → diseased, alive → dead)
- ***Goals*** of survival analysis include:
  - Computing the ***probability of an event*** occurring by some time
  - Detecting ***associations between risk factors*** and time-to-event
  - ***Predicting*** failure times from complex data sources
- ***Survival data*** are frequently encountered in biomedical studies



Credit: https://xkcd.com/2311/

Survival Analysis Preliminaries

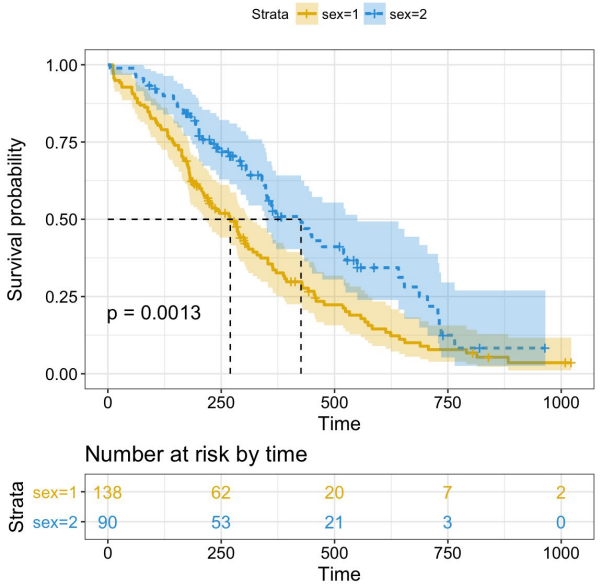
Censoring distinguishes survival outcomes from other outcomes…

- Events will be ***unobserved*** for some individuals, and events that are not observed are said to be ***censored***
- ***Right censoring*** occurs when a subject's follow-up time ends ***before the event can be observed***
- The ***fraction of events*** that are censored can be ***substantial*** (sometimes > 90%)
- Survival analysis relies on information extracted from ***all subjects***, censored or not

Statistical Challenges

Standard regression methods may not work for survival data…

- ***Failure times*** are ***not normally distributed*** (skewed to the right)
- Transformation may help; one needs to be ***cautious*** about ***interpretation*** of parameter estimates
- ***Events may not be observed*** for all subjects (censoring)
- Most crucially, some concepts or functions were ***uniquely designed for survival data***



http://www.sthda.com/english/wiki/survival-analysis-basics

Notation and Definitions

Survival Function

- Assume $T_1, \ldots, T_n$ are i.i.d. copies of T
- ***Survival Function***: $S(t) = Pr(T > t)$
    - Non-Increasing: $S(t_2) \leq S(t_1)$ for $t_2 \geq t_1$
    - $S(0) = 1$
    - $lim_{t \to \infty} S(t) = 0$

Hazard Function

- ***Formal Definition***: $\lambda(t) = lim_{u \to 0} 1/u \, Pr(t \leq T < t + u \mid T \geq t)$
- ***Working Definition***: $\lambda(t) = P(T = t \mid T \geq t)$

Common Notation:

- i : Subject (i = 1, . . . , n)
- t : Time
- T_i : Survival time of subject i
- C_i : (Right) censoring time
- $Y_i = T_i \wedge C_i$: Observed time
- $\delta_i = I(T_i \leq C_i)$: Event indicator
- X_i : Covariate vector

Assume *independent censoring*:

- $T_i \perp C_i$ given X_i

Definitions, Continued

Hazard Function

- "***Rate of death*** (time *t*) among subjects alive at time *t*"
- ***Difference*** between *λ(t)* and the ***density function,*** *f(t),* is ***conditionality*** on $T \geq t$
- They are ***connected*** with $\lambda(t) = f(t)/S(t)$
- With ***independent censoring***, death rate equals ***observed death rate*** among those ***alive*** and ***uncensored*** at *t*

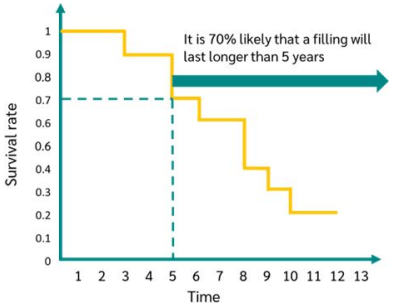
  $Pr(T = t \mid T \geq t) = P(T = t, \delta = 1 \mid T \geq t)$

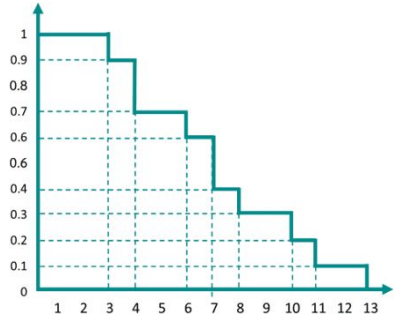
  implying we can use data ***observed*** at *t* (even though it is a ***filtered version*** of the true data) to ***estimate*** λ(t)

Definitions, Continued

Cumulative Hazard Function and Kaplan-Meier Estimator

- **Cumulative Hazard Function:** $\Lambda(t) = \int_0^t \lambda(u)du$
- Cumulative hazard and survival have an ***important connection***: $S(t) = e^{-\Lambda(t)}$
- ***Note***: $e^{-\lambda(u)du} = 1 - \lambda(u)du$
- Taking product (between 0 and t) on both sides: $\hat{S}(t) = e^{-\Lambda(t)} = \prod_{u \in (0,t]} \{1 - \lambda(u)du\}$
- The basis for the ***Kaplan-Meier***, or 'product limit' estimator: $S(t) = \prod_{tj \le t} \{1 - Dj / Yj\}$
  - Y_j = Number "at risk" and D_j = Number of failures at $t = t_j$
  - $\hat{S}(t)$ is a ***step function***, with jumps only at ***failure times*** and drops to 0 if and only if the ***last observation*** time is a ***death***





https://datatab.net/tutorial/kaplan-meier-curve

Low-Dimensional Survival Approaches

Parametric AFT Regression Models

- General formulation for the ***accelerated failure time*** model is: $log\ T_i = \beta^T X_i + e_i$
- ***Log-transformation*** ensures parameter space of β is ***unconstrained***
- Distribution of e_i ***induces*** distribution of T_i

Distribution of e_i	Induced distribution of T_i
Normal distribution	Log-normal distribution
Extreme value distribution	Weibull distribution
Logistic distribution	Log-logistic distribution

e_i is the residual term, and T_i is the survival time.

Low-Dimensional Survival Approaches

Semi-Parametric AFT Regression Models

- With the distribution of e_i ***unspecified***, models are ***semi-parametric*** and MLE is ***difficult***
- ***Buckley-James (1979)***: Estimate e_i's distribution, impute censored outcomes, conduct least square estimation
  - ***Define*** $e_i(b) = log(Y_i) - b^T X_i$
  - Compute ***Kaplan-Meier estimate*** $\hat{S}(\cdot, b)$ based on $(e_i(b), \delta_i), i = 1, \ldots, n$
  - ***Impute*** $T_i^*(b) = log(Y_i) + (1 - \delta_i)\int_{ei(b)}^{\infty} \hat{S}(s, b)\, ds \,/\, \hat{S}(e_i(b), b)$
  - ***Solve*** for β: $\sum_{i=1}^{n}(X_{ij} - X_j)(T_i^*(\beta) - \beta^T X_i) = 0$; for $j = 1, \ldots, p$

Low-Dimensional Survival Approaches

Cox Proportional Hazards Model

- ***Cox (1972)*** is among the ***most cited papers*** in science and ***dominates survival analysis*:**
  - A novel ***hazard model***
  - A novel method for estimating parameters with ***partial likelihood***
- ***Cox Proportional Hazards (PH) Model:*** Given a covariate vector, X_i

  $\lambda_i(t) = \lambda(t \mid X_i) = \lambda_0(t)\, exp\{\beta^T X_i\}$
- $\lambda_0(t) = \lambda(t \mid X_i = 0)$ is (unspecified) ***baseline hazard***
- Model is ***semi-parametric***: parametric covariate effects $(\beta^T X_i)$ and non-parametric baseline hazard

