Identify shortcomings of Estimators of Discriminative Performance in Time-to-Event Analyses: A Comparison Study

Ying Jin, Andrew Leroux

February 11, 2022

Introduction

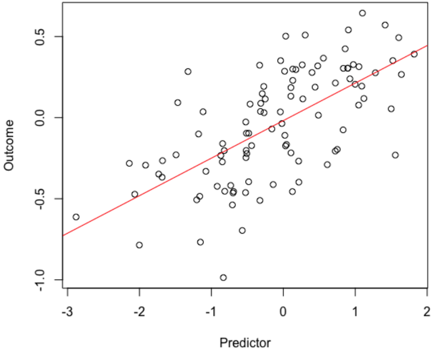
Question

If the evaluation metric reveals perfect out-of-sample performance of the underlying model, should we conclude that the model fits the population well?

Introduction

Example

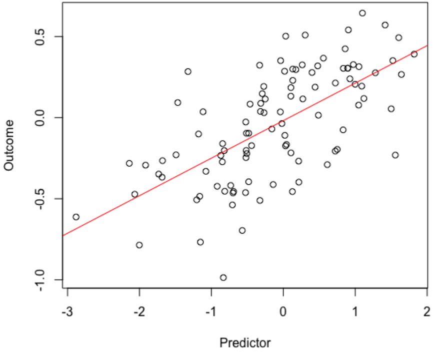
A linear regression model with out-of-sample $R^2 = 1$ or MSE $= 0$?

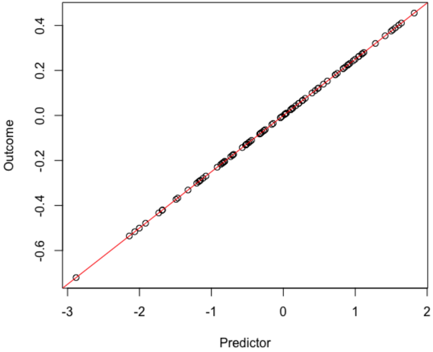


Introduction

Example

A linear regression model with out-of-sample $R^2 = 1$ or MSE $= 0$?

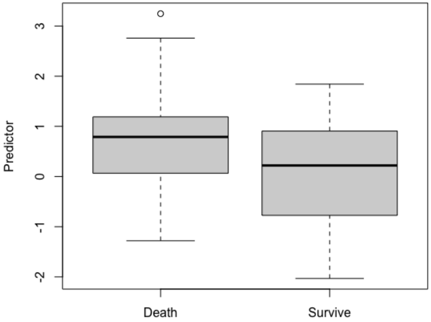




Introduction

Example

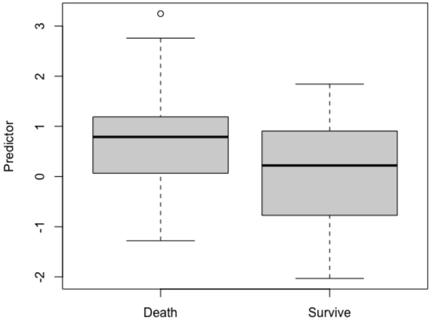
A logistic regression model with out-of-sample AUC $= 1$, indicating perfect separation?

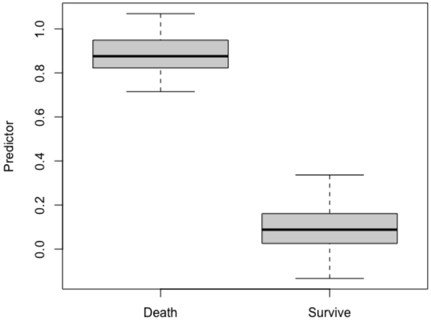


Introduction

Example

A logistic regression model with out-of-sample AUC $= 1$, indicating perfect separation?



