Semi-supervised Algorithms for Risk Assessment with Noisy EHR Data

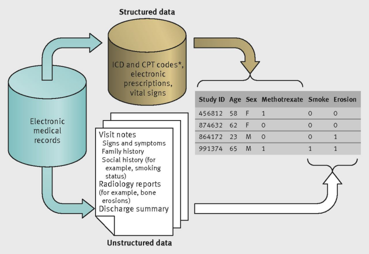
Chuan Hong

(Joint work with Tianxi Cai from Harvard University)

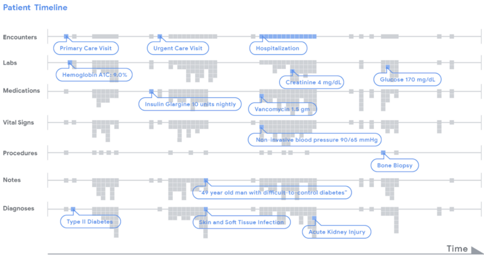
Department of Biostatistics and Bioinformatics
Duke University

September 8, 2022

EHR Structure



(Figure from Liao et al. 2015)



(Figure from ai.googleblog.com)

- **Structured data**: ICD billing codes; lab results etc
- **Unstructured text data**: extracted via natural language processing (NLP)
- **Detailed longitudinal patient level data**

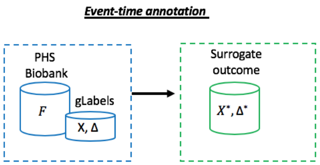
Challenges in EHR-linked Survival Analysis

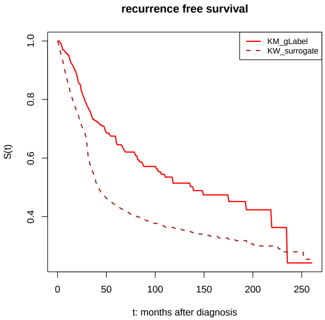
Disease status and event time information is not readily available.

- Annotating **event time** requires labor extensive chart review
- Time to first ICD codes **inaccurate**
- **Surrogate event time**: derived from label+codes+NLP
- $\rightsquigarrow$ power loss, biased estimates

Motivating Example: PHS Lung Cancer Study

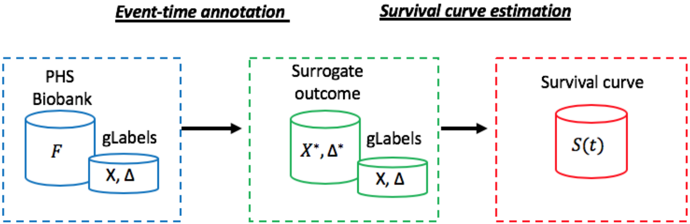
- PHS: Partners Healthcare contains both a wealth of clinical and also biological measurements.





- Aim to estimate recurrence free survival for lung cancer patients.
- Lung cancer data mart
- 70K patients from PHS biobank EMR
- 40K patients identified as lung cancer
- 5K early stage patients
- 300 gLabels for event time manually annotated by domain experts
- Surrogate outcomes derived by using Uno et al. 2018

Motivating Example: PHS Lung Cancer Study



Solution: Develop a calibrated survival curve that combines imperfect sources of information on event time in EHR, together with the exact event time.

Sources of information on event times in EHRs

- T_i: failure time, the time that the patient developed the event
- C_i: censoring time

Labeled data: $i = 1, \cdots, n$

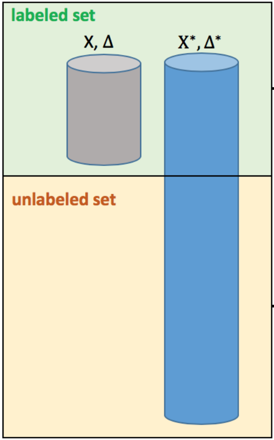
- $X_i = \min(T_i, C_i)$: observed event time:
- $\Delta_i = \mathbb{1}(T_i \leq C_i)$: the censoring indicator (whether the patient developed the event prior to the last visit)

Unlabeled data: $i = 1, \cdots, n + N$

- X^*: imperfect estimates of event times
  – time to the first ICD code related to the event
  – time to the first NLP mention of the event
  – algorithm annotated event time
- $\Delta_i^* = \mathbb{1}(T_i^* \leq C_i)$

Goal: estimate the survival function $S(t)$.

