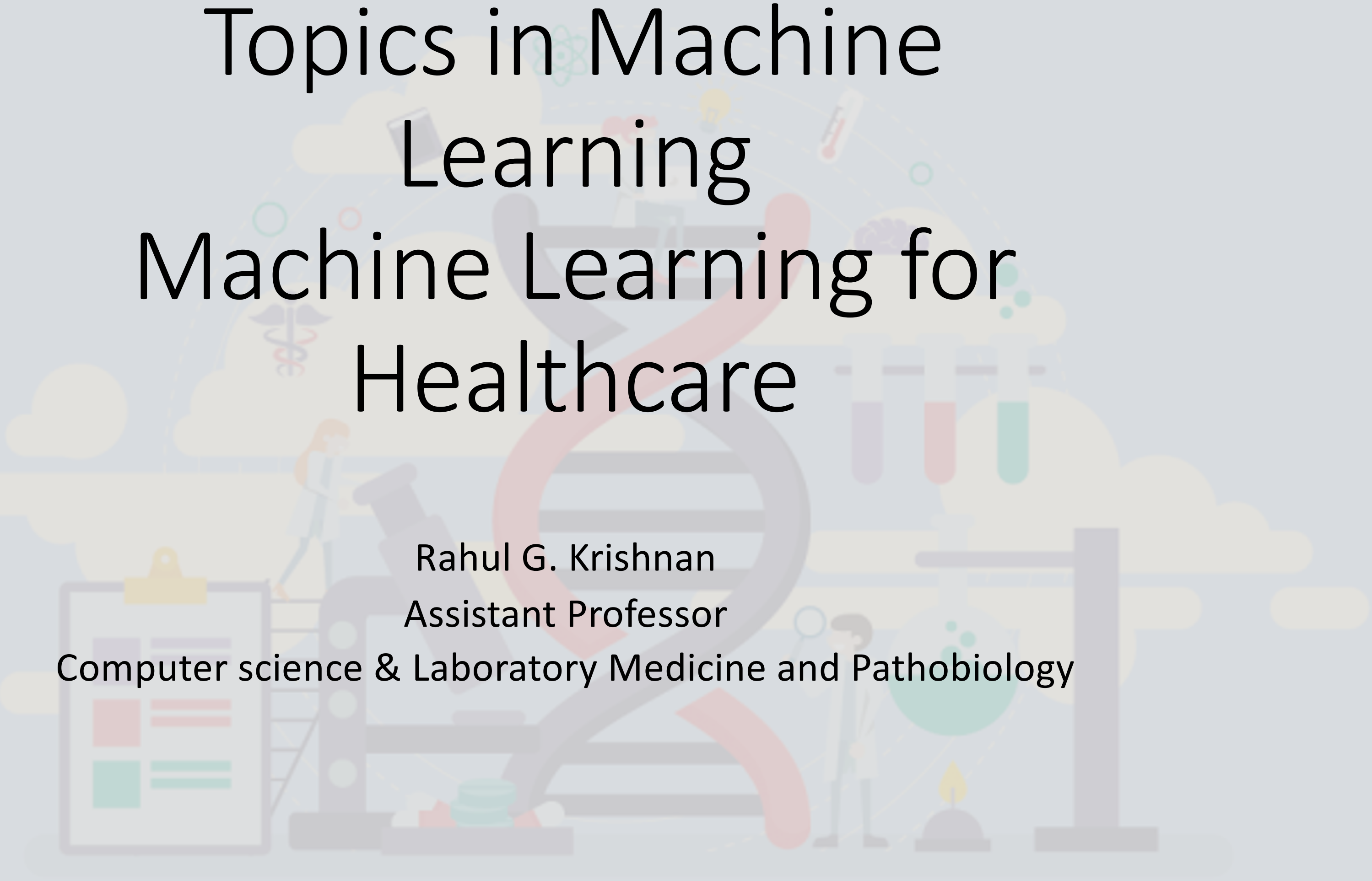
UNIVERSITY OF TORONTO

VECTOR INSTITUTE



Topics in Machine Learning
Machine Learning for Healthcare

Rahul G. Krishnan
Assistant Professor
Computer science & Laboratory Medicine and Pathobiology

Last week

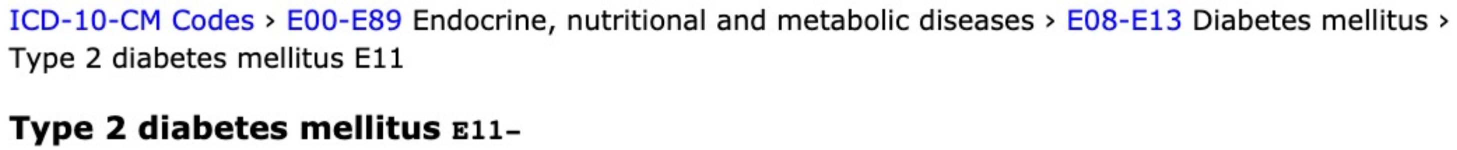
- Supervised machine learning
- Risk stratification
  - Stratification as a prediction problem
  - **Case study:** Predicting the onset of diabetes

Outline

- Clarification to questions
- Risk stratification
    - Deriving labels
    - Evaluating models
- Survival analysis:
    - From binary to continuous valued outcomes
    - Parametric
    - Non-parametric
    - Semi-parametric

Questions from last week

- Is there structure among diagnosis codes:
  - Yes! They are organized in a hierarchy. View ICD10 codes here