Low-Dimensional Survival Approaches

Partial Likelihood

- ***Partial likelihood*** contributions occur only at ***death times***. For subject *i*, who is observed to die:

$$PL_i(\beta) = Pr(\text{subject } i \text{ dies at } Y_i \mid \text{someone dies at } Y_i, R(Y_i))$$

$$= Pr(i \text{ dies at } Y_i \mid i \text{ survives until } Y_i) / \sum_{j \in R(Yi)} Pr(j \text{ dies at } Y_i \mid j \text{ survives until } Y_i)$$

- where *R(t)* is set of subjects at risk at time *t*, then

$$PL_i(\beta) = \lambda_i(Y_i) / \sum_{j \in R(Yi)} \lambda_j(Y_i) = \exp\{\beta^T X_i\} / \sum_{j \in R(Yi)} \exp\{\beta^T X_j\}$$

- Baseline hazard ***cancels out***, so this can ***estimate β*** without estimating $\lambda_0(t)$ (***no intercept*** included in model)

Low-Dimensional Survival Approaches

Partial Likelihood, Continued

- Taking ***product across all subjects***,

$$PL(\beta) = \prod^{n}_{i=1}\{exp\{\beta^{T}X_i\} / \sum^{n}_{j=1} I(Y_j > Y_i)\, exp\{\beta^{T}X_j\}\}^{\delta i}$$

- Not a ***usual likelihood***
- Called ***partial likelihood*** as it only uses ***part of data*** that involves β (i.e., rank of the observed times)
- ***Estimator*** has many of the ***MLE properties*** (e.g., consistency, asymptotic normality)

Low-Dimensional Survival Approaches

Censored Quantile Regression (Powell, 1984, 1986)

- For any $\tau \in (0, 1)$, the τ-***quantile*** is a value at or below which a τ-fraction of population lies
- Denote the τ-th conditional quantile of $\tilde{T} = \log T$ given X by $Q_{\tilde{T}}(\tau \mid X)$; the ***model stipulates***

  $Q_{\tilde{T}}(\tau \mid X) = \beta_0(\tau) + \beta_X(\tau)^T X$
- ***Note***: AFT is a special case when $\beta_X(\tau)$ does not depend on τ
- To deal with ***censoring***, you can ***reweight*** the censored observations (Portnoy, 2003), use Efron (1967)'s ***redistribution-of-mass*** idea, or use ***martingale-based estimating equations*** (Peng and Huang, 2008)

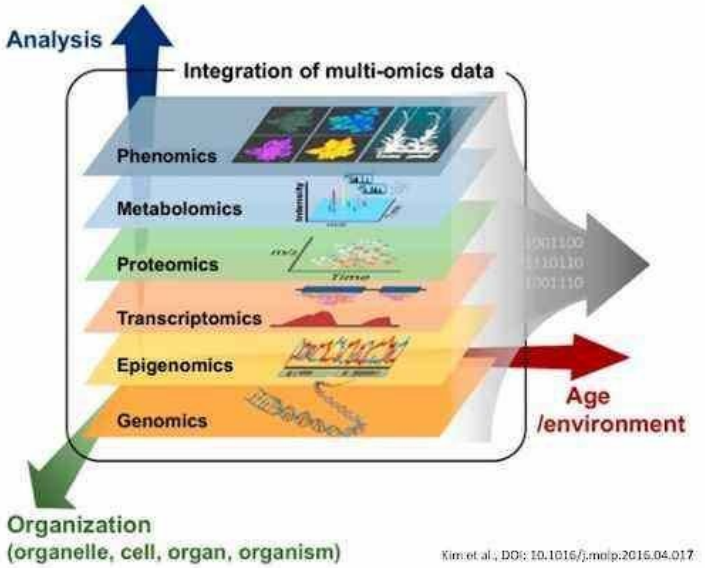
High-Dimensional Survival Data

Regression Strategies & Regularization

High-Dimensional Survival Data

What happens when p > n?

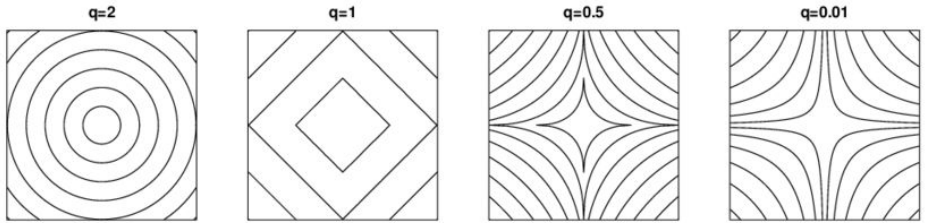
- ***High-Dimensional Data***: The number of variables, *p*, exceeds the number of subjects, *n*
- Generally unwise to build prediction models with ***all features*** because of ***overfitting***
- A useful strategy is to select “vital” features under the ***sparsity assumption*** that most of the features are “unimportant” (Friedman et al, 2001)
- Research focuses on how to perform ***variable selection*** and ***estimation*** simultaneously



Regularized Cox Models

Regularization with l_q *Penalty*

- Add $||\beta||_q = (\sum^p_{j=1} |\beta_j|^q)^{1/q}$, $q \geq 1$ to objective functions to ***enforce sparsity***
- When $0 \leq q < 1$, the ***penalty is non-convex***, making optimization challenging
- Various q and ***combinations of norms*** give a ***variety of regularized methods*** (e.g., elastic net)
- $||\beta||_0 = \sum^p_{j=1} I(\beta_j \neq 0)$: ***Number of non-zero coefficient***s, corresponding to ***AIC or BIC***

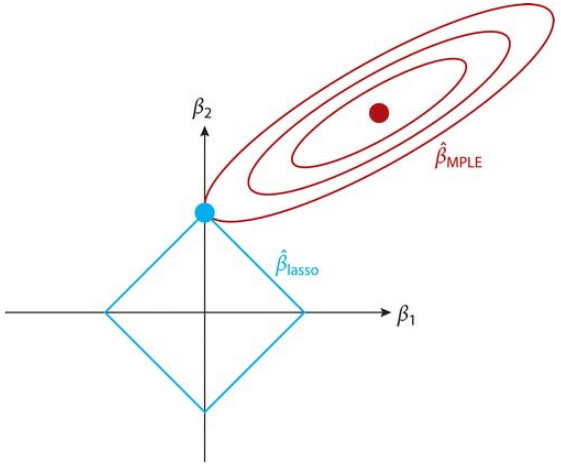


Basu, Tathagata & Einbeck, Jochen & Troffaes, Matthias. (2021). Uncertainty Quantification in Lasso-Type Regularization Problems. 10.1007/978-3-030-60166-9_3.

Cox LASSO

Least Absolute Shrinkage and Selection Operator (Tibshirani, 1997)

- $q = 1$ leads to the ***LASSO estimator***
- $argmin_{\beta}\{-\mathscr{L}(\beta) + \lambda||\beta||_1\}$, with $\lambda > 0$ where $\mathscr{L}(\beta) = log\ PL(\beta)$
- Originated from LASSO for ***linear regression***
- Performs feature selection and estimation ***simultaneously***
- Has a ***Bayesian interpretation***, i.e., β has a “double exponential” prior
- select λ via ***cross-validation***



Salerno S, Li Y. 2023
Annu. Rev. Stat. Appl. 10:25–49

Other Choices of Penalty Terms

Method	Penalty	Constraints
Ridge	$\|\|\boldsymbol{\beta}\|\|_2^2$	NA
Lasso	$\|\|\boldsymbol{\beta}\|\|_1$	NA
Elastic net	$\alpha\|\|\boldsymbol{\beta}\|\|_1 + (1-\alpha)\|\|\boldsymbol{\beta}\|\|_2^2$	$0 < \alpha < 1$
Adaptive lasso	$\sum_j w_j \beta_j$	$w_j \geq 0$
SCAD ($Pen(\|\beta\|)$)	$Pen'(\|\beta\|) = \eta \left\{ \mathbb{I}(\|\beta\| \leq \eta) + \frac{(\alpha\eta - \|\beta\|)_+}{(\alpha-1)\eta} \mathbb{I}(\|\beta\| > \eta) \right\}$	$\alpha > 2, \eta > 0$
Group lasso	$\sum_g \|\|\boldsymbol{\beta}_g\|\|_1$	$\boldsymbol{\beta}_g = (\beta_{g1}, \ldots, \beta_{gj_g})^T$
Fused lasso	$\sum_j \|\beta_j\|$ and $\sum_j \|\beta_j - \beta_{j-1}\|$	NA

Abbreviations: Lasso, least absolute shrinkage and selection operator; NA, not applicable; SCAD, smoothly clipped absolute deviation.