Conditional Nelson-Aalen Estimator



$$\hat{S}(t|X_i^* = x, \Delta_i^* = d) = \exp\{-\widehat{\Lambda}(t;x,d)\},$$

$$where, \widehat{\Lambda}(t;x,d) = \int_0^t \frac{\sum_{i;\Delta_i^*=d} K_h(X_i^* - x)dN_i(z)}{\sum_{i;\Delta_i^*=d} K_h(X_i^* - x)Y_i(z)}$$

$$with\ N_i(t) = I(X_i \leq t)\Delta_i, and\ \ Y_i(t) = I(X_i \geq t)$$

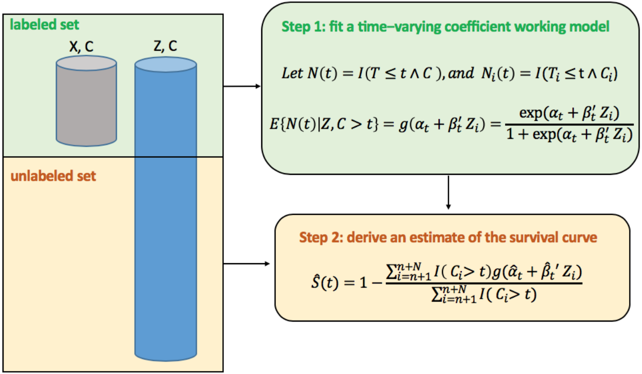
$$\hat{S}(t) = N^{-1} \sum_{i=n+1}^{n+N} \hat{S}(t|X_i^*, \Delta_i^*)$$

- Parast et al. 2014: baseline covariates
- In our case, T and C may no longer be independent conditional on X^* and Δ^* if there is no restriction on X^* and Δ^*.
- As a result, the Nelson-Aalen type estimator defined above may not be consistent.

Semiparametric Estimator

Let $\mathbf{Z}$ denote a variable related to both T and C.
As $T \perp\!\!\!\perp C$, we have

$$
\begin{aligned}
\pi_t &= E\{\mathbb{1}(T \le t)\} = E\{\mathbb{1}(T \le t) \mid C > t\} = E\{\mathbb{1}(T \le t)\Delta \mid C > t\} \\
&= E[E\{N(t) \mid \mathbf{Z}, C > t\} \mid C > t].
\end{aligned}
$$



Combined Estimator

To improve the efficiency, we combine proposed semiparametric estimator with the KM.

Let $\widehat{\boldsymbol{\mu}} = (\widehat{S}_{\mathrm{Semi}}, \widehat{S}_{\mathrm{KM}})^{\mathsf{T}}$, and $\boldsymbol{\Sigma}$ denote their covariance matrix, then the combined estimator is constructed as

$$(\mathbf{1}^{\mathsf{T}}\boldsymbol{\Sigma}^{-1}\mathbf{1})^{-1}\mathbf{1}^{\mathsf{T}}\boldsymbol{\Sigma}^{-1}\widehat{\boldsymbol{\mu}}$$

- unbiased
- smallest possible variance among all linear combinations

Asymptotic Properties of Proposed Estimator

Let $\mathbf{W}_i = (1, \mathbf{Z}_i^{\mathsf{T}})^{\mathsf{T}}$.

Step 1.

$$\sqrt{n}(\widehat{\boldsymbol{\theta}}_t - \boldsymbol{\theta}_t) \to \text{Normal}\{\mathbf{0}, \mathbf{A}^{-1}\mathbf{B}(\mathbf{A}^{-1})^{\mathsf{T}}\},$$

where

$$\begin{aligned} \mathbf{A} &= -E\{\mathbb{1}(C > t) g'(\boldsymbol{\theta}_t^{\mathsf{T}}\mathbf{W})\mathbf{W}\mathbf{W}^{\mathsf{T}}\}; \\ \mathbf{B} &= \text{cov}[\mathbb{1}(C > t)\{N(t) - g(\boldsymbol{\theta}_t^{\mathsf{T}}\mathbf{W})\}\mathbf{W}]; \\ g'(x) &= \exp(x)/\{1 + \exp(x)\}^2. \end{aligned}$$