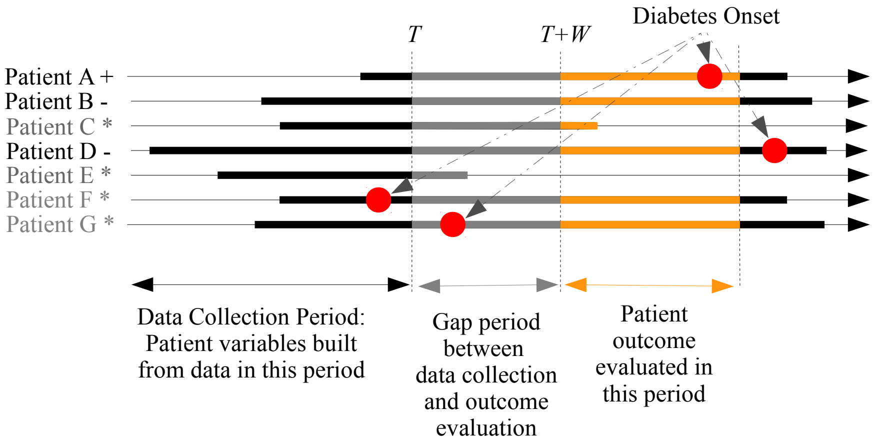


ICD-10-CM Codes › E00-E89 Endocrine, nutritional and metabolic diseases › E08-E13 Diabetes mellitus › Type 2 diabetes mellitus E11

Type 2 diabetes mellitus `E11-`

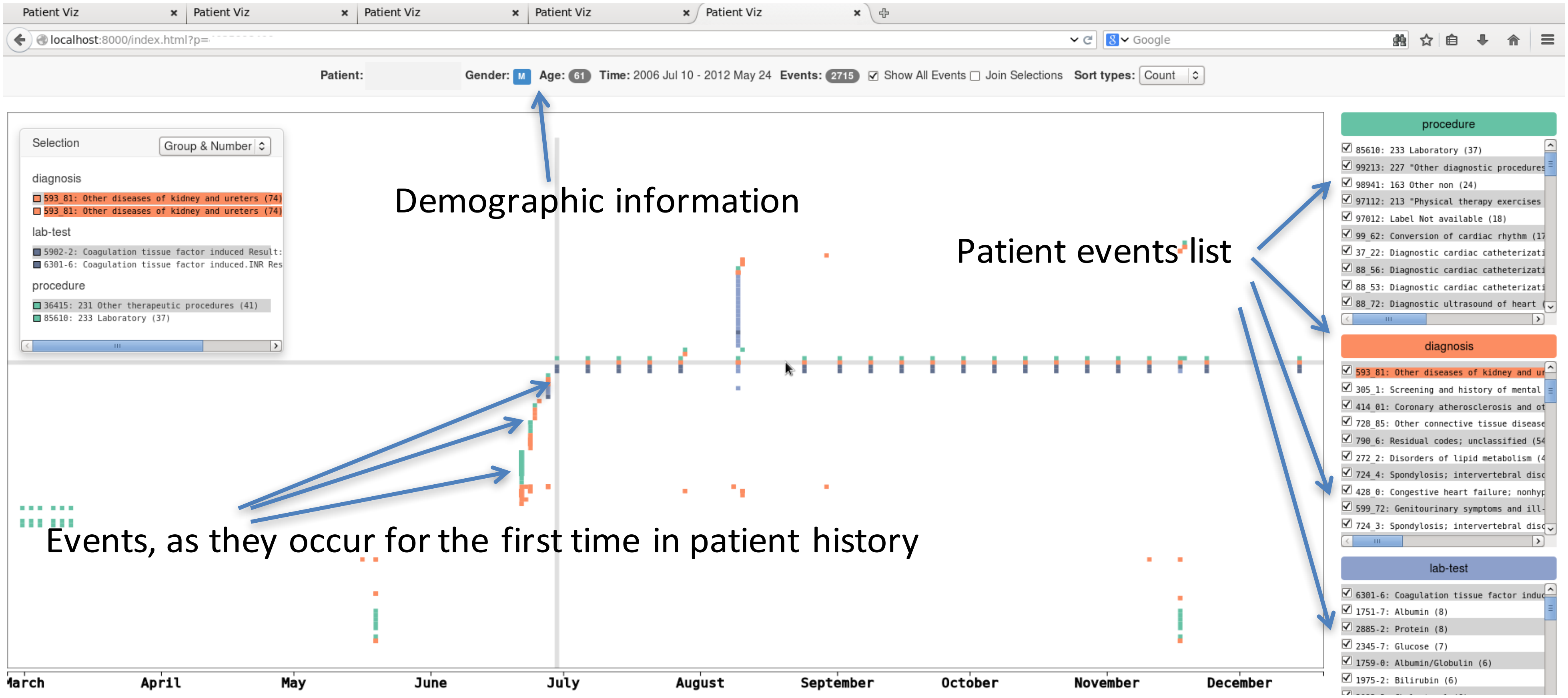
- Using predictive models in the future (on different data):
  - Non-stationarity of data is a challenging problem
    - Means that data distribution changes in unpredictable ways over time
  - Covariate shift can tank good machine learning models deployed in clinics
  - Still a lot of research on good techniques for detection of covariate shift

Deriving labels for risk stratification


Diabetes Onset
T
T+W
Patient A +
Patient B -
Patient C *
Patient D -
Patient E *
Patient F *
Patient G *
Data Collection Period: Patient variables built from data in this period
Gap period between data collection and outcome evaluation
Patient outcome evaluated in this period

- Typically done via chart review
- Work with doctor to assess criteria that constitute **Diabetic Onset**
  - e.g. does the patient have ICD10 code for diabetes

Evaluation of risk stratification models



Demographic information

Patient events list

Events, as they occur for the first time in patient history

Patient: **Gender:** M **Age:** 61 **Time:** 2006 Jul 10 - 2012 May 24 **Events:** 2715 ☑ Show All Events ☐ Join Selections **Sort types:** Count

Selection: Group & Number

diagnosis
- 593_81: Other diseases of kidney and ureters (74)
- 593_81: Other diseases of kidney and ureters (74)

lab-test
- 5902-2: Coagulation tissue factor induced Result:
- 6301-6: Coagulation tissue factor induced.INR Res

procedure
- 36415: 231 Other therapeutic procedures (41)
- 85610: 233 Laboratory (37)