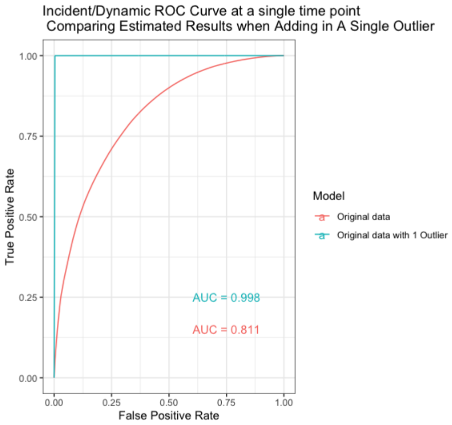


Introduction

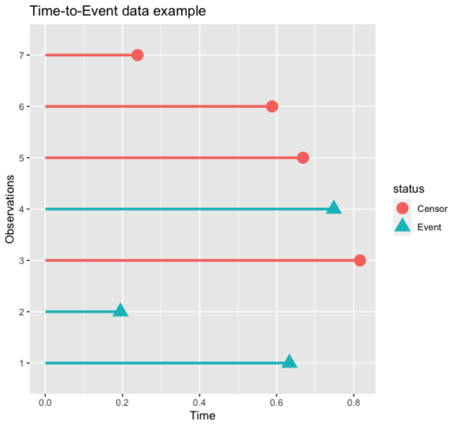
Alert

This can actually happen when we have noisy or abnormal observations!

- Example: With the same model, adding one outlier inflates the AUC estimates to 1
- Estimator does not reflect true prediction performance of model



Time-to-event data



- Observed outcomes include timing and occurrence of a specific event.
  e.g. death, disease onset, relapse, etc.
- Observations are usually subject to right censoring, which means event has not happened until censoring time, but no information is available after censoring.
  e.g. Fixed follow-up period, patient dropout, etc

Notation

- let i = 1...N indicates individuals
- For each individual, T_i^* is true event time subject to independent right censoring; C_i is censoring time
- Observed time $T_i = min(T_i^*, C_i)$; observed status $\delta_i = I(T_i^* < C_i)$
- Observed covariate $\boldsymbol{X}_i = (X_{i1}, ...X_{ip})^t$; risk score $\eta_i = \boldsymbol{X}_i^t\boldsymbol{\beta}$
- Assume survival time is associated with risk score through a propotional hazard model as follows:

$$\begin{aligned}\log \lambda_i(t|\boldsymbol{X}_i) &= \log \lambda_0(t) + \boldsymbol{X}_i^t\boldsymbol{\beta}\ , \qquad t > 0\\ &= \log \lambda_0(t) + \eta_i\end{aligned}$$

  where $\lambda_0(t)$ is constant across population
- We aim to evaluate discriminative performance of models, which is the ability to predict timing of event with covariates

Estimands: Time-varying AUC

We use Incident/dynamic AUC to measure discriminative performance of a model.
At a specific time t and a fixed threshold of risk score c:

- Incident sensitivity

$$\text{sensitivity}^{\mathbb{I}}(c,t) = \text{TP}_t^{\mathbb{I}}(c) = \Pr(\eta_i > c \mid T_i^* = t)$$

Estimands: Time-varying AUC

We use Incident/dynamic AUC to measure discriminative performance of a model.
At a specific time t and a fixed threshold of risk score c:

- Incident sensitivity

$$\text{sensitivity}^{\mathbb{I}}(c, t) = \text{TP}_t^{\mathbb{I}}(c) = \Pr(\eta_i > c | T_i^* = t)$$

- Dynamic specificity

$$1 - \text{specificity}^{\mathbb{D}}(c, t) = \text{FP}_t^{\mathbb{D}}(c) = 1 - \Pr(\eta_i \leq c | T_i^* > t)$$

Estimands: Time-varying AUC

We use Incident/dynamic AUC to measure discriminative performance of a model.
At a specific time t and a fixed threshold of risk score c:

- Incident sensitivity

$$\text{sensitivity}^{\mathbb{I}}(c,t) = \text{TP}_t^{\mathbb{I}}(c) = \Pr(\eta_i > c \mid T_i^* = t)$$

- Dynamic specificity

$$1 - \text{specificity}^{\mathbb{D}}(c,t) = \text{FP}_t^{\mathbb{D}}(c) = 1 - \Pr(\eta_i \leq c \mid T_i^* > t)$$

- For all possible thresholds c: $\text{ROC}_t^{\mathbb{I}/\mathbb{D}}(p) = \text{TP}_t^{\mathbb{I}}\{[\text{FP}_t^{\mathbb{D}}]^{-1}(p)\}$