Step 2.

$$\sqrt{n}(\widehat{\pi}_t - \pi_t) = G(t)^{-1} E\{\mathbb{1}(C > t) g'(\boldsymbol{\theta}_t^{\mathsf{T}}\mathbf{W})\mathbf{W}\}^{\mathsf{T}} \sqrt{n}(\widehat{\boldsymbol{\theta}}_t - \boldsymbol{\theta}_t) + o_p(1),$$

where $G(t) = P(C > t)$.

Asymptotic Properties of Kaplan Meier Estimator

We have

$$\sqrt{n}(\widehat{\pi}_t^{\mathrm{KM}} - \pi_t) \quad = \quad \frac{1}{\sqrt{n}} \sum_{i=1}^{n} (1 - \pi_t) \int_0^t \frac{dM_i(u)}{P(X \geq u)} + o_p(1).$$

In practice, we have to replace all the unknown quantities in the above influence function with their estimations. This leads to

$$(1 - \widehat{\pi}_t^{\mathrm{KM}}) \int_0^t \frac{d\widehat{M}_i(u)}{\overline{Y}(u)} \quad = \quad (1 - \widehat{\pi}_t^{\mathrm{KM}}) \left\{ \frac{\Delta_i \mathbb{1}(X_i \leq t)}{\overline{Y}(X_i)} - \sum_j \frac{\Delta_j \mathbb{1}(X_j \leq t \wedge X_i)}{n\overline{Y}^2(X_j)} \right\},$$

where $\overline{Y}(u) = n^{-1} \sum Y_i(u)$ and $Y_i(u) = \mathbb{1}(X_i \geq u)$.

Combined Estimator

To improve the efficiency, we combine proposed semiparametric estimator with the KM.

Let $\widehat{\boldsymbol{\mu}} = (\widehat{S}_{\mathrm{Semi}}, \widehat{S}_{\mathrm{KM}})^{\mathsf{T}}$, and $\boldsymbol{\Sigma}$ denote their covariance matrix, then the combined estimator is constructed as

$$(\mathbf{1}^{\mathsf{T}}\boldsymbol{\Sigma}^{-1}\mathbf{1})^{-1}\mathbf{1}^{\mathsf{T}}\boldsymbol{\Sigma}^{-1}\widehat{\boldsymbol{\mu}}$$

where

$$\begin{aligned}
\Sigma_{1,1} &= \mathbf{C}^{\mathsf{T}} E[\mathbb{1}(C_j > t)\{N_j(t) - g(\boldsymbol{\theta}_t^{\mathsf{T}}\mathbf{W}_j)\}^2 \mathbf{W}_j \mathbf{W}_j^{\mathsf{T}}]\mathbf{C}. \\
\Sigma_{2,2} &= E[\mathbb{1}(T_i \wedge t \le C_i)\{\pi_t^{\mathrm{KM}} - \mathbb{1}(X_i < t)\}^2 / G^2(X_i \wedge t)]. \\
\Sigma_{1,2} &= -\mathbf{C}^{\mathsf{T}} E[\mathbb{1}(C_i > t)\{N_i(t) - g(\boldsymbol{\theta}_t^{\mathsf{T}}\mathbf{W}_i)\}\mathbf{W}_i w_i(t)\{\pi_t^{\mathrm{KM}} - \mathbb{1}(X_i < t)\}]
\end{aligned}$$