Dantzig Selector for AFT Models

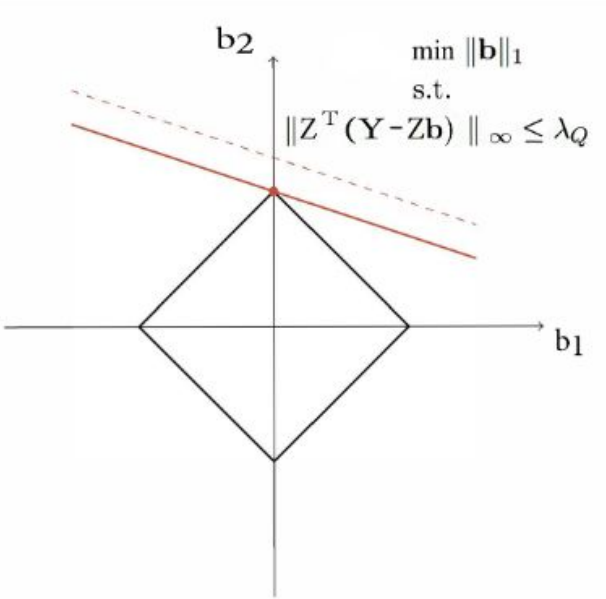
Li, Dicker, and Zhao (2014)

- Consider an ***AFT model*** with $p > n$
- Likelihood functions involve ***infinite dimensional parameters*** and LASSO estimation is difficult
- Can use ***Buckley-James imputation*** to express AFT estimation as a ***least squares*** estimation problem
- Then apply ***Dantzig selector*** to the least squares estimation:

  min $||\beta||_1$, subject to $||X^T P_n(T^*(\beta) - X\beta)||_\infty \leq \lambda_Q$

  where P_n is a centering matrix and $\lambda_Q > 0$ is a tuning parameter

- ***Weighted Dantzig selector*** has model selection consistency and oracle properties for the estimates



Beyond This …

Some additional important topics in high-dimensional survival

- ***Feature screening*** for ***ultrahigh-dimensional*** covariates (p >> n)
  - Principled sure independence screening
  - Conditional screening and forward regression
  - Non-parametric screening

- ***Inference*** with high-dimensional covariates - two main categories:
  - Post-selection inference
  - Debiased LASSO

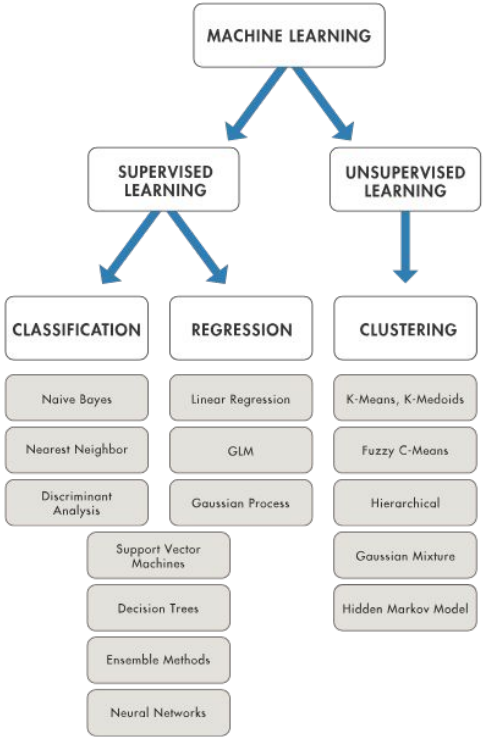
Machine Learning & Deep Learning

A Brief Survey of Approaches

Machine Learning Techniques

Can readily be applied to survival data

- Significant work has gone into the development of ***machine learning algorithms*** for survival data
- ***Non-parametric*** learning approaches handle ***non-linear relationships*** or ***higher-order interaction***s
- ***Improve accuracy in prediction*** for survival outcomes
    - Support Vector Machines
    - Tree Based Methods
    - Ensemble Learners
    - Artificial Neural Networks
    - And More!

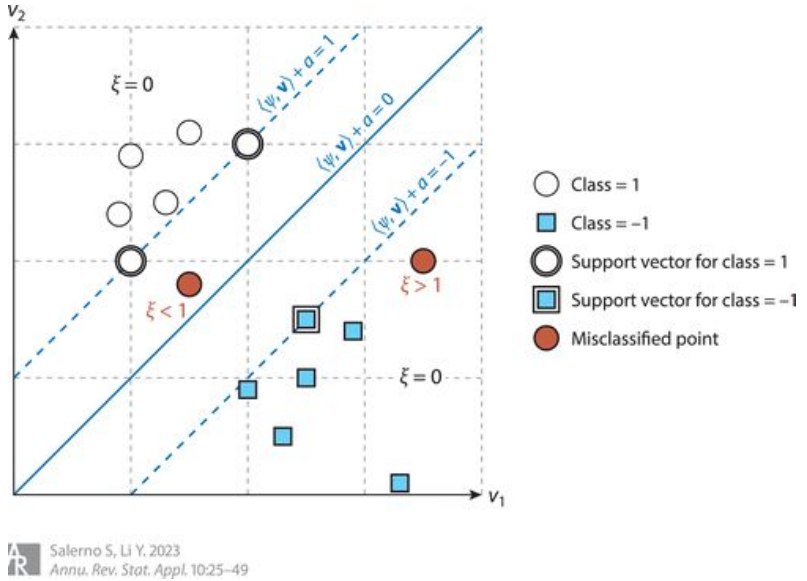


https://www.mathworks.com/discovery/machine-learning.html

Support Vector Machines

And Survival SVM

- SVMs fall under the ***supervised learning family*** (Vapnik et al., 1995; Noble, 2006)
- Goal is to find a ***hyperplane*** that provides ***maximal separation*** between groups
- Maps original predictors to a ***higher-dimensional space*** where outcomes can be ***distinguished***
- ***Survival SVMs*** use ***rank concordance*** between ***predicted and observed times*** among ***comparable individuals***
- See Van Belle et al. (2011) and Pölsterl et al. (2015)

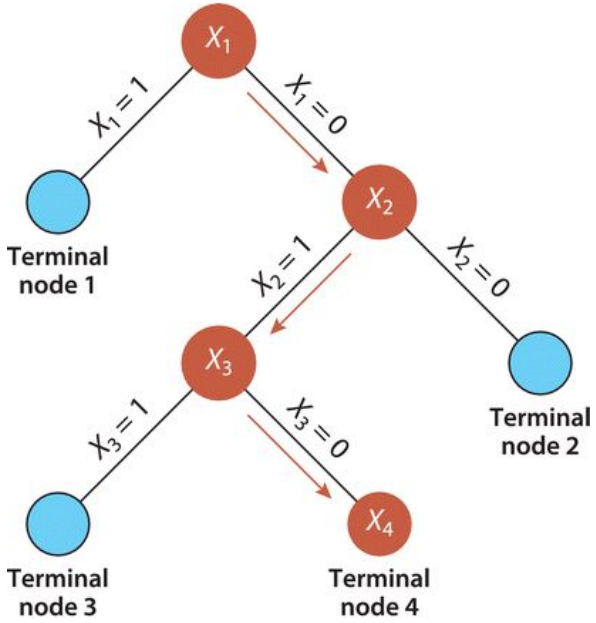


Salerno S, Li Y. 2023
Annu. Rev. Stat. Appl. 10:25–49

Tree-Based Methods

Decision Trees

- SVMs do not have ***clear interpretation*** for classifying data
- Decision trees classify based on ***hierarchical relationships*** between predictors (recursive partitioning)
- ***Survival trees*** (Gordon & Olshen, 1985; Ciampi et al., 1986, 1987) establish splitting criteria based on:
    - Log-rank test statistic
    - Likelihood ratio test statistic
- Terminal nodes split data into distinct groups, which are ***more homogeneous*** in terms of ***survival estimates***