Estimands: Time-varying AUC

We use Incident/dynamic AUC to measure discriminative performance of a model.
At a specific time t and a fixed threshold of risk score c:

- Incident sensitivity

$$\text{sensitivity}^{\mathbb{I}}(c,t) = \text{TP}_t^{\mathbb{I}}(c) = \Pr(\eta_i > c \,|\, T_i^* = t)$$

- Dynamic specificity

$$1 - \text{specificity}^{\mathbb{D}}(c,t) = \text{FP}_t^{\mathbb{D}}(c) = 1 - \Pr(\eta_i \leq c \,|\, T_i^* > t)$$

- For all possible thresholds c: $\text{ROC}_t^{\mathbb{I}/\mathbb{D}}(p) = \text{TP}_t^{\mathbb{I}}\{[\text{FP}_t^{\mathbb{D}}]^{-1}(p)\}$
- Incident/dynamic AUC: $\text{AUC}^{\mathbb{I}/\mathbb{D}}(t) = \int_0^1 \text{ROC}_t^{\mathbb{I}/\mathbb{D}}(p)dp$

Estimands: Concordance

We use concordance to measure discriminative performance of model over the entire follow-up period:

- Concordance

$$C = \Pr(\eta_i < \eta_j | T_i^* > T_j^*) \ .$$

Estimands: Concordance

We use concordance to measure discriminative performance of model over the entire follow-up period:

- Concordance

$$C = \Pr(\eta_i < \eta_j | T_i^* > T_j^*) \ .$$

- Concordance subject to administrative censoring

$$C^\tau = \Pr(\eta_i < \eta_j | T_i^* > T_j^*, T_j^* < \tau) \ ,$$

  where τ is the end of follow-up period

Estimands: Concordance

We use concordance to measure discriminative performance of model over the entire follow-up period:

- Concordance

$$C = \Pr(\eta_i < \eta_j | T_i^* > T_j^*) \ .$$

- Concordance subject to administrative censoring

$$C^\tau = \Pr(\eta_i < \eta_j | T_i^* > T_j^*, T_j^* < \tau) \ ,$$

where τ is the end of follow-up period

- Concordance by integrating AUC

$$\mathrm{C}^\tau = \int_0^\tau \mathrm{AUC}^{\mathbb{I}/\mathbb{D}}(t) w^\tau(t) dt$$

where $w^\tau(t) = 2f(t)S(t)/1 - S^2(\tau)$

Estimators

- Now let's talk about ways to estimator these evaluation metrics.
- Many estimators proposed, roughly categorized as semi-parametric and non-parametric.
- Semi-parametric estimators suffer from out-of-sample inflation
- Non-parametric estimators are highly variable

Estimators

Incident/dynamic AUC: at time t and a specific threshold for risk threshold c,

- Dynamic False-positive rate

$$\hat{\mathrm{FP}}_t^{\mathbb{D}}(c) = \frac{\sum_k I(\eta_k > c) I(T_k > t)}{\sum_j I(T_j > t)}$$

Estimators

Incident/dynamic AUC: at time t and a specific threshold for risk threshold c,

- Dynamic False-positive rate

$$\hat{\mathrm{FP}}_t^{\mathbb{D}}(c) = \frac{\sum_k I(\eta_k > c) I(T_k > t)}{\sum_j I(T_j > t)}$$

- Incident True-positive rate:
  Non-parametric:

$$\hat{\mathrm{TP}}_t^{\mathbb{I}}(c) = \frac{\sum_k I(\eta_k > c) I(T_k = t) I(\delta_k = 1)}{\sum_j I(T_j = t) I(\delta_j = 1)}$$

Estimators

Incident/dynamic AUC: at time t and a specific threshold for risk threshold c,

- Dynamic False-positive rate

$$\hat{\mathrm{FP}}_t^{\mathbb{D}}(c) = \frac{\sum_k I(\eta_k > c) I(T_k > t)}{\sum_j I(T_j > t)}$$

- Incident True-positive rate:
  Non-parametric:

$$\hat{\mathrm{TP}}_t^{\mathbb{I}}(c) = \frac{\sum_k I(\eta_k > c) I(T_k = t) I(\delta_k = 1)}{\sum_j I(T_j = t) I(\delta_j = 1)}$$