- unbiased
- smallest possible variance among all linear combinations

Simulation Setup

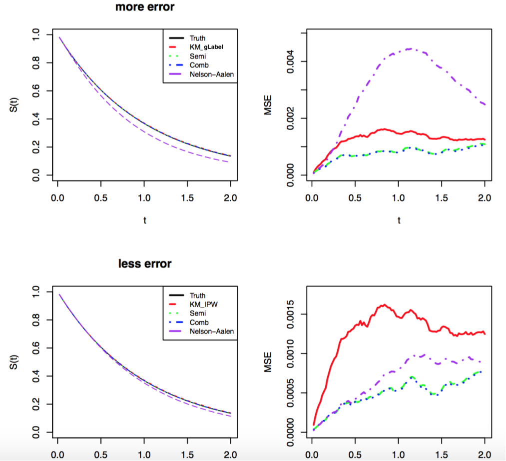
- 400 datasets, each has 200 labeled subjects and 2000 unlabeled subjects
- For each dataset, generate
  - $T_i \sim$ Exponential(1)
  - $C_i \sim$ Exponential(1) or $C_i \sim$ Uniform(3)
  - $Z_i = h(\lambda, T_i, C_i) + e_i$, with $e_i \sim$ Normal$(0, \sigma)$
  - λ large vs. small

Settings

1. $T_i \sim \text{Exponential}(1)$, $C_i \sim \text{Exponential}(1)$,
   $Z_i = \log(T_i) + \lambda \log(C_i) + e_i$, with $e_i \sim \text{Normal}(0, 1)$;
   $\lambda = 1$ and $\lambda = 0.1$.
2. $T_i \sim \text{Exponential}(1)$, $C_i \sim \text{Exponential}(1)$,
   $Z_i = \lambda \log(T_i) + (1 - \lambda) \log(C_i) + e_i$, with $e_i \sim \text{Normal}(0, 1)$;
   $\lambda = 1/2$ versus $\lambda = 0.9$.
3. $T_i \sim \text{Exponential}(1)$, $C_i \sim \text{Exponential}(1)$,
   $Z_i = \lambda \log(T_i) + (1 - \lambda) \log(C_i) + e_i$, with $e_i \sim \text{Normal}(0, 0.25)$;
   $\lambda = 1/2$ versus $\lambda = 0.9$.
4. $T_i \sim \text{Exponential}(1)$, $C_i \sim \text{Uniform}(0, 3)$,
   $Z_i = \lambda \log(T_i) + (1 - \lambda) \log(C_i) + e_i$, with $e_i \sim \text{Normal}(0, 0.25)$;
   $\lambda = 1/2$ versus $\lambda = 0.9$.
5. $T_i \sim \text{Exponential}(1)$, $C_i \sim \text{Uniform}(0, 3)$,
   $Z_i = \log\{\min(T_i^*, C_i)\}$, with $T_i^* = T_i + e_i$ and $e_i \sim \text{Exponential}(\lambda)$;
   $\lambda = 2$ versus $\lambda = 5$.

Simulation Results

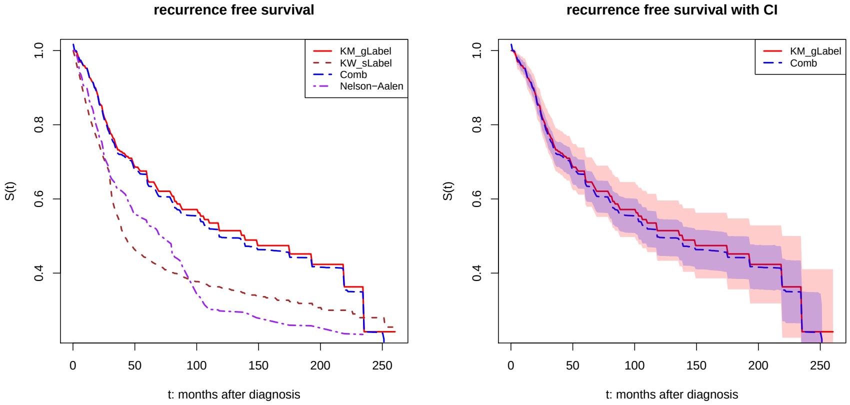
$T_i \sim \text{Exponential}(1),\ C_i \sim \text{Uniform}(0,3),$
$Z_i = \lambda \log(T_i) + (1-\lambda)\log(C_i) + e_i$, with $e_i \sim \text{Normal}(0, 0.25)$;
$\lambda = 1/2$ versus $\lambda = 0.9$



Real data example

- Goal: estimate the recurrence-free survival curve of lung cancer patients
- Dataset
  - 37,021 total lung cancer patients
  - 5K early stage patients
  - 340 had recurrence status and observed time to recurrence labels from chart review (Δ_i, X_i)
  - Surrogate outcomes (Δ_i^*, Z_i) are available for each patient obtained by Uno's two-step estimator
    –binary recurrence status
    –predict event times using peaks of specific features
  - Accuracy of Δ_i^* is only 0.79;