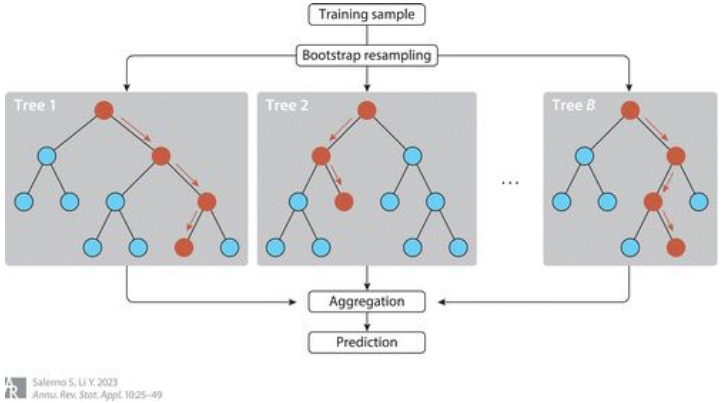


Salerno S, Li Y. 2023
Annu. Rev. Stat. Appl. 10:25–49

Ensemble Learners

Bootstrap Aggregation (Bagging)

- ***Ensemble learners*** overcome tree instability by using techniques such as ***bagging***, ***boosting***, ***random forests***
- ***Bagging*** resamples data ***with replacement***, trains individual survival trees, and ***combines their results***
- Bagging for ***survival trees*** (Hothorn et al., 2004) aggregates by averaging ***survival predictions***
  - Each survival tree is grown so that every ***terminal node*** has ***enough events***
  - Predict ***survival function*** node-wise at terminal nodes

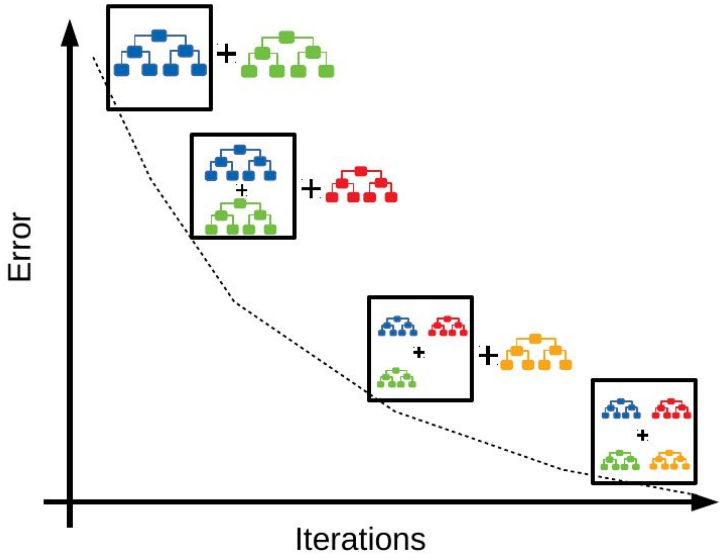


Salerno S, Li Y. 2023
Annu. Rev. Stat. Appl. 10:25–49

Ensemble Learners

Gradient Boosting

- ***Boosting***, like bagging, trains a series of weak learners (individual trees) to create an ***ensemble***
- Boosting is done ***sequentially***, updating weights placed on learners ***iteratively***
    - Hothorn et al. (2006) proposed a gradient boosting algorithm for ***survival settings***
    - Prediction is made based on ***previous prediction*** and an ***additional weak learner***
- ***Bagging*** individual weak learners such as survival trees are trained ***independently*** and ***in parallel***
- ***Boosting*** is better for ***low variability*** and ***high bias***, ***bagging*** is better for ***high variability*** but ***low bias***



https://www.analyticsvidhya.com/blog/2022/11/top-10-interview-questions-on-gradient-boosting/

Ensemble Learners

Random Forests

- Like bagging, ***random forests*** aggregate predictions from individual trees on ***bootstrap resampled datasets***
- However, random forests ***randomly select*** a ***subset of features***, say p′ < p features, for each tree
    - ***Reduces correlations*** among individual trees, leading to gains in ***accuracy*** (Breiman 2001)
    - The choice of p′ is problem-specific, so it can also be viewed as a ***tuning parameter***
    - Ishwaran et al. (2008) averages the predicted ***cumulative hazard functions*** into an ensemble prediction
- Ishwaran et al. (2011) extended random survival forests to ***high dimensions*** by incorporating regularization
- Ishwaran & Lu (2019) provided ***standard errors*** and ***confidence intervals*** for ***variable importance***
- Steingrimsson et al. (2019) proposed ***censoring unbiased*** regression trees and ensembles

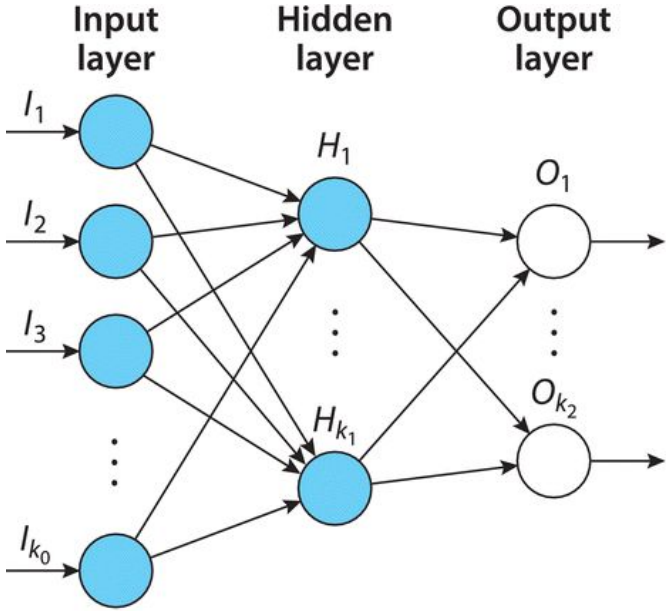
Deep Learning

Artificial Neural Networks

- Mirrors how the ***human brain*** functions (Rosenblatt 1958)
- ***Neurons*** connected in ***network*** as weighted sum of inputs through ***affine transformations*** and ***nonlinear activations***
- A fully connected, ***feed-forward artificial neural network*** is made up of L layers, with kl neurons in the lth layer
- Predictions are made based on an L-fold composite function, $f_L \circ f_{L-1} \circ \cdots \circ f_1(\cdot)$, *with* $(g \circ f)(\cdot) = g(f(\cdot))$, with

$$f_l(\mathbf{v}) = \sigma_l(\mathbf{W}_l\mathbf{v} + \mathbf{b}_l) \in \mathbb{R}^{k_l},$$

- where W_l is a ***weight matrix***, b_l is a ***bias vector***, and $\sigma(\cdot)$ is an ***activation function***



Salerno S, Li Y. 2023
Annu. Rev. Stat. Appl. 10:25–49

Deep Learning

Variations for Survival Prediction

- Faraggi & Simon (1995): Fully connected, feed-forward neural network to extend the Cox model to nonlinear predictions.
- Other feed-forward neural networks (Liestbl et al., 1994; Brown et al., 1997; Biganzoli et al., 1998; Eleuteri et al., 2003) use ***survival status*** as a ***training label*** and output predicted ***survival probabilities***
- Further developments in ***Bayesian networks*** (Lisboa et al., 2003; Bellazzi & Zupan, 2008; Fard et al., 2016)
- ***Convolutional neural networks*** (Katzman et al., 2017, 2018; Ranganath et al., 2016; Yao et al., 2017)
- ***Recurrent neural networks*** (Yang et al. 2018).