  Semi-parametric [Heagerty and Zheng, 2005]:

$$\hat{\mathrm{TP}}_t^{\mathbb{I}}(c) = \frac{\sum_k I(\eta_k > c) I(T_k \geq t) exp(\eta_k)}{\sum_j I(T_j \geq t) exp(\eta_j)}$$

Estimators

Concordance:

- Integrating $\text{AUC}^{\mathbb{I}/\mathbb{D}}(t)$ estimates

$$\text{C}^{\tau} = \int_0^{\tau} \text{AUC}^{\mathbb{I}/\mathbb{D}}(t) w^{\tau}(t) dt$$

  Requires estimates of $\text{AUC}^{\mathbb{I}/\mathbb{D}}(t)$, $S(t)$ and $f(t)$

Estimators

Concordance:

- Integrating $\text{AUC}^{\mathbb{I}/\mathbb{D}}(t)$ estimates

$$C^{\tau} = \int_0^{\tau} \text{AUC}^{\mathbb{I}/\mathbb{D}}(t) w^{\tau}(t) dt$$

  Requires estimates of $\text{AUC}^{\mathbb{I}/\mathbb{D}}(t)$, $S(t)$ and $f(t)$

- Gonen-Heller [Gonen and Heller, 2005]: semi-parametric

$$C = \frac{2}{n(n-1)} \sum_{i<j} \frac{I(\eta_j - \eta_i < 0)}{1 + exp(\eta_j - \eta_i)} + \frac{I(\eta_i - \eta_j < 0)}{1 + exp(\eta_i - \eta_j)}$$

Estimators

Concordance:

- Integrating $\text{AUC}^{\mathbb{I}/\mathbb{D}}(t)$ estimates

$$C^{\tau} = \int_0^{\tau} \text{AUC}^{\mathbb{I}/\mathbb{D}}(t) w^{\tau}(t) dt$$

  Requires estimates of $\text{AUC}^{\mathbb{I}/\mathbb{D}}(t)$, $S(t)$ and $f(t)$

- Gonen-Heller [Gonen and Heller, 2005]: semi-parametric

$$C = \frac{2}{n(n-1)} \sum_{i<j} \frac{I(\eta_j - \eta_i < 0)}{1 + exp(\eta_j - \eta_i)} + \frac{I(\eta_i - \eta_j < 0)}{1 + exp(\eta_i - \eta_j)}$$

- Harrell's C index: non-parametric

$$C = \frac{\sum_{i<j} I(T_i < T_j) I(\eta_i > \eta_j) I(\delta_i = 1) + I(T_i > T_j) I(\eta_i < \eta_j) I(\delta_j = 1)}{\sum_{i<j} I(T_i < T_j) I(\delta_i = 1) + I(T_i > T_j) I(\delta_j = 1)}$$

Simulation setup

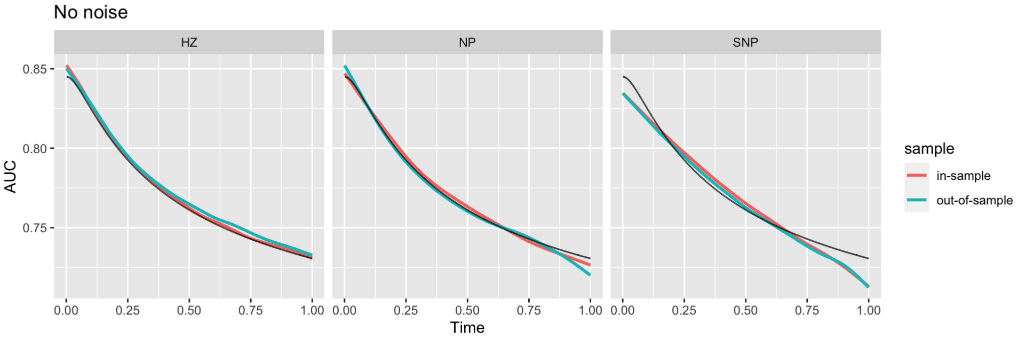
- 1000 simulated data sets with sample size N $=$ 500
- Generate three signals independently from N(0, 1)
- Survival generated from Cox propotional hazard model with Weibull baseline hazard

$$\log \lambda_i(t|\boldsymbol{X}_i) = \log p\theta t^{p-1} + X_{i1}\beta_1 + X_{i2}\beta_2 + X_{i3}\beta_3$$

