UNIVERSITY OF ALBERTA
SCHOOL OF PUBLIC HEALTH

This is PUBLIC HEALTH

# A Summary Index of Prediction Accuracy for Binary and Censored Time to Event Data

Yan Yuan, PhD

April 23, 2019

OICR Biostatistics Seminar

# Clinical Prediction: Examples of Prevention and Planning

- WHO risk charts for cardiovascular disease for most countries
- Numerous risk score systems (n>40) for diabetes risk in general population
- Sepsis risk prediction (CMAJ 2019)

# Risk Score as a Screening Tool

- Characteristics of typical condition that risk scores are developed for
  - seriousness (may result in mortality or significantly affect the quality of life);
  - early detection/intervention can make a difference in disease prognosis but may be expensive or invasive;
  - the event rate is low

# Motivating Data – Binary outcome

Digital Mammography Imaging Screening Trial (Pisano et al. 2005 *New England Journal of Medicine*)

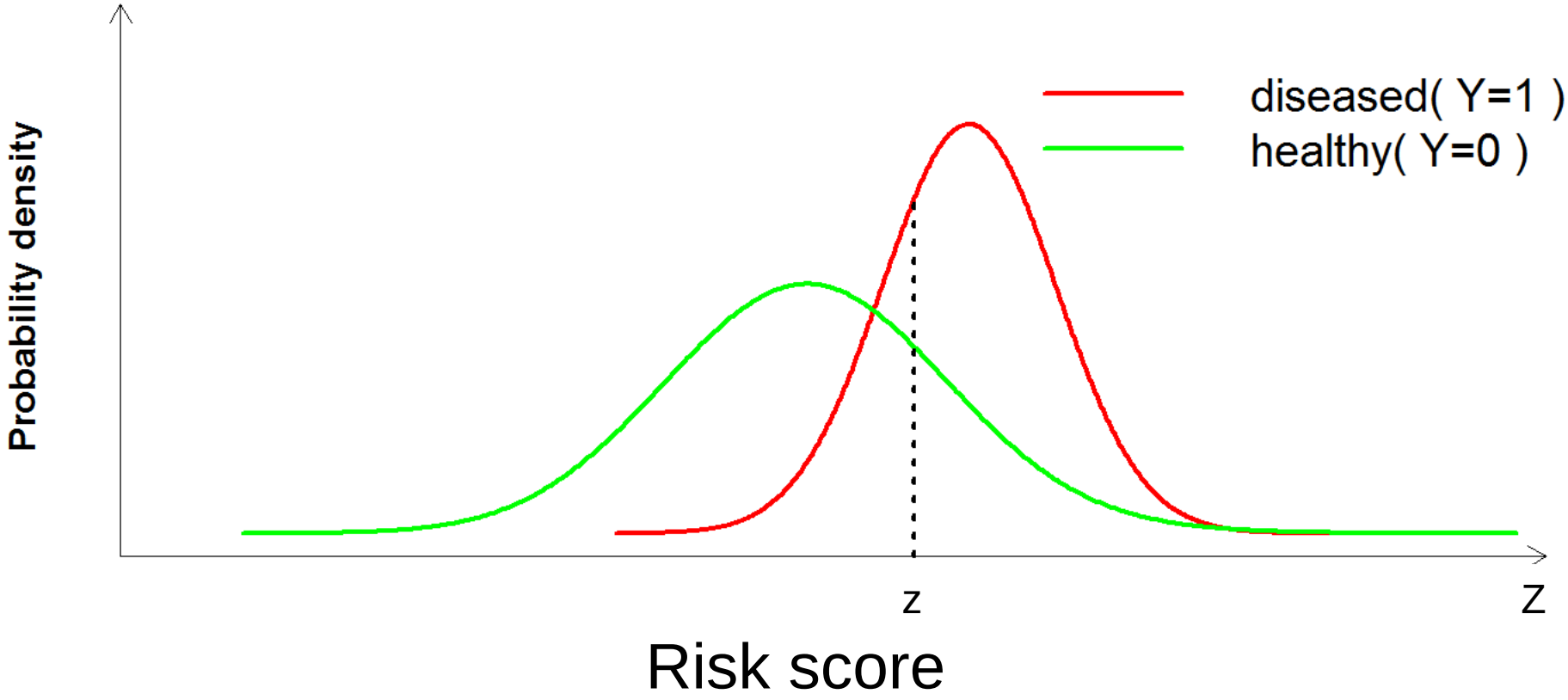
| Malignancy score | | 7 | 6 | 5 | 4 | 3 | 2 | 1 | Total |
|---|---|---|---|---|---|---|---|---|---|
| Digital M | Category Total | 11 | 29 | 69 | 1061 | 2224 | 6588 | 32588 | 42570 |
| | Cancers | 10 | 18 | 25 | 85 | 49 | 25 | 122 | 334 |
| Film M | Category Total | 17 | 29 | 70 | 942 | 2291 | 6910 | 32486 | 42745 |
| | Cancers | 13 | 24 | 25 | 74 | 35 | 33 | 131 | 335 |