(Semi-) Competing Risk Data

… and getting closer to the truth

Motivation and Background

What happens in practice versus what happens in reality

- ***Mortality*** is often the ***endpoint of choice*** for clinical trials and cohort studies
- Many ***survival processes*** in real applications involve ***multiple competing events***, however
    - Mortality is often considered ***without other events***
    - Or composite endpoints such as ***progression-free survival*** are used
- This can lead to ***biased results*** if progression is not accounted for (Jazić et al., 2016)
- If ***progression*** is the endpoint and a ***strong precursor*** to death, need to account for ***dependent censoring***

Motivation and Background

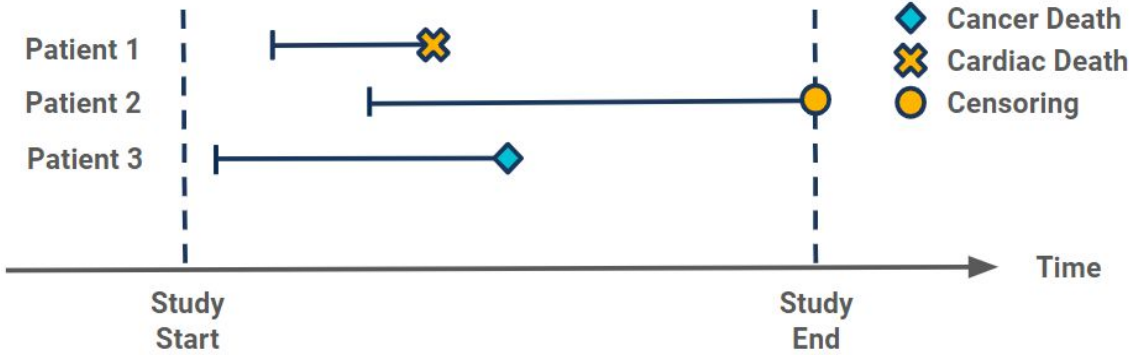
Competing and Semi-Competing Risks

- Risk prediction in settings of ***multiple endpoints*** is an up-and-coming field with many ***potential developments***
- We focus on two common settings, ***competing*** and ***semi-competing risks***

Competing Risks

Recognizing Multiple, Competing Events

- In a ***competing risk setting***, observing an event type, labeled by $c \in \{1, \ldots K\}$, effectively ***eliminates*** the chance of ***observing other event types*** happening to the same individual (e.g., cancer-related vs. cardiac death)



Methods for Competing Risks

Cause-Specific vs. Subdistribution Hazards

- Both are ***commonly used*** statistical metrics, but ***target different counterfactual scenarios***
- ***Cause-Specific Hazards***: Describe the risk under ***hypothetical elimination*** of competing events
- ***Subdistribution Hazards***: Describe the observable risk ***without elimination*** of any competing events by

$$\lambda_c(t) = \lim_{\Delta \to 0} \frac{\Pr(t \leq T_i < t + \Delta, C_i = c \mid T_i \geq t \cup (T_i < t \wedge C_i \neq c))}{\Delta} = \frac{dF_c(t)/dt}{1 - F_c(t)},$$

- ***Instantaneous risk of failure*** from event type *c* among those who have not experienced ***this type of event***
- $F_c(t) = Pr(T_i < t, C_i = c)$ is the ***cumulative incidence function***

Subdistribution Hazards Model

High-Dimensional Methods based on Fine & Gray (1999)

- The ***subdistribution hazard model*** (Fine & Gray 1999) links a subdistribution hazard function to covariates

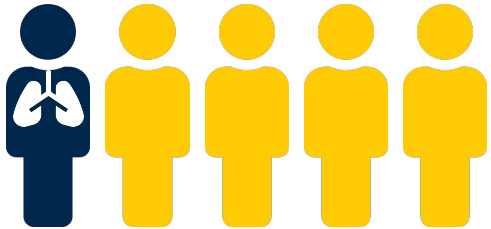
$$\lambda_c(t|\boldsymbol{X}_i) = \lambda_{0c}(t)\exp(\boldsymbol{X}_i^T\boldsymbol{\beta}),$$

- Useful for predicting the ***probability of a type of event by a given time***, which reflects an individual's actual risks and prognosis (Lau et al., 2009; Koller et al. 2012)
- ***Regularized*** models for variable selection (Kawaguchi et al. 2019, Ha et al. 2014, Ahn et al. 2018)
- One-step ***debiased lasso estimator*** for performing ***inference*** (Hou et al., 2019)
- For ***prediction***, several deep learning methods for competing risks have been proposed based on CIFs:
  - ***DeepHit*** (Lee et al., 2018): Multi-task network based on the joint distribution of first hitting times
  - ***Dynamic DeepHit*** (Lee et al., 2019): Incorporates longitudinal information for dynamic predictions
  - ***Other Approaches***: DeepCompete (Aastha & Liu 2020), hierarchical approaches (Tjandra et al., 2021)

Some of my work…

Motivation

Lung Cancer Prognosis

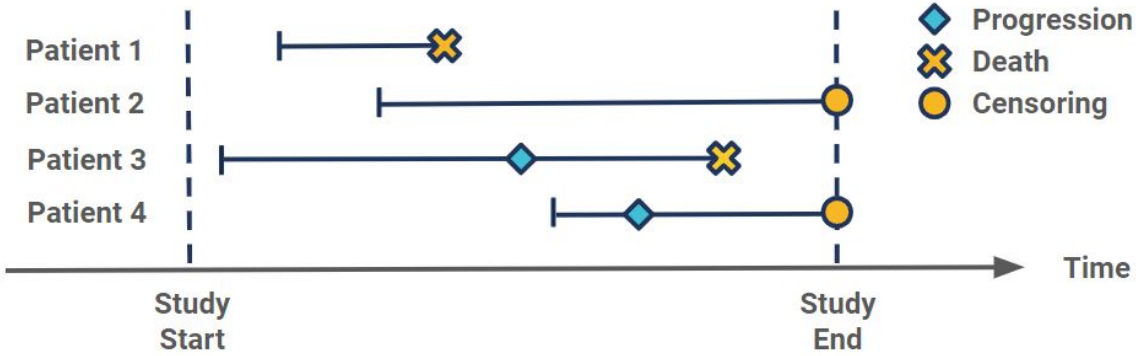


- Approximately ***1 in 5 cancer deaths*** are attributed to ***lung cancer***
- ***5-year survival rate*** of ***1 in 5*** (Bade and Cruz, 2020)
- Prognosis depends on ***individualized risk factors*** (Ashworth et al., 2014)
- Motivation comes from the ***Boston Lung Cancer Study*** (BLCS), a large cancer epidemiology cohort examining:
  - ***Complex mechanisms*** governing ***relationships*** between risk factors
  - ***Efficacy*** of treatments
  - Methods for accurately ***predicting*** survival

Semi-Competing Risks

Additional complications when considering terminal vs. non-terminal events

- ***Semi-Competing***: Occurrence of non-terminal event is subject to occurrence of terminal event, not vice versa
- As the non-terminal event (e.g., cancer progression) is often a ***strong precursor*** to the terminal event (death), the terminal event may ***informatively censor*** the non-terminal event (Jazić et al., 2016)



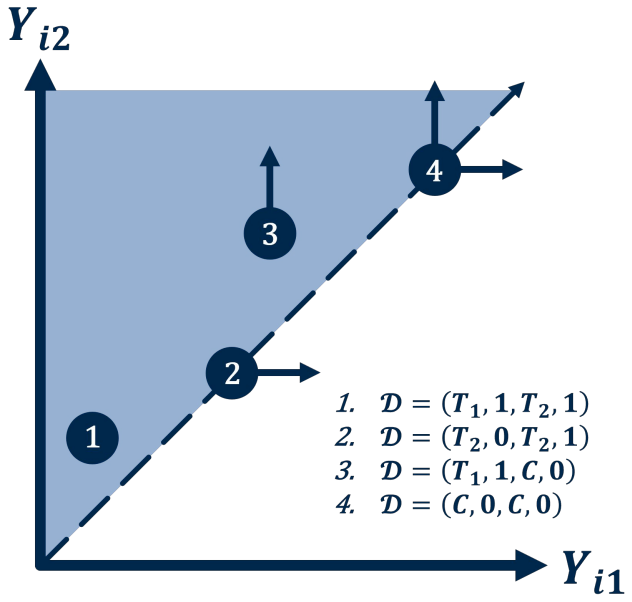
Semi-Competing Risks

Observable Data

- Now consider T_{i1}, T_{i2}, and C_i as the time to ***progression***, ***death***, and ***censoring*** for the *i*th individual. We observe:

  $D = \{(Y_{i1}, \delta_{i1}, Y_{i2}, \delta_{i2}, X_i); i = 1, \ldots, n\}$

- ***Observed Data*** Definitions:
  - $Y_{i2} = min(T_{i2}, C_i)$
  - $\delta_{i2} = I(T_{i2} \leq C_i)$
  - $Y_{i1} = min(T_{i1}, Y_{i2})$
  - $\delta_{i1} = I(T_{i1} \leq Y_{i2})$
  - X_i = Covariates