- Censoring time generated uniformly from (0.5, 1)

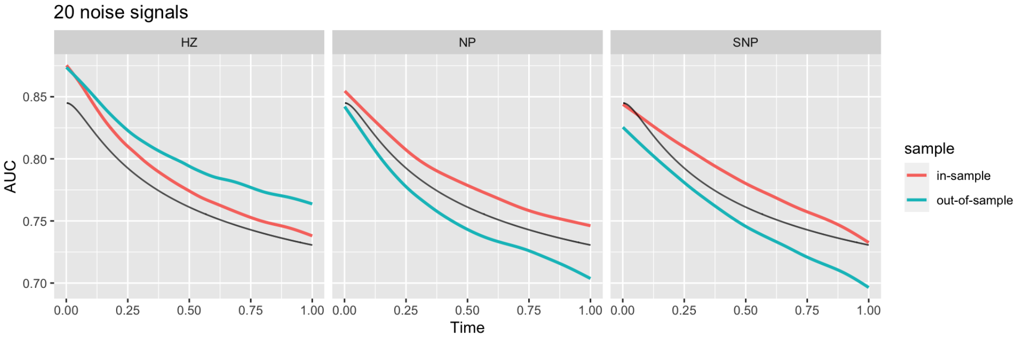
Results: Incident/Dynamic AUC

Note: Incident/Dynamic AUC is smoothed across all simulation



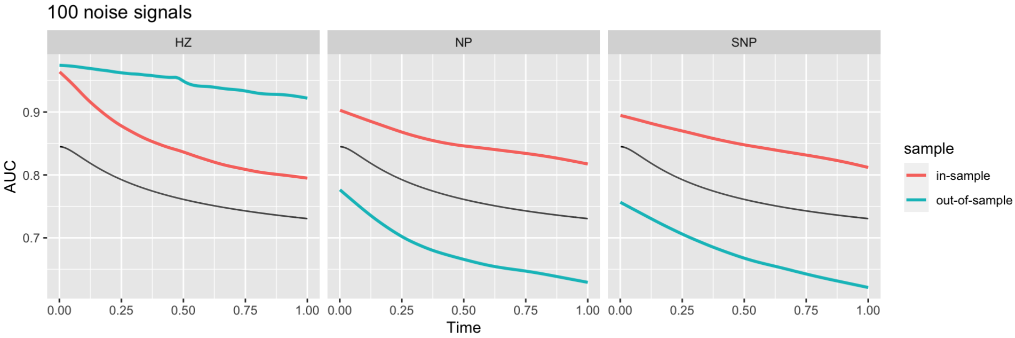
Results: Incident/Dynamic AUC

Note: Incident/Dynamic AUC is smoothed across all simulation

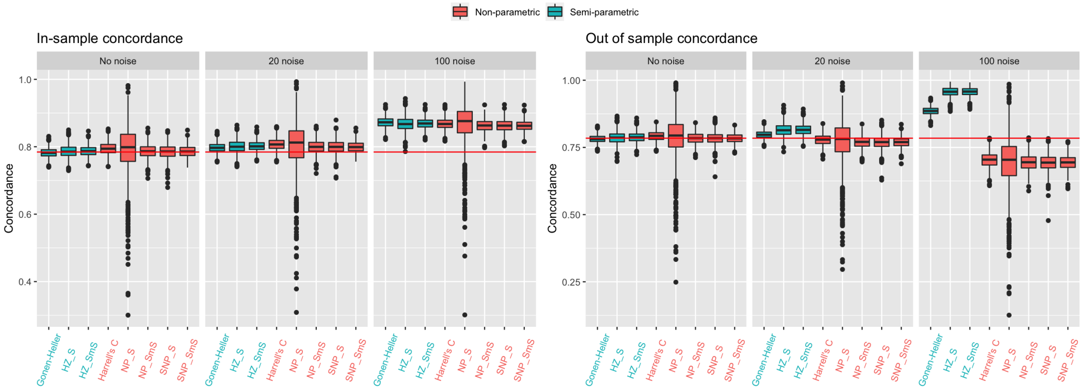


Results: Incident/Dynamic AUC

Note: Incident/Dynamic AUC is smoothed across all simulation



Results: Concordance



Conclusion

Estimator	Bias	Variability	Inflation with noise
Semi-parametric (HZ)	No bias	Small	Large
Non-parametric (NP)	No bias	Large	No inflation
Smoothed non-parametric (SNP)	Underestimate at edges and overestimate in the middle	Medium	No inflation

