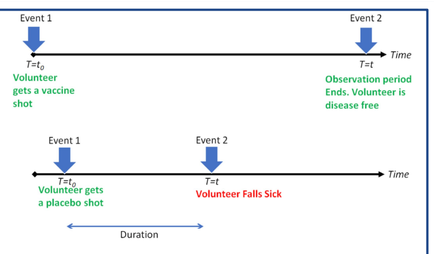
CS 291k Differential Privacy in Survival Analysis

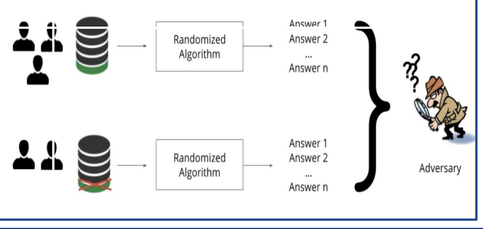
Esha Singh
University of California, Santa Barbara

1 Problem Definition

- Define Survival Analysis (SA)
  - Estimation of survival function & understanding effects of explanatory variables modelled by different survival models.

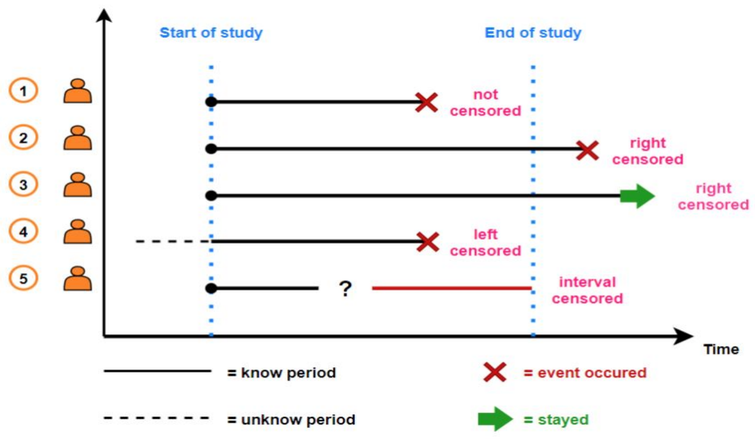


- Privacy issues of survival models
  - Aggregated results shared after SA leads to privacy risk
- Differential Privacy (DP) to the rescue!
  - None of the previous work can guarantee privacy



2 Survival Analysis

- Set of statistical methods for analyzing the time until an event occurs in a population.

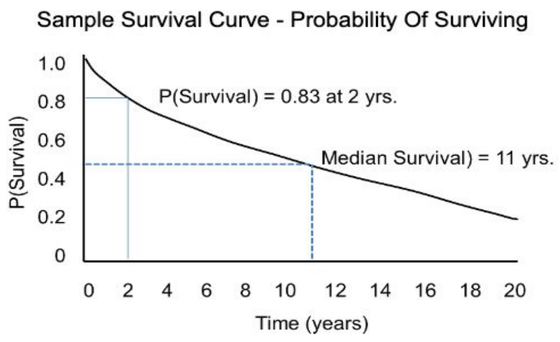


Ref: Survival Analysis, Jim Gruman

- Major components of SA

1	Time T	A non-negative random variable representing time till an event of interest occurs
2	Survival Function	S(t): with *t* as input and outputs probability of survival time greater than *t* S(t) = 1 - F(t) = P(T>t) for t>0
3	Hazard Function	h(t): probability of occurrence of the event at T = t, assuming that the event has not occurred up through t. h(t) = f(t)/S(t-)

- Survival curve



- Estimating the Survival function
  - Non-parametric methods: Kaplan-Meier Estimator model

$$\hat{S}_{KM}(t) = \prod_{l:t_l \geq t}(1 - \frac{D_t}{N_l})$$

  - Semi-parametric methods: Cox proportional hazards model

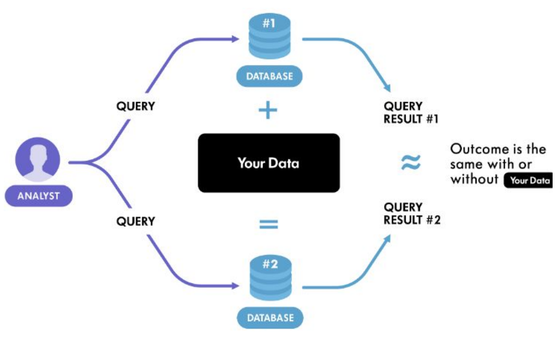
$$\lambda_i(t|x_i) = \lambda_0(t)exp\{x_i'\beta\}$$

- Evaluation metrics

AIC (Akaike Information Criterion) | Concordance index | Mean Relative Error (MRE)

3 Differential Privacy

- Theoretical framework to ensure robust privacy of individual-level data when performing statistical analysis of sensitive datasets.



- A randomized algorithm is (ε, δ) - differentially private if:

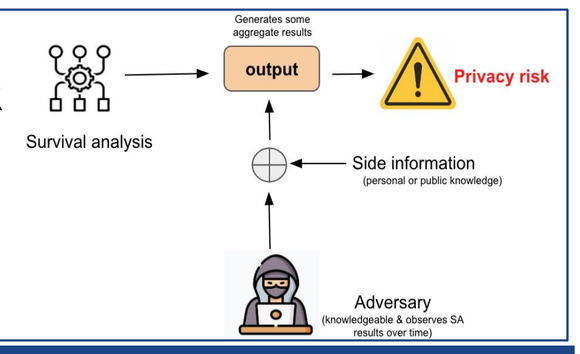
$$Pr[\mathcal{M}(x) \in S] \leq exp(\epsilon)Pr[\mathcal{M}(y) \in S] + \delta$$