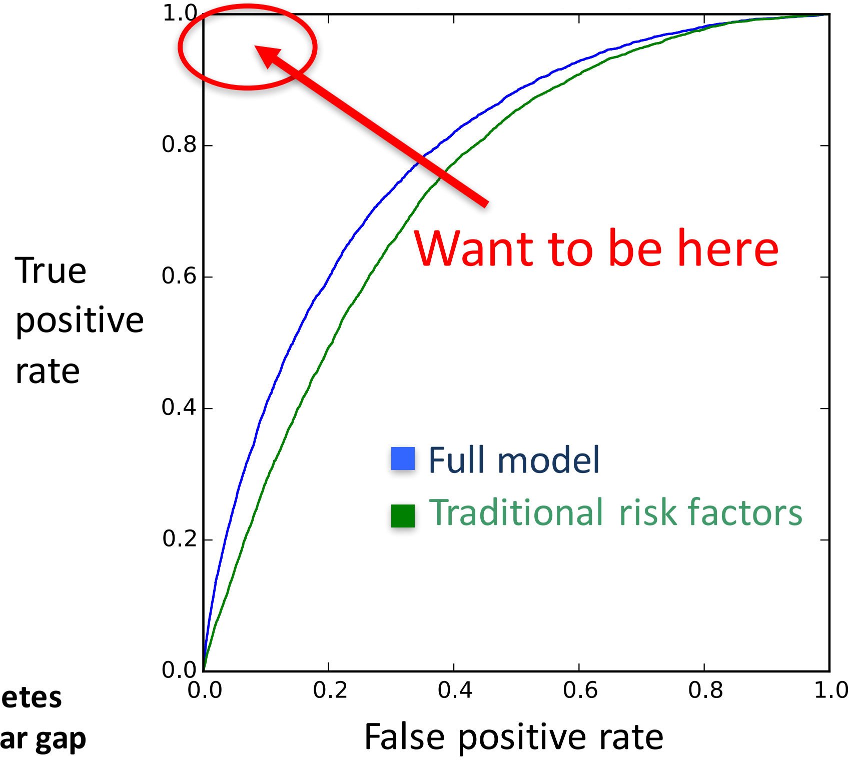
procedure
- ☑ 85610: 233 Laboratory (37)
- ☑ 99213: 227 "Other diagnostic procedures
- ☑ 98941: 163 Other non (24)
- ☑ 97112: 213 "Physical therapy exercises
- ☑ 97012: Label Not available (18)
- ☑ 99_62: Conversion of cardiac rhythm (17
- ☑ 37_22: Diagnostic cardiac catheterizati
- ☑ 88_56: Diagnostic cardiac catheterizati
- ☑ 88_53: Diagnostic cardiac catheterizati
- ☑ 88_72: Diagnostic ultrasound of heart (

diagnosis
- ☑ 593_81: Other diseases of kidney and ur
- ☑ 305_1: Screening and history of mental
- ☑ 414_01: Coronary atherosclerosis and ot
- ☑ 728_85: Other connective tissue disease
- ☑ 790_6: Residual codes; unclassified (54
- ☑ 272_2: Disorders of lipid metabolism (4
- ☑ 724_4: Spondylosis; intervertebral disc
- ☑ 428_0: Congestive heart failure; nonhyp
- ☑ 599_72: Genitourinary symptoms and ill-
- ☑ 724_3: Spondylosis; intervertebral disc

lab-test
- ☑ 6301-6: Coagulation tissue factor induc
- ☑ 1751-7: Albumin (8)
- ☑ 2885-2: Protein (8)
- ☑ 2345-7: Glucose (7)
- ☑ 1759-0: Albumin/Globulin (6)
- ☑ 1975-2: Bilirubin (6)

https://github.com/nyuvis/patient-viz

Evaluation of risk stratification models



Want to be here

Diabetes
1-year gap

AUC = Area under the ROC curve

- Invariant to class imbalance
- Interpretable as the probability that an algorithm ranks a positive patient over a negative patient

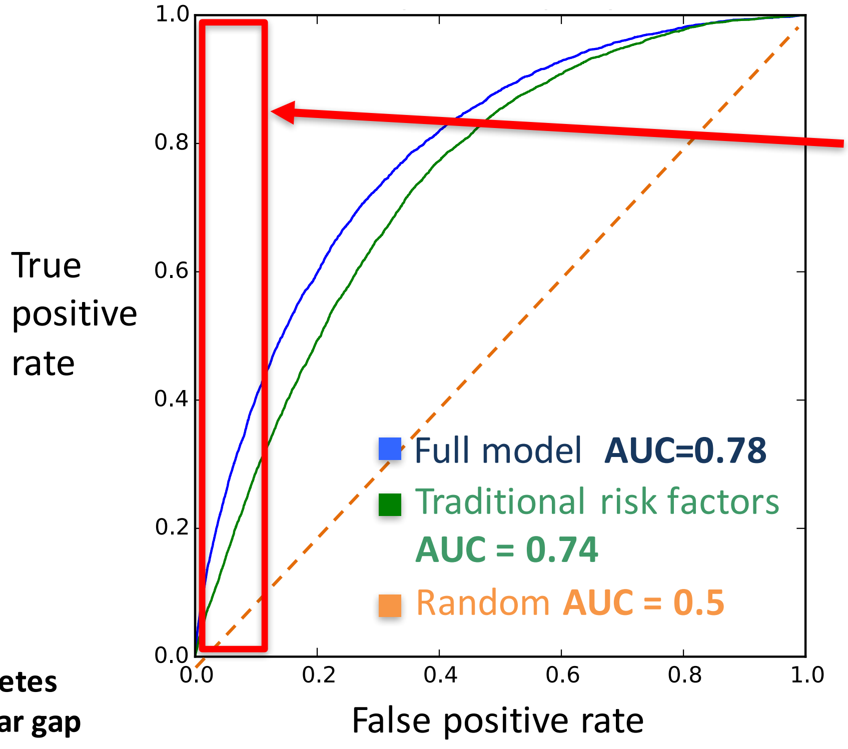
$$TPR = \frac{TP}{Actual\ Positive} = \frac{TP}{TP + FN}$$

$$FNR = \frac{FN}{Actual\ Positive} = \frac{FN}{TP + FN}$$

$$TNR = \frac{TN}{Actual\ Negative} = \frac{TN}{TN + FP}$$

$$FPR = \frac{FP}{Actual\ Negative} = \frac{FP}{TN + FP}$$

Where you want to be on ROC curve



Risk stratification usually focuses on just this region

(because of the cost of interventions)