Comparison of I/D AUC estimators

Conclusion

Estimator	**Bias**	**Variability**	**Inflation with noise**
Semi-parametric	No bias	Small	Large
Non-parametric	Overestimation	Large	No inflation
Non-parametric with smoothed weight	No bias	Medium	No inflation

Comparison of concordance estimators

Data Application

- Data:
  - NHANES 2011-2014
  - 3556 participants age 50-80
  - Analytic dataset
    - Complete data: age, mortality follow-up, BMI, PIR
    - ≥ 3 days of accelerometry data with 95% wear
- Goals:
  - Predict time to all-cause mortality using PA features and other covariate data using a complicated mean model
  - Compare estimated in- and out-of-sample performance for discrimination
  - Compare results to a linear regression and $\hat{R}^2$

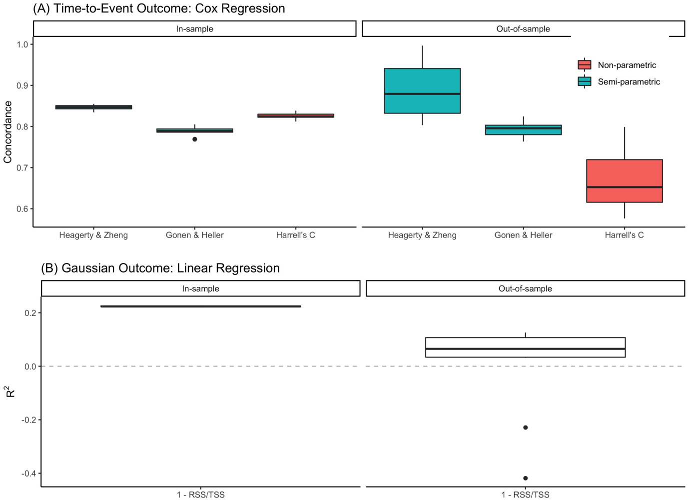
Data Application

- Suppose we have a covariate vector $[x_{i1}, \ldots, x_{i6}]^t = \boldsymbol{X}_i \in \mathbb{R}^6$
- Models:

$$\text{Cox Regression: } \log \lambda_(t|X_i) = \log \lambda_0(t) + f(\boldsymbol{X}_i)$$
$$\text{Linear Regression: } E[\text{age}_i] = \beta_0 + f(\boldsymbol{X}_i)$$

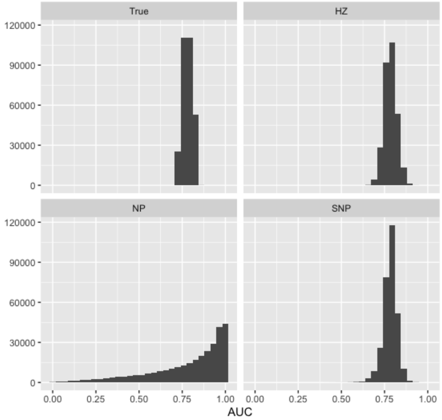
- $f(\cdot)$ is a smooth function of the 6 covariates modelled as a linear combination of 200 (penalized) thin plate regression splines
- 200 coefficients (parameters) with $N = 3556$ will tend to overfit, particularly for non-Gaussian models (e.g. log hazard)

Data Application



Discussion

- Use ridge regression to reduce overfit
- Mitigate bias of smoothed non-parametric estimator (SNP) of $\mathrm{AUC}^{\mathbb{I}/\mathbb{D}}(t)$
  - Skewed distribution of $\mathrm{AUC}^{\mathbb{I}/\mathbb{D}}(t)$
  - Neither transformation of outcome nor weighted regression by variance was helpful
  - Bounded regression led to slight improvement
- Smoothness of survival function seems to reduce bias of non-parametric concordance estimator



References

Gonen, M. and Heller, G. (2005).
Concordance probability and discriminatory power in proportional hazards regression.
Biometrika, 92(4):965–970.