$\mathcal{M}$: randomized algorithm
$N^{|\mathcal{X}|}$: domain of $\mathcal{M}$
ϵ, δ : privacy budget, delta
$S \subseteq$ Range($\mathcal{M}$)
$x,y \in N^{|N|}$, $x - y \leq 1$

- if δ=0, then $\mathcal{M}$ is ϵ-DP

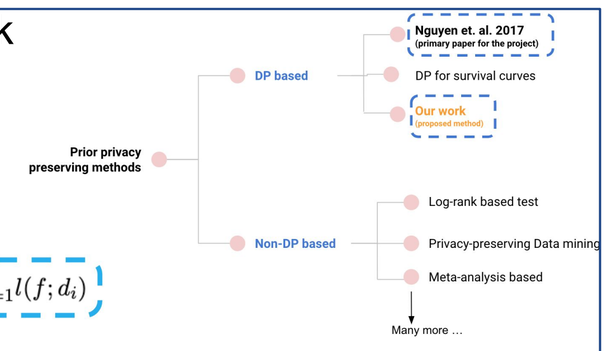
3 Privacy Issues in Survial analysis

- Survival curves along with side-information increase PR
- **No prior work** provides provable gaurnette against PR for SA



4 Survival analysis with DP

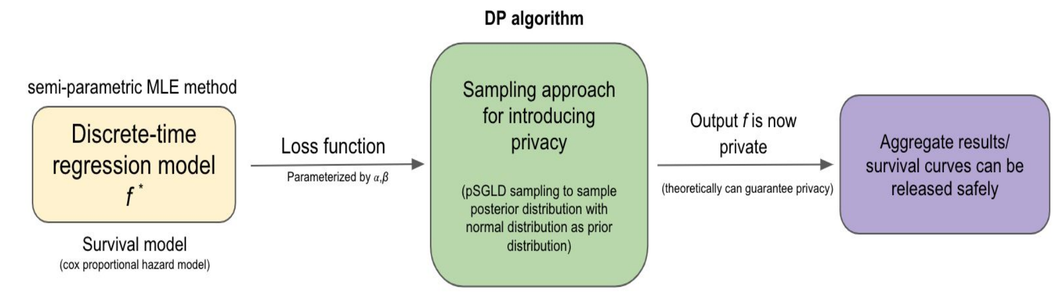
- Most recent & vital prior work in this field Nguyen et.al.[1]



- Methodology:
  - Estimate MLE

$$f^* = arg\ min\ \Sigma_{i=1}^{n} l(f; d_i)$$

  - Privatize the loss function



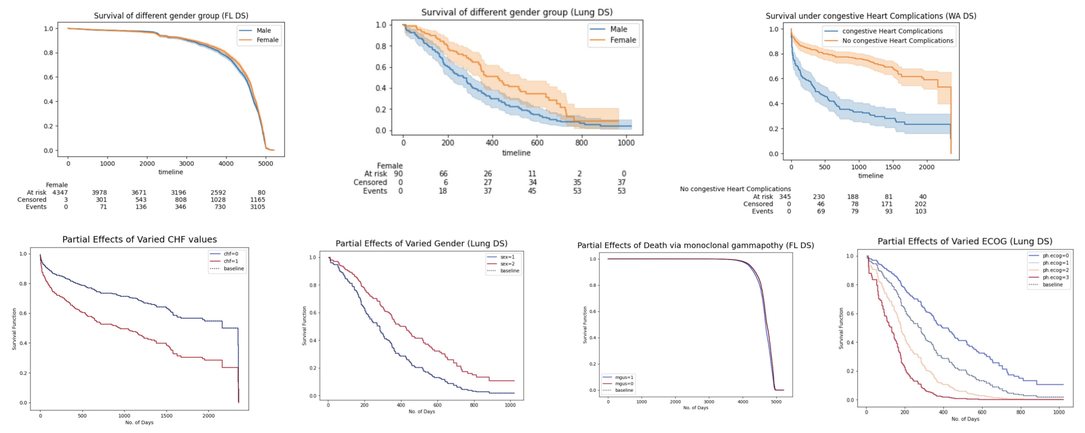
- Prior works with KM estimator for MLE estimation but many drawbacks with using such methods practically.
- MCMC sampling methods only approximately ε-DP. Instead use more robust DP mechanisms eg: DP-SGD.
- **Contributions**
  - First ever implementation of Nguyen et.al. [1]
  - Stronger privacy mechanism proposed.

5 Experiments

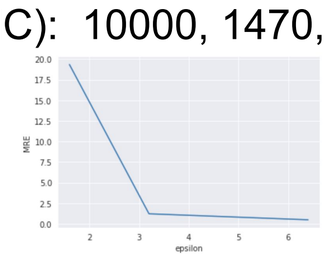
- Data

Dataset	Size	# uncensored	# explanatory variables
Lung cancer	228	165	7
FL	7874	2169	8
WA	500	215	14

- KM & Cox hazard model survival plots for 3 datasets (DS):



- Evaluation:
  - Concordance statistic: 85%, 65%, 78%
  - AIC (or partial AIC): 10000, 1470, 2200
- MRE $= \frac{1}{t}\sum_{i=1}^{t}\frac{\|f_i - f^*\|}{\|f^*\|}$



Concordance → How discriminant the model is
AIC → Estimate of the prediction error of the model

  - Figure shows MRE values wrt to ϵ for only WA DS.
  - Plotting MRE for all 3 DS & w.r.t increasing epochs show similar trend. Results stabilize with more epochs.

5 Conclusion & Future work

- Cox PH allow to account for multiple factors for observations
- MCMC sampling method for privacy-preserving SMs
- Need comparative analysis with other classical DP mechanisms

Code repository coming soon: https://github.com/esha-singh