**42,760 screening participants underwent two screening technology, 335 were diagnosed with breast cancer by the end of 15 months follow-up.**

# Evaluating Model Performance when Predicting Low Prevalence Events

- Threshold Dependent Measure (predictor needs to be binary)
  - Misclassification rate
  - Sensitivity (TPF): P(test positive | disease present) = $\mathrm{P}(\hat{Y} = 1 \,|Y = 1)$
  - Specificity (FPF): P(test negative | disease absent) = $\mathrm{P}(\hat{Y} = 0 \,|Y = 0)$
  - Positive Predictive value (PPV): $\mathrm{P}(Y = 1|\hat{Y} = 1)$
  - Negative Predictive Value (NPV): $\mathrm{P}(Y = 0|\hat{Y} = 0)$

# When risk score is continuous or ordinal



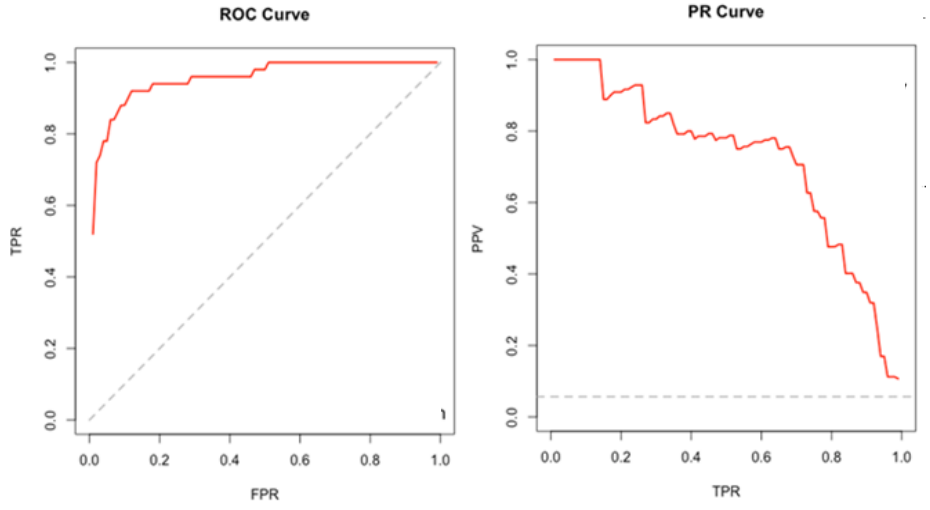
# Threshold-free Summary Measure

- Area Under the ROC Curve (AUC)

$$\mathrm{AUC} \equiv \int_R \mathrm{TPF}(z) d\mathrm{FPF}(z)$$

- Area under the Precision-Recall curve

$$AP \equiv \int_R \mathrm{PPV}(z) d\mathrm{TPF}(z)$$



# MLE of AP