Heagerty, P. J. and Zheng, Y. (2005).
Survival model predictive accuracy and roc curves.
Biometrics, 61(1):92–105.

Uno, H., Cai, T., Pencinac, M. J., D'Agostinod, R. B., and Weib, L. J. (2011).
On the c-statistics for evaluating overall adequacy of risk prediction procedures with censored survival data.
Statistics in medicine, 30(10), 1105–1117. https://doi.org/10.1002/sim.4154, 30(10):1105–1117.

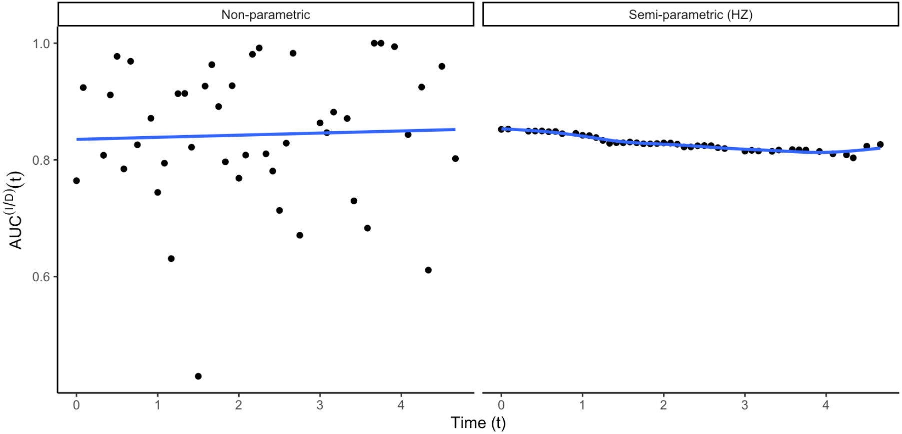
Thank you!

Supplemental Material

- Methodologic contributions/work for follow-up papers
  - Propose "optimal" estimation procedures for measures of discrimination
  - Propose a method for deriving true $\mathrm{AUC}^{\mathbb{I}/\mathbb{D}}(t)$ to use for comparisons in simulation
- Estimation procedures for discrimination
  - For estimating C^{τ} using integrated AUC we need to estimate the quantities
    1. $\mathrm{AUC}^{\mathbb{I}/\mathbb{D}}(t)$
    2. $w^{\tau}(t)$
       a. $S(t)$
       b. $f(t)$
  - We propose to estimate each quantity using penalized regression splines
- Deriving true $\mathrm{AUC}^{\mathbb{I}/\mathbb{D}}(t)$
  - Some Bayes
  - Some numeric approximations

Supplemental Material: Estimating $\text{AUC}^{\mathbb{I}/\mathbb{D}}(t)$

- Clearly the SP estimator of $\text{AUC}^{\mathbb{I}/\mathbb{D}}(t)$ is inappropriate, but it is relatively "stable"
- NP estimator is highly variable



Supplemental Material: Estimating $\text{AUC}^{\mathbb{I}/\mathbb{D}}(t)$

- Numeric integration a noisy function is generally a bad idea
- Integrate a smoothed version of $\text{AUC}^{\mathbb{I}/\mathbb{D}}(t)$ instead (blue line) instead
- Penalized regression splines

$$\begin{aligned}\widehat{\text{AUC}}^{\mathbb{I}/\mathbb{D}}(t) &= f_0(t) + \epsilon(t) \\ &= \sum_{k=1}^{K} \xi_k B_k(t) + \epsilon(t)\end{aligned}$$

Subject to a second derivative penalty on $f_0(t)$

- $\hat{f}_0(t) = \widetilde{\text{AUC}}^{\mathbb{I}/\mathbb{D}}(t)$ is the smoothed estimate of the non-parametric estimator