Results - Survival Curve Comparisons



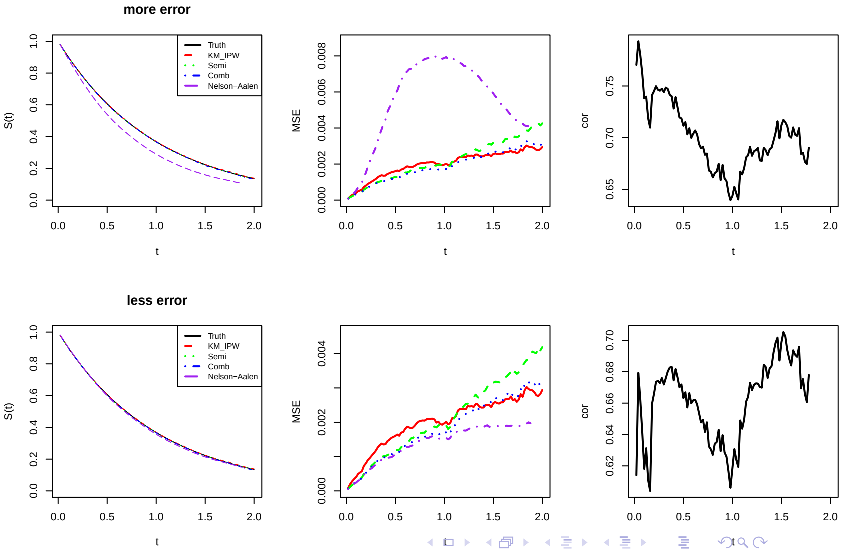
Remarks

- a semi-supervised calibrated survival curve
- fully utilize both labeled and unlabeled data
- next step: estimate the survival function among different risk groups (e.g., different treatment group), and test for difference.

Thank you!

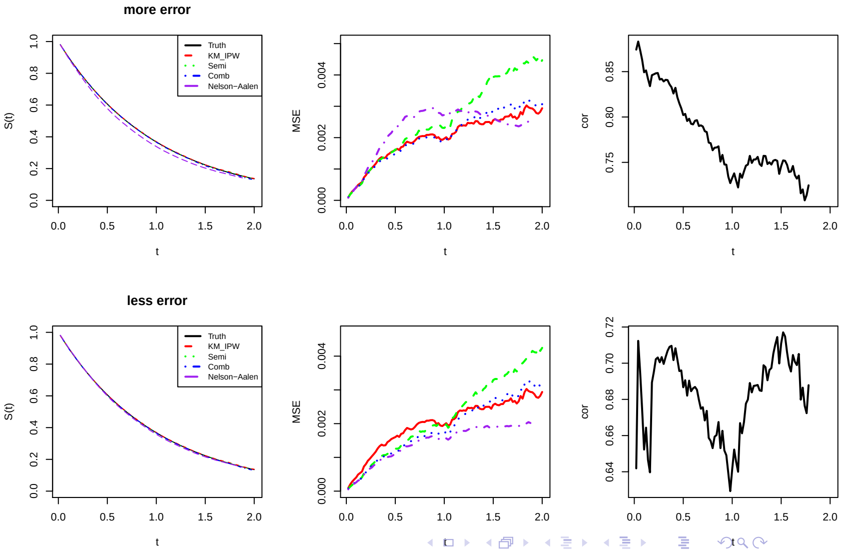
Setting 1

$T_i \sim \text{Exponential}(1), \; C_i \sim \text{Exponential}(1),$
$Z_i = \log(T_i) + \lambda \log(C_i) + e_i$, with $e_i \sim \text{Normal}(0,1)$;
$\lambda = 1$ and $\lambda = 0.1$.