Diabetes
1-year gap

Many other important statistical considerations when building models

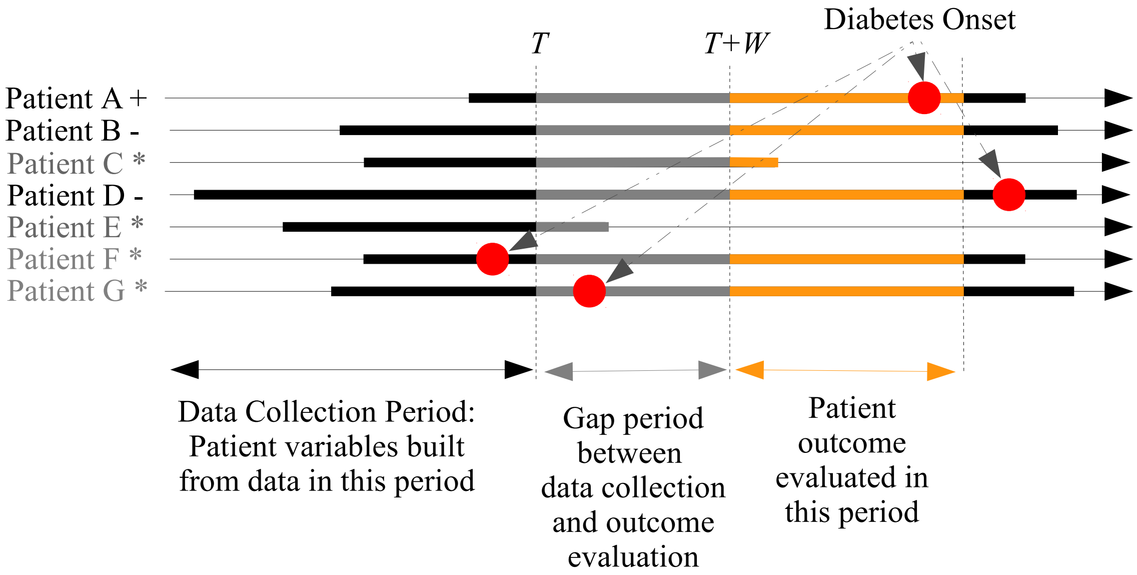
- Calibration
- Sensitivity analysis
- Error bars and confidence intervals on prediction estimates
- Heterogeneity of results:
  - Does the model only work well for a subpopulation?
- Model introspection:
  - For a linear model, are the features used by the models the ones you might expect?
  - Do root nodes in a decision tree make sense?
  - More challenging to do for deep neural networks

The importance of interpretability

- Intelligible Models for HealthCare: Predicting Pneumonia Risk and Hospital 30-day Readmission, Caruana et. Al, KDD 2015
- Used generalized additive models to make predictions of pneumonia and readmission
- Learn HasAsthma(x) = LowerRiskOfDying(x)
- Why?
  - Asthmatics w/ pneumonia are prioritized
  - Get aggressive treatment, faster, in ICU
  - Treatment lowers risk of death compared to general population
- Scenario where the prescription of an intervention taints the outcomes
- The consequence:
  - Automated methods might flag asthmatics as not being problematic!

Questions?

Continuous valued outcomes for risk stratification


Diabetes Onset
T
T+W
Patient A +
Patient B -
Patient C *
Patient D -
Patient E *
Patient F *
Patient G *
Data Collection Period: Patient variables built from data in this period
Gap period between data collection and outcome evaluation
Patient outcome evaluated in this period

More natural to define patient outcomes as the time taken to reach a clinical endpoint

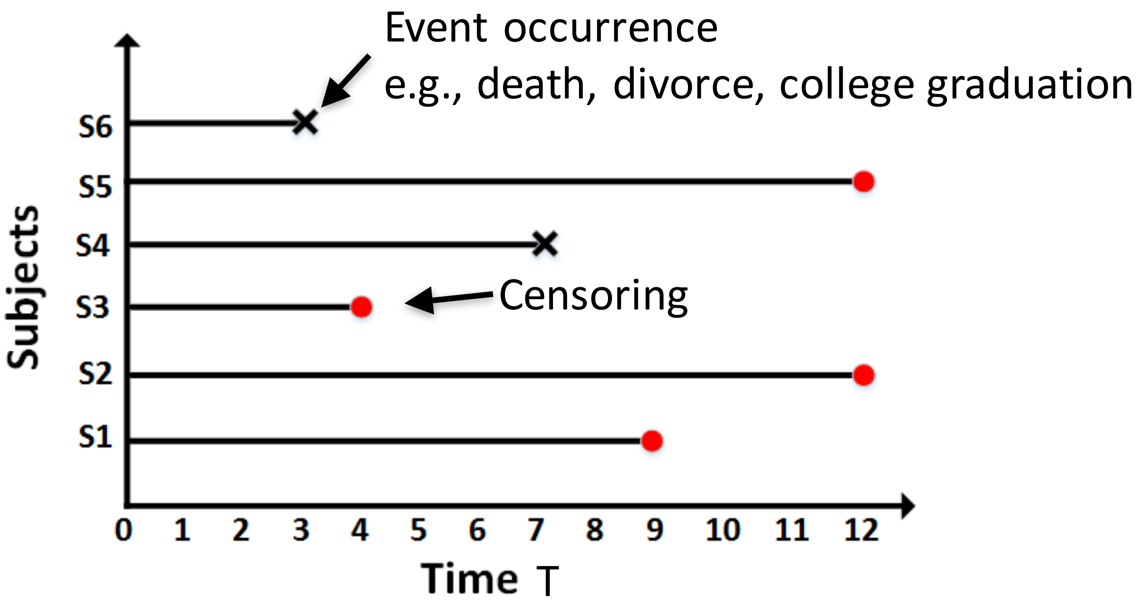
Time to re-admission

Time to death

Time to progression of disease

Labels are partially missing!