Observable data - arrows in the direction of censoring

Handling Semi-Competing Risks

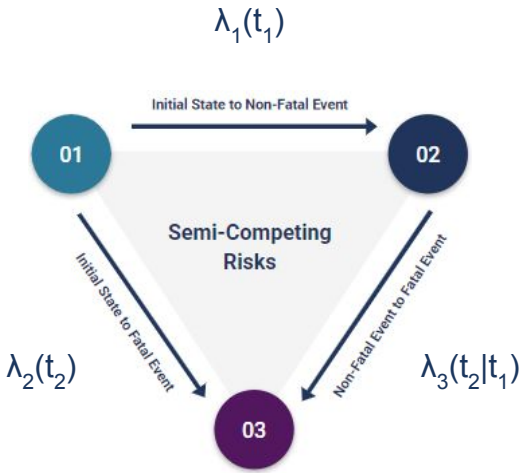
Illness-Death Model Framework

- Consider the ***Illness-Death Model***, a ***compartment-type model*** for the rates at which individuals ***transition*** between states (Andersen et al. 2012)

$$\lambda_1(t_1 \mid \gamma_i, x_i) = \gamma_i \times \lambda_{01}(t_1) \times \exp\{h_1(x_i)\}; \quad t_1 > 0$$

$$\lambda_2(t_2 \mid \gamma_i, x_i) = \gamma_i \times \lambda_{02}(t_2) \times \exp\{h_2(x_i)\}; \quad t_2 > 0$$

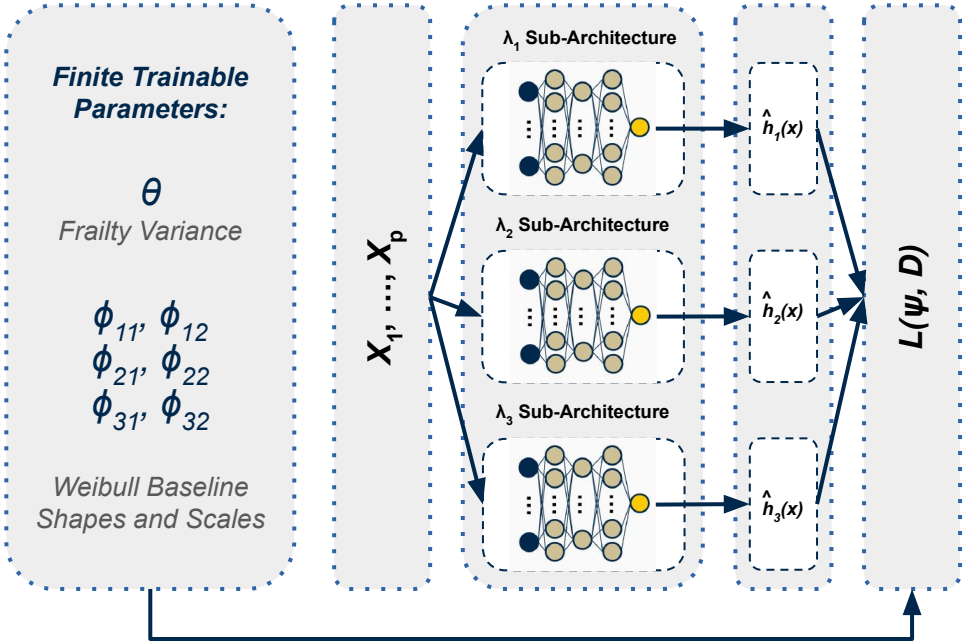
$$\underbrace{\lambda_3(t_2 \mid t_1, \gamma_i, x_i)}_{\text{Hazard Function}} = \underbrace{\gamma_i}_{\text{Frailty}} \times \underbrace{\lambda_{03}(t_2 - t_1)}_{\text{Baseline Hazard}} \times \underbrace{\exp\{h_3(x_i)\}}_{\text{Risk Function}}; \quad t_2 > t_1 > 0$$



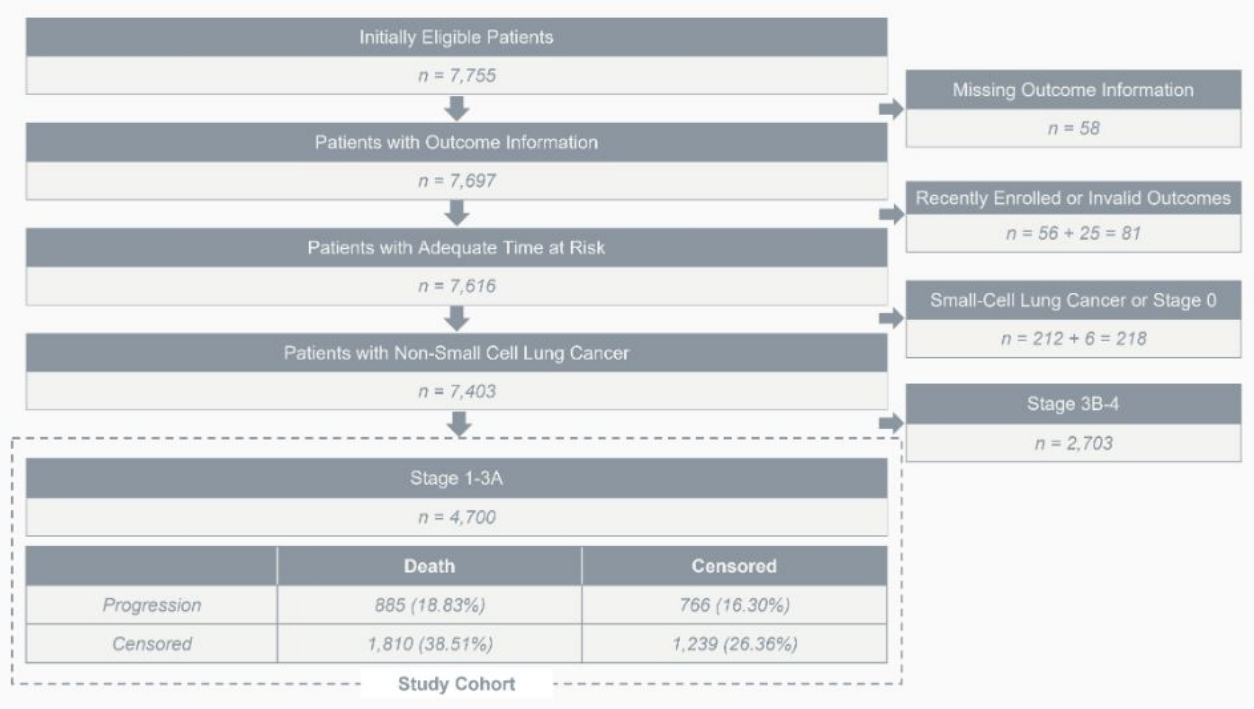
- ***Hazard = Frailty x Baseline Hazard x Risk Function***

Deep Learning for Semi-Competing Risks

Use deep learning to estimate the risk functions for each hazard (i.e, state transition)