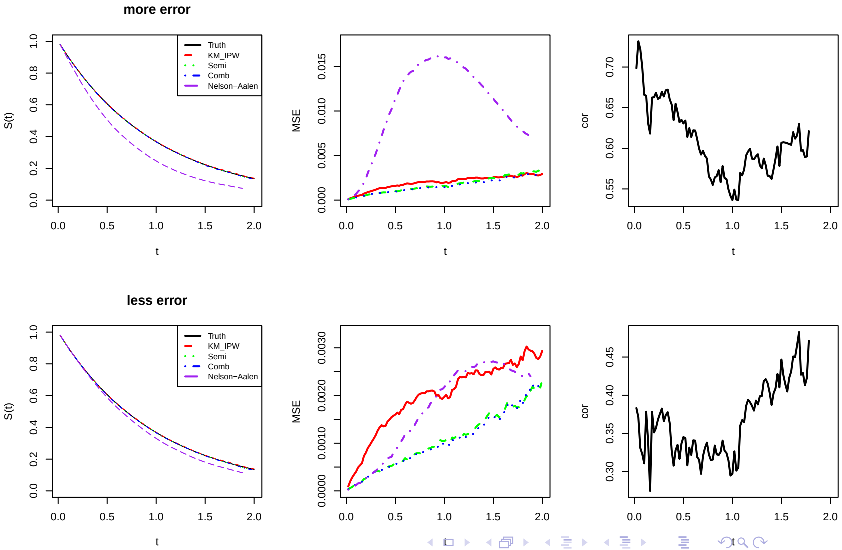
Setting 2

$T_i \sim \text{Exponential}(1),\ C_i \sim \text{Exponential}(1),$
$Z_i = \lambda \log(T_i) + (1-\lambda)\log(C_i) + e_i$, with $e_i \sim \text{Normal}(0,1)$;
$\lambda = 1/2$ versus $\lambda = 0.9$.



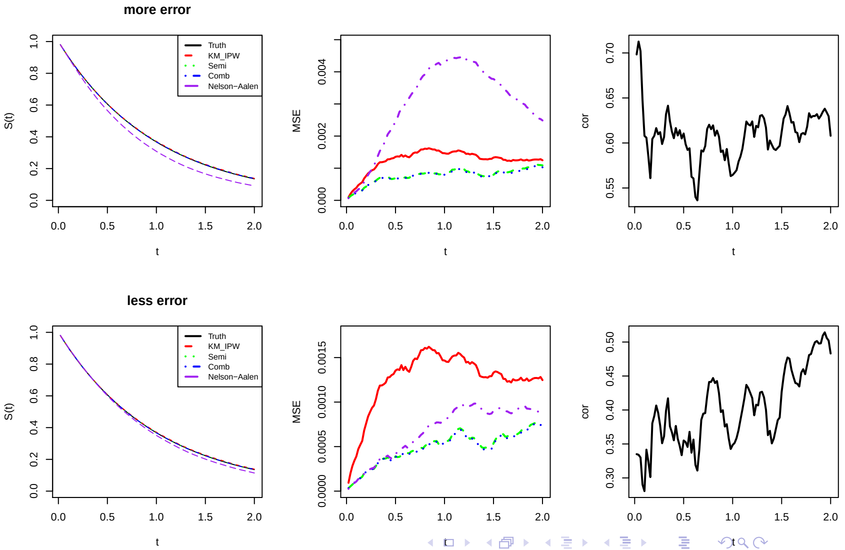
Setting 3

$T_i \sim \text{Exponential}(1),\ C_i \sim \text{Exponential}(1),$
$Z_i = \lambda \log(T_i) + (1 - \lambda) \log(C_i) + e_i$, with $e_i \sim \text{Normal}(0, 0.25)$;
$\lambda = 1/2$ versus $\lambda = 0.9$.



Setting 4

$T_i \sim \text{Exponential}(1),\ C_i \sim \text{Uniform}(0,3),$
$Z_i = \lambda \log(T_i) + (1-\lambda)\log(C_i) + e_i$, with $e_i \sim \text{Normal}(0, 0.25)$;
$\lambda = 1/2$ versus $\lambda = 0.9$.



Setting 5

$T_i \sim \text{Exponential}(1)$, $C_i \sim \text{Uniform}(0,3)$,
$Z_i = \log\{\min(T_i^*, C_i)\}$, with $T_i^* = T_i + e_i$ and $e_i \sim \text{Exponential}(\lambda)$;
$\lambda = 2$ versus $\lambda = 5$.