- We may not observe the outcome in the dataset – life, administrative challenges etc.


Event occurrence
e.g., death, divorce, college graduation
S6
S5
S4
S3
S2
S1
Censoring
Subjects
0
1
2
3
4
5
6
7
8
9
10
11
12
Time T

Event observed

Event not observed

This kind of missingness is known as right censorship!

Censorship

- Censorship is an important to know about when handling longitudinal data
- Three types of censorship:
  - Left censored data:
    - We don't observe the *start* of an event but we do observe longitudinal data after it
    - **Example**: ICU patient's vitals are continuously recorded from when the enter. If one of the sensor fails and is later fixed, their data is left censored
  - Right censored data [focus for today]
    - We don't observe the *incidence* of an event but we know it occurs after the last observed time
    - **Example:** We want to predict time-to-death in the ICU as our outcome but not all people die, some survive and leave the ICU at a certain time
  - Interval censored data:
    - Both left and right censorship
    - **Example:** Neonatal unit is tracking data on children, observe data sometime after they are born and until they leave the unit (for those who survive)

What can we do when we do not observe when the event occurs?

- What do we know:
  - x: features
  - y: last observed time
  - b: whether or not the event occurs
- Option: why not throw away all datapoints for which we don’t know when the event occurs:
  - Wasteful, might end up with very little data
- **Key idea** behind survival analysis:
  - Learn to predict time-to-event using all the available data that we have

Survival analysis

- To develop the ideas around survival analysis, we'll need some tools from probability theory,
- Our goal is to predict a continuous outcome:
  - We'll use random variable T to denote event time
  - We'll assume that an event can only occur in the future i.e. $T > 0$
- Next slide will introduce probabilistic concepts

Preliminaries – (1) - Notation

- (x, T, b) = (features, time, censoring)
    - b = 0 if censored and b = 1 if event is observed
- f(t) = p(t) = probability of death at time t; F(t) = P(T<=t) = CDF of t
- Survival function: $S(t) = P(T > t) = \int_t^\infty f(x)dx$
- Hazard function: $h(t) = \lim_{\epsilon \to 0} p(T \in (t, t + \epsilon] | T \geq t)$
- Cumulative hazard function: $H(t) = \int_0^t h(u)du = \int_0^t \frac{f(u)}{S(u)}du = \int_0^t \frac{-dS(u)}{S(u)}du = -\log\{S(t)\}$
- Hazard function & survival function: $h(t) = \frac{f(t)}{S(t)}$

$$f(t) = h(t)S(t) = h(t)\exp\{-H(t)\}$$

$$S(t) = \exp\{-H(t)\}.$$

Slide credit: Lu Tian and Richard Olshen’s course on survival analysis

Preliminaries – (2) - Visualization

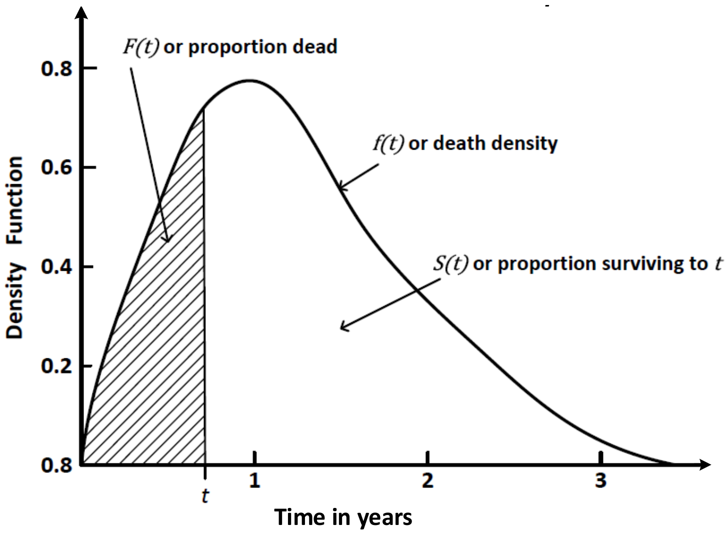


Fig. 2: Relationship among different entities $f(t)$, $F(t)$ and $S(t)$.

[Wang, Li, Reddy. Machine Learning for Survival Analysis: A Survey. 2017]

[Ha, Jeong, Lee. Statistical Modeling of Survival Data with Random Effects. Springer 2017]

What is a hazard function?

- Different from the probability density of event time (t)
- h(t) dt is approximately the conditional probability of the event occurring in an infinitesimal interval around t conditional on it not having occurred before t

Lecture restart [Sept 17]
Announcements

- Continue our discussion on survival analysis
- Watch MIMIC data and analysis tutorials [posted to course website]
  - Discussion with Alistair Johnson, Scientist, SickKids Hospital
- **Upcoming deadline [October 1, 2021]: Project proposal due**
  - **10% of your grade**
  - Piazza now has a pinned post to help you start looking for teammates to work on problems – make use of it
  - Start to form groups of 3-4 people to work together on a problem and come up with a candidate proposal,
  - Does not have to be exactly what the final project will look like, just some ideas on what are interesting problems and how you'll tackle them,
  - Use TA office hours! Dataset questions, project idea brainstorming