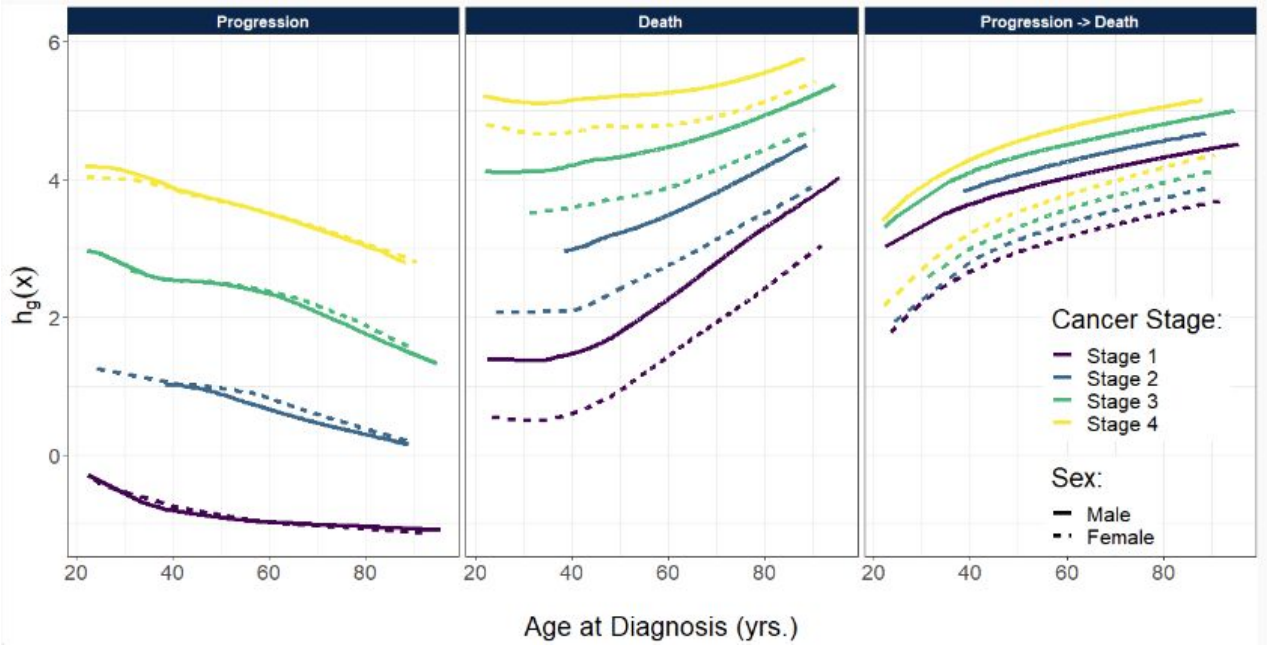


Boston Lung Cancer Study



Estimated Log-Risk Functions

Some Interactions and Non-Linear Relationships



BLCS Analysis Results

Some Conclusions

- Estimated θ to be 3.15 (bootstrapped 95% CI: 3.02-3.29), suggesting ***progression is correlated with death***
- Our approach reveals a ***nonlinear effect*** of ***age*** that differs by type of event transition, cancer stage, and sex
    - Younger age, advanced stage had higher hazards for progression
    - Older age related to higher hazards of death
    - Male patients more likely to die than female after diagnosis or progression

Causal Inference & Beyond

Time permitting, a detour into causality

Focusing on Progression

Some initial thoughts…

- ***Mortality*** is often the primary endpoint for studying ***treatment efficacy***
- ***Non-fatal events*** impact ***treatment decisions*** and ***disease management***
- ***Cancer recurrence*** alters remaining available treatments, making it an ***important endpoint*** in patients who have undergone curative treatment
- Understanding ***patient-specific*** treatment efficacy is crucial when considering ***individualized care approaches***
- ***Observational studies*** such as BLCS provide a wealth of information on individualized risk factors

Causal Inference

Potential Outcomes Framework

- Z_i: ***Causal variable of interest*** (e.g., binary treatment indicator);
    - $Z_i = 1$ for surgical resection and $Z_i = 0$ for other first-line treatment options
- X_i: p-vector of additional ***confounding variables***
- T_{i1}: Time to recurrence (censored by death at T_{i2} or independently at C_i)
- A ***potential outcomes framework***:
    - T^z_{i1}: ***Potential*** time to recurrence had *i*th patient received treatment $z \in \{0, 1\}$
    - ***Causal inference*** estimates the 'true' effect of an intervention on time to disease recurrence by comparing T^1_{i1} vs T^0_{i1}

Key Assumptions

Counterfactual Framework

1. ***Consistency***: $T_{i1} = T^{Zi}_{i1}$ almost surely

   An individual's potential outcome under their assigned treatment group is the outcome is observed

2. ***Positivity***: $Z_i \in \{0, 1\} \ \forall X_i$

   Every individual has a non-zero probability of being assigned to either treatment group

3. ***No Interference***: T^z_{i1} is unaffected by the value of z for another subject, j

   The potential outcomes of one individual are not affected by the treatment assignment of other individuals

4. ***Exchangeability***: $T^1_{i1}, T^0_{i1} \perp Z_i | X_i$

   There is no unmeasured confounding

Additionally, assume ***non-informative censoring***, i.e., $T_{i1} \perp C_i | Z_i, X_i$

Causal Quantity of Interest

Average Treatment Effect (ATE)

- A ***causal quantity of interest*** is the ***ATE*** - the expected difference in potential outcomes
- For time-to-recurrence, consider the average causal ***risk difference*** at time *t*:

$$E[I(T^1_{i1} > t) - I(T^0_{i1} > t)]$$

- With a ***consistent estimator*** of $S_1(t)$, we can use ***direct standardization*** to estimate the ATE on disease recurrence via an ***S-Learner***

Some Challenges

We are hoping to address

- Semi-competing risks complicate causal inference by introducing ***dependent censoring***, where death ***precludes*** disease recurrence
- Approaches for semi-competing risks necessitate ***complicated loss functions***, ***strong assumptions***, or techniques such as ***principal stratification***
- Parametric or semi-parametric methods are limited in their ability to model ***complex relationships*** and ***interactions*** between covariates
- Integrating ***causal inference*** and ***machine learning*** has shown great promise, but little work has been done in settings with ***dependent censoring***

Our Proposal

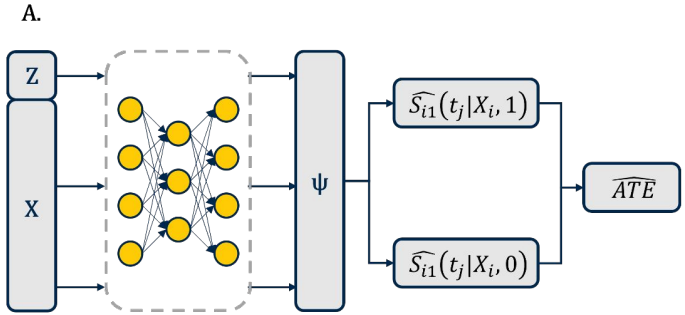
A Three-Stage Approach

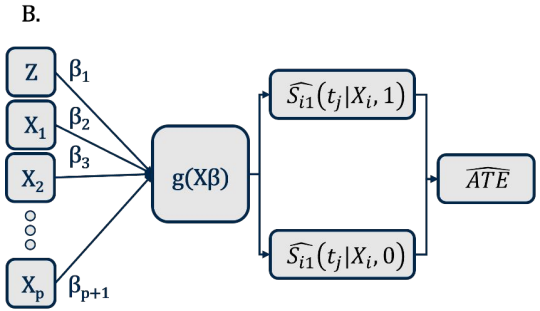
1. We estimate the ***survival function*** for time-to-recurrence based on a ***Clayton copula*** representation of the joint survival function
2. We estimate ***pseudo-survival probabilities*** at fixed time points as ***target values*** to circumvent the need for a complex loss function
    - Facilitates the development of causal estimators for such targets
    - Shown to be consistent and do not impose common assumptions such as proportional hazards across all time points
3. We relate our pseudo-outcomes to our ***causal variable of interest*** and additional confounders in a ***deep neural network***
    - Estimate survival average causal effect estimates with an S-learner