Supplemental Material: Estimating $\text{AUC}^{\mathbb{I}/\mathbb{D}}(t)$

- Some issues
  1. Residuals $\epsilon(t)$ non-Gaussian (data bounded)
  2. Data are correlated: $\widehat{\text{AUC}}^{\mathbb{I}/\mathbb{D}}(t_1)$, $\widehat{\text{AUC}}^{\mathbb{I}/\mathbb{D}}(t_2)$ calculated using overlapping individuals (intersection of $R(t_1), R(t_2)$)
  3. $\text{Var}(\widehat{\text{AUC}}^{\mathbb{I}/\mathbb{D}}(t_1)) < \text{Var}(\widehat{\text{AUC}}^{\mathbb{I}/\mathbb{D}}(t_2))$ for $t_1 < t_2$
  4. Biased results
- Some possible solutions (work in progress)
  - Non-Gaussian data
    - Response transformation
    - Generalized additive model
  - Correlated/heteroskedastic data
    - (easier) Bootstrap estimates of covariance/bias
    - (harder) derive distributional results

Supplemental Material: Estimating $S(t), f(t)$

- Same strategy as with $\widehat{\text{AUC}}^{\mathbb{I}/\mathbb{D}}(t_1)$
- Smooth the Kaplan-Meier Estimator

$$\hat{S}(t) = \prod_{i:t_i \leq t} \left(1 - \frac{d_i}{|R_i(t)|}\right)$$

- Subject to
  - Montonicity constraints ($S(t_1) > S(t_2)$ for $t_1 < t_2$)
  - Point constraint $S(0) = 1$
  - Positivity ($S(t) > 0$)
- For $f(t)$, given an estimate for $\tilde{S}(t)$

$$\tilde{f}(t) = -\frac{d}{dt}\tilde{S}(t)$$

Supplemental Material: Deriving True $\text{AUC}^{\mathbb{I}/\mathbb{D}}(t)$

Non-trivial to obtain "true" $\text{AUC}^{\mathbb{I}/\mathbb{D}}(t)$ except from extremely simple models. Here we propose a general solution. Recall we simulate from the model

$$f(t|\eta) = h(t|\eta)S(t|\eta)$$
$$\eta \sim N(0, \sigma_\eta^2)$$

With Weibull baseline hazard this becomes

$$= (\theta e^\eta)\gamma t^{\gamma-1} e^{-(\theta e^\eta)t^\gamma}$$

Supplemental Material: Deriving True AUC$^{\mathbb{I}/\mathbb{D}}(t)$

Incident sensitivity

$$\begin{aligned}
\Pr(\eta_i > c | T_i = t) &= E[1(\eta > c) | T = t] \\
&= \int 1(\eta > c) f(\eta | t) d\eta \\
&= \int 1(\eta > c) \frac{f(t|\eta) f(\eta)}{\int f(t|\eta) f(\eta) d\eta} d\eta
\end{aligned}$$

Dynamic specificity

$$\begin{aligned}
\Pr(\eta_i \leq c | T_i > t) &= \frac{\Pr(\eta_i \leq c \cap T_i > t)}{\Pr(T_i > t)} \\
&= \frac{\int_t^\infty \int_{-\infty}^c f(t|\eta) f(\eta) d\eta dt}{\int_t^\infty [\int f(t|\eta) f(\eta) d\eta] dt}
\end{aligned}$$

Supplemental Material: Deriving True $\text{AUC}^{\mathbb{I}/\mathbb{D}}(t)$

- For any t we can estimate $\Pr(\eta_i > c | T_i = t)$, $\Pr(\eta_i \leq c | T_i > t)$ for a range of c
- Estimate $\text{AUC}^{\mathbb{I}/\mathbb{D}}(t)$ as

$$
\begin{aligned}
\text{AUC}^{\mathbb{I}/\mathbb{D}}(t) &= \int_0^1 \text{ROC}_t^{\mathbb{I}/\mathbb{D}}(p)dp \\
&= \int_0^1 \text{TP}_t^{\mathbb{I}}\{[\text{FP}_t^{\mathbb{D}}]^{-1}(p)\}dp \\
&\approx \sum_l \delta_l \text{TP}_t^{\mathbb{I}}\{[\text{FP}_t^{\mathbb{D}}]^{-1}(l)\}
\end{aligned}
$$

- Where δ_l are quadrature weights