Project resources

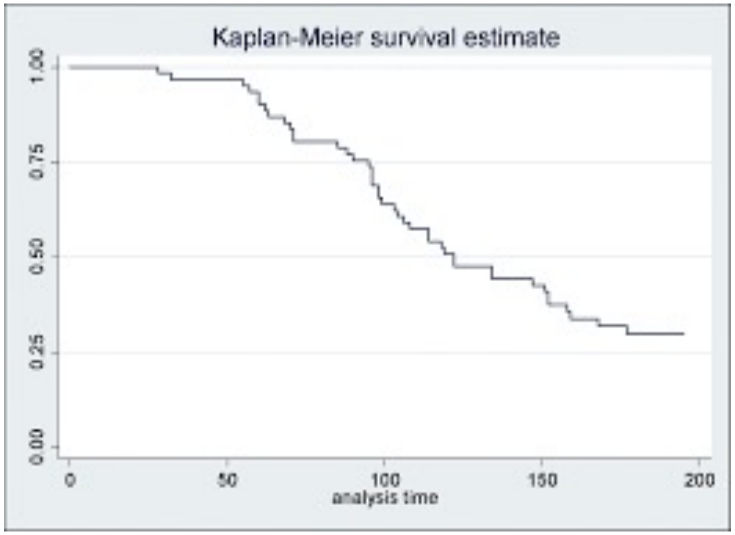
- Cancer:
  - COMMpass study: https://themmrf.org/finding-a-cure/our-work/the-mmrf-commpass-study/
  - Starter code for processing data from the study: https://github.com/clinicalml/ml_mmrf
- Parkinson's disease
  - https://www.ppmi-info.org/
- **APPLY EARLY FOR ACCESS!**

Non-parametric survival analysis

- Let start by ignoring our features and asking about computing S(t)
  - S(t) is an integral $S(t) = P(T > t) = \int_t^{\infty} f(x)dx$
  - **Idea:** If we had access to f(x) we could discretize time and evaluate f(x) in each bin and sum them up.
  - **Issue:** We don’t have access to f(x) but we **do** have samples!
- **Kaplan Meier curves:**
  - **Non-parametric** estimator of the survival function S(t)
    - We do not assume anything about the underlying distribution of S(t)
  - We’ll use our entire dataset to approximate the shape of S(t)

Kaplan Meier estimator

- Derivation out of scope for this class
  - Survival analysis is a rich area of research and is often a course in and of itself
  - E.g. Lu Tian and Richard Olshen at Stanford



Observed event times

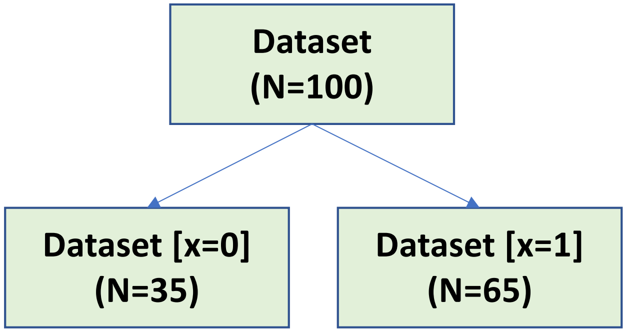
$$y_{(1)} < y_{(2)} < \cdots < y_{(D)}$$

$d_{(k)}$ = # events at this time

$n_{(k)}$ = # of individuals alive and uncensored

$$\widehat{S}_{K-M}(t) = \prod_{k:y_{(k)} \le t} \left\{ 1 - \frac{d_{(k)}}{n_{(k)}} \right\}$$

What do we do if we have features (x)?



Dataset (N=100)

Dataset [x=0] (N=35)

Dataset [x=1] (N=65)

Evaluate KM estimator on each strata

Right [survival probability of patients who have multiple myeloma stratified by genetic marker]

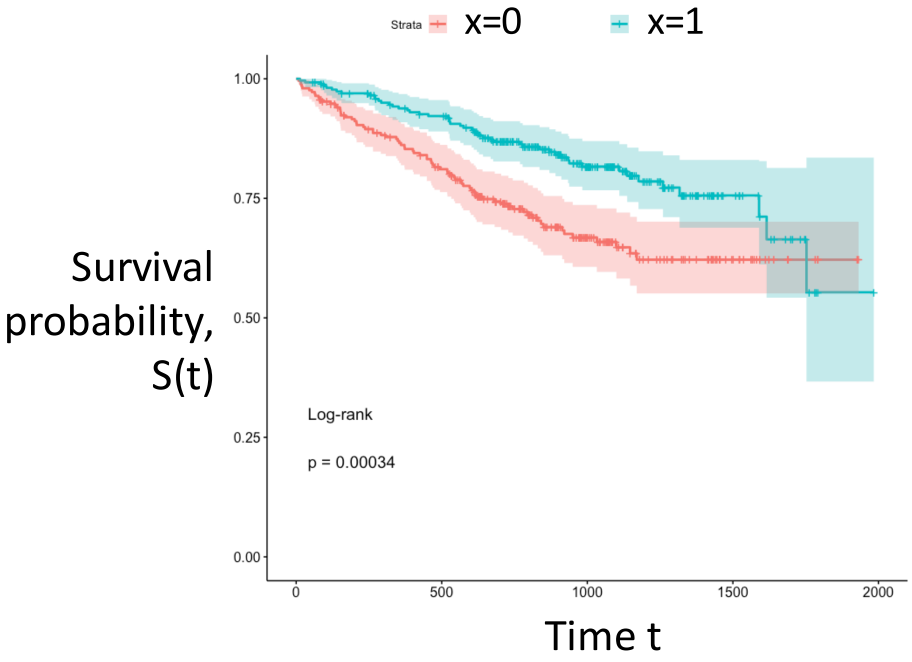
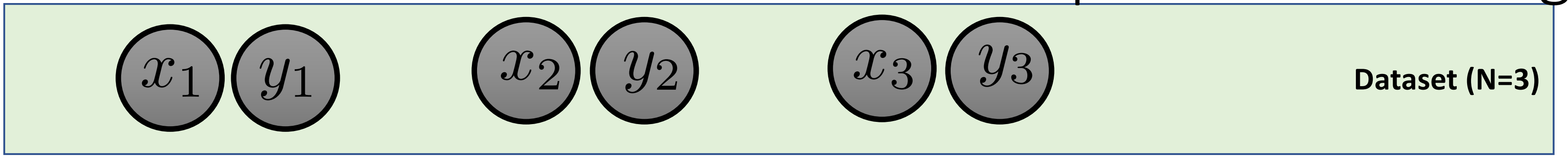


Figure credit: Rebecca Boiarsky

What if x is high-dimensional?