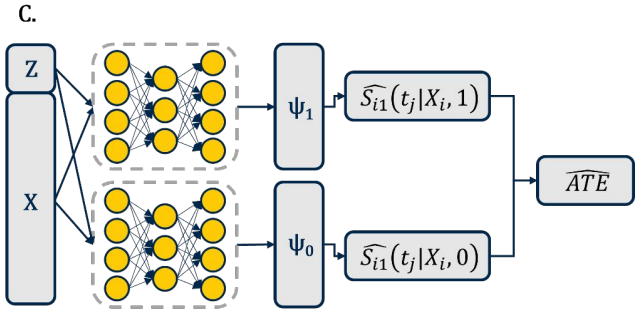
Causal Learner Illustrations

Different Computational Graphs

A. S-Learner
B. Q-Model
C. T-Learner

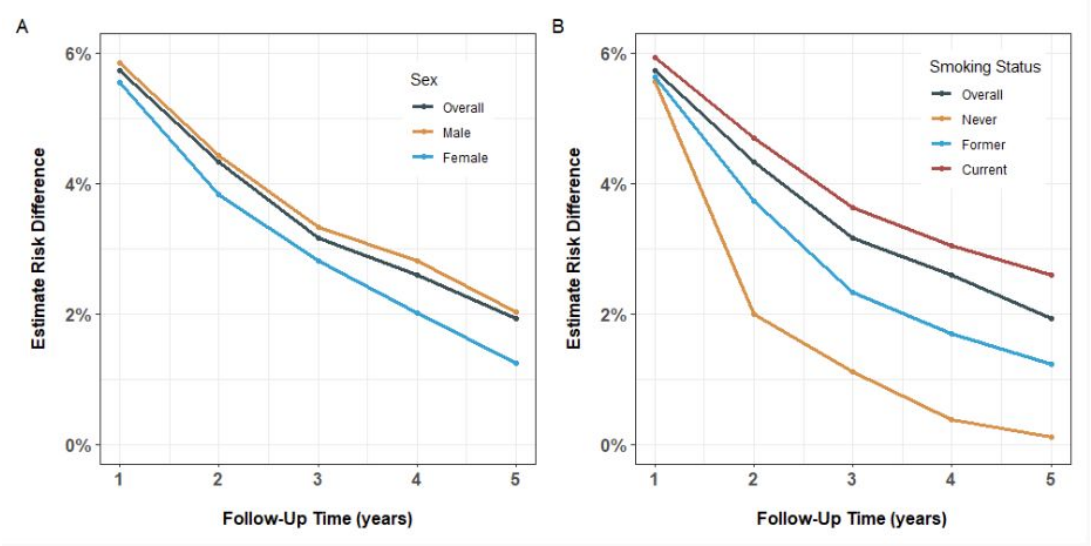






BLCS Results

Estimated average causal difference in the risk of recurrence between surgery and other first-line treatments among patients with stage 1-3A non-small cell lung cancer, over time and (A) stratified by sex; (B) stratified by smoking status



BLCS Results

Preliminary Conclusions

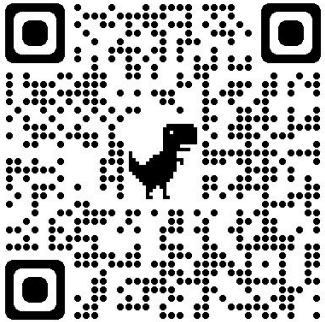
- ***Overall difference*** in risk of recurrence between first-line therapies ***attenuates*** over time: 5.7% at 1 year vs. 1.9% at 5 years
- ***Stratified by sex*** risk difference is ***slightly higher*** among ***male patients***
  - Among ***males***, 1-year difference of 5.9% attenuating to 2.0% at 5 years
  - Among ***females*** patients, 1-year difference of 5.6% attenuating to 1.3% at 5 years
- Larger treatment differences were observed when stratifying by ***smoking status***
  - Differences ***slightly higher*** among ***current smokers*** (one year difference = 5.9% vs. 2.5% at five years)
  - Differences were ***less*** among ***former*** (range: 5.6% to 1.2%) and ***never smokers*** (range: 5.6% to 0.1%)

High-dimensional survival is an exciting area with many open problems!

Some Helpful Review Papers

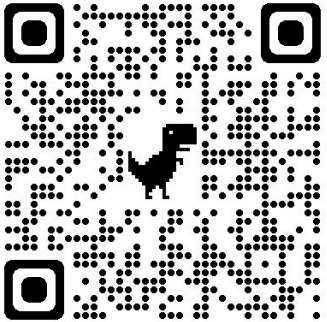
High-Dimensional Survival Analysis: Methods and Applications

https://www.annualreviews.org/doi/abs/10.1146/annurev-statistics-032921-022127



Shared Frailty Methods for Complex Survival Data: A Review of Recent Advances

https://www.annualreviews.org/doi/abs/10.1146/annurev-statistics-032921-021310



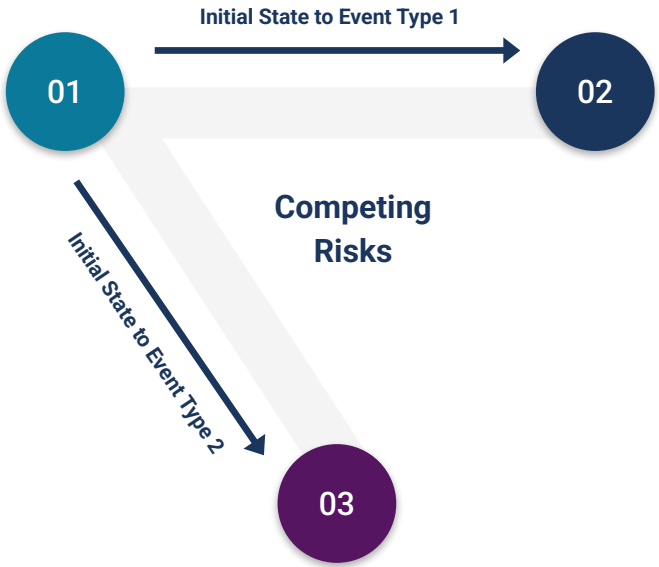
Fifty Years of the Cox Model

https://www.annualreviews.org/doi/abs/10.1146/annurev-statistics-033021-014043

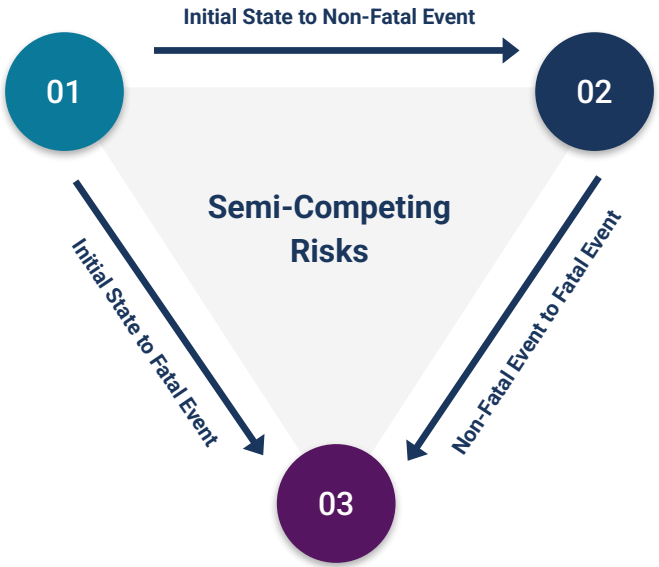
Fred Hutch
Cancer Center

Thank You!

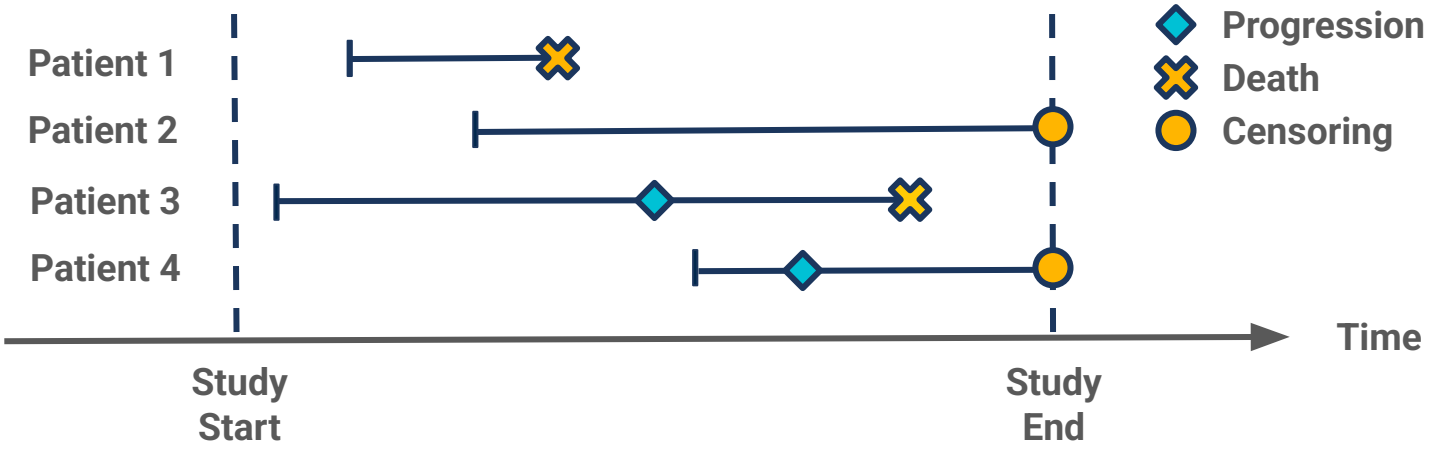
Univariate Risk



Competing Risks



Semi-Competing Risks


Patient 1
Patient 2
Patient 3
Patient 4
Progression
Death
Censoring
Time
Study Start
Study End