- Option 1: Cluster x and stratify based on clusters of x
- Option 2: Let the survival function **depend on x**
- This idea is used in linear regression!
- In linear regression: $y \sim \mathcal{N}(w^T x + b; 1)$
  - Outcome is a Gaussian function centered around [w^Tx + b]
  - Known as a parametric model for y:
    - There are some parameters that govern the behavior of y as a function of x

Maximum likelihood estimation for supervised learning



- Given a dataset, the model parameters are learned via **maximum likelihood estimation**

$$\mathcal{L}(y, x) = \log p(y|x; \theta)$$

Score function (high is good, low is bad)

$$\theta = \arg\max_{\theta} \sum_{i=1}^{N} \mathcal{L}(y_i, x_i)$$

Solve this optimization problem to **learn** the model. Often formulated as a minimization of the negative of the log-likelihood function

Maximum likelihood estimation for survival analysis

x_1 t_1 b_1 x_2 t_2 b_2 x_3 t_3 b_3

Dataset (N=3)

- Given a dataset, the model parameters are learned via **maximum likelihood estimation**

$$p(T = t|x;\theta) = f(t)$$

Uncensored likelihood

$$p(T > t|x;\theta) = S(t)$$

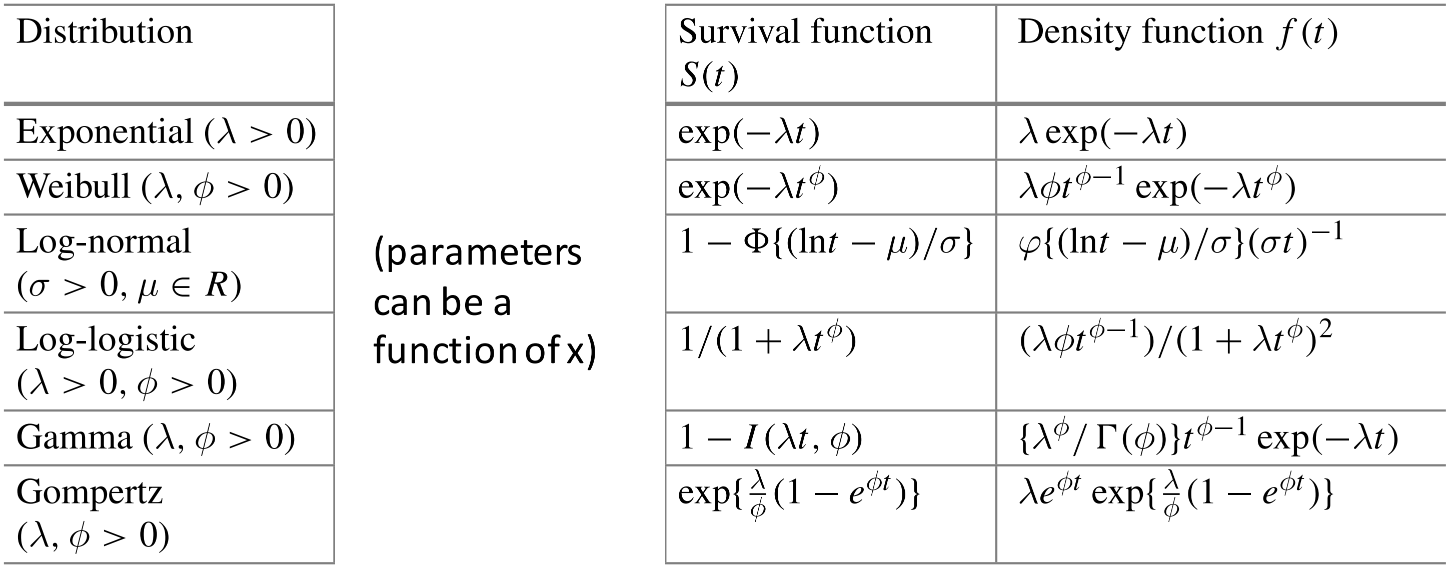
Censored likelihood

$$\sum_{i=1}^{N} b_i \log p(T = t_i|x_i;\theta) + (1 - b_i) \log p(T > t_i|x_i;\theta)$$

Maximize the following objective function to learn model parameters

What distribution should I use for T?

Table 2.1 Useful parametric distributions for survival analysis



Distribution	Survival function $S(t)$	Density function $f(t)$
Exponential ($\lambda > 0$)	$\exp(-\lambda t)$	$\lambda \exp(-\lambda t)$
Weibull ($\lambda, \phi > 0$)	$\exp(-\lambda t^{\phi})$	$\lambda \phi t^{\phi-1} \exp(-\lambda t^{\phi})$
Log-normal ($\sigma > 0, \mu \in R$)	$1 - \Phi\{(\ln t - \mu)/\sigma\}$	$\varphi\{(\ln t - \mu)/\sigma\}(\sigma t)^{-1}$
Log-logistic ($\lambda > 0, \phi > 0$)	$1/(1 + \lambda t^{\phi})$	$(\lambda \phi t^{\phi-1})/(1 + \lambda t^{\phi})^2$
Gamma ($\lambda, \phi > 0$)	$1 - I(\lambda t, \phi)$	$\{\lambda^{\phi}/\Gamma(\phi)\} t^{\phi-1} \exp(-\lambda t)$
Gompertz ($\lambda, \phi > 0$)	$\exp\{\frac{\lambda}{\phi}(1 - e^{\phi t})\}$	$\lambda e^{\phi t} \exp\{\frac{\lambda}{\phi}(1 - e^{\phi t})\}$

(parameters can be a function of x)

[Ha, Jeong, Lee. Statistical Modeling of Survival Data with Random Effects. Springer 2017]

CoxPH: Interpretability in survival analysis

- Parametric models that depend on x change parameters of a distribution in linear/non-linear ways as a function of x
- **Goal:**
  - Link variation to covariates directly to the survival function
- The Cox Proportional Hazard’s model is one of the most popular tools in survival analysis

$$h(t|X=x;\theta) = \underbrace{h_0(t)}_{\text{Baseline hazard}} \exp(\beta^T x)$$

Baseline hazard reflects the hazard for subjects with all covariates equal to 0

Interpretation in the univariate case

$$\frac{h(t|X = x_1;\theta)}{h(t|X = x_2;\theta)} = \frac{\exp(\beta^T x_1)}{\exp(\beta^T x_2)} \quad \frac{h(t|X = x+1)}{h(t|X = x)} = \exp(\beta)$$

Hazard ratio is independent of time

Parameters have an intuitive meaning

CoxPH: Linear model for log of the hazard ratio

CoxPH for binary data

- X = [received drug (0 no, 1 yes) , gender (0 male, 1 female)]

$$h(t|x_1, x_2) = h_0(t)\exp(\beta_1 z_1 + \beta_2 z_2)$$

$h(t\|X) = h_0(t)$	No treatment	Male
$h(t\|X) = h_0(t)\exp(\beta_1)$	Yes treatment	Male
$h(t\|X) = h_0(t)\exp(\beta_2)$	No treatment	Female
$h(t\|X) = h_0(t)\exp(\beta_1 + \beta_2)$	Yes treatment	Female

Key advantage of the CoxPH model

- We can estimate the model parameters β
- Notably we can do so **without** estimating the baseline hazard
- This is a **semi-parametric model**
  - We make no assumptions about the baseline hazard rate
  - However, we learn parameters that dictate how it is modified based on patient covariates
- How do we learn this model?
  - Won't derive from scratch in this class but we'll discuss the algorithm
    - a useful exercise if this is your area of research
    - come to my office hours if you're interested in pursuing a project around this!
  - Useful reference: Course notes by Ronghui (Lily) Xu

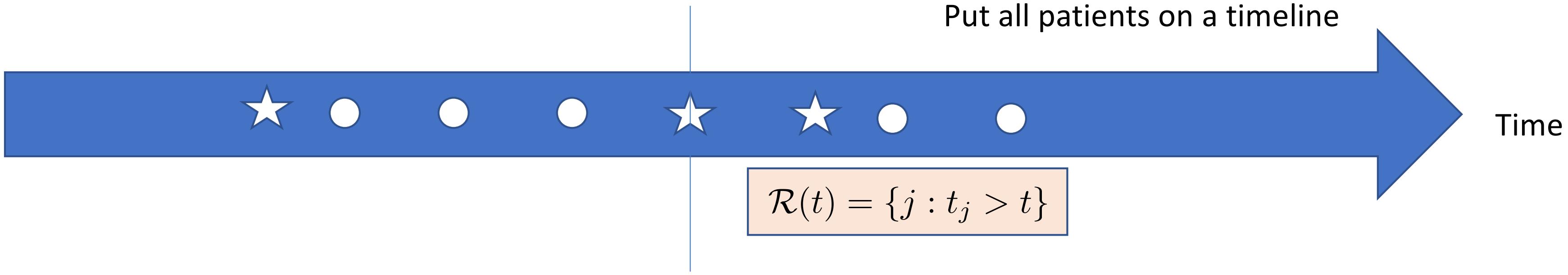
Cox Partial Likelihood

1. Loss function used for learning the Cox Proportional Hazards model
2. Scan from left to right in time, at each discrete point, calculate the risk set and the loss
3. Sum up the losses and use gradient based methods for parameter estimation

$$\mathcal{R}(t) = \{j : t_j > t\}$$

$$\mathcal{L}(\beta) = \sum_{i=1}^{K} b_i \log \frac{\exp(\beta^T X_i)}{\sum_{l \in \mathcal{R}(t_i)} \exp(\beta^T X_l)}$$

Visualizing the computation of the partial likelihood



$\mathcal{R}(t) = \{j : t_j > t\}$

$$\mathcal{L}(\beta) = \sum_{i=1}^{K} b_i \log \frac{\exp(\beta^T X_i)}{\sum_{l \in \mathcal{R}(t_i)} \exp(\beta^T X_l)}$$

○ Censored event

☆ Event

Intuition: How likely are the features of this patient to explain their elevated risk of having the event occur now compared to all the individuals whose event occurs later!

Advances in machine learning for survival analysis

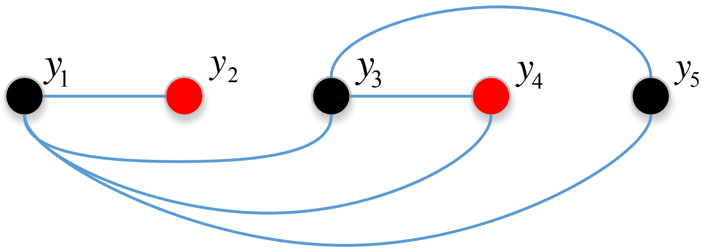
- DeepSurv, Katzman et. al, 2017
  - One of the readings for this week uses a deep neural network to parameterize the modification to the hazard function
  - Parameter estimation by taking derivatives of the
- Advanced reading Deep survival analysis, Ranganath et. al, 2016
  - What if x is very high dimensional?
  - Rather than condition on x directly, learn a latent representation of x while jointly modeling survival time

Evaluation in survival analysis

- Concordance index (aka C-statistic) – predicts how well the model ranks patients based on survival (i.e. predicts relative survival time)
- Equivalent to AUC (when there is no censoring)

$$\hat{c} = \frac{1}{num} \sum_{i:\bar{b}_i = 0} \sum_{j:y_i < y_j} I[S(\hat{y}_j|X_j) > S(\hat{y}_i|X_i)]$$

Black = uncensored
Red = censored



Other ways to evaluate models

- Mean squared error [for just those who are uncensored]
- Held out likelihood (censored + uncensored)

Questions?

Why not use classification?

This can be a reasonable option when data is scarce,

Thresholds for classification may not be known at training time,

Why not use regression?

When outcomes are missing [event time not observed] you may have to throw data out

- Leads to limited training data
- Might introduce bias into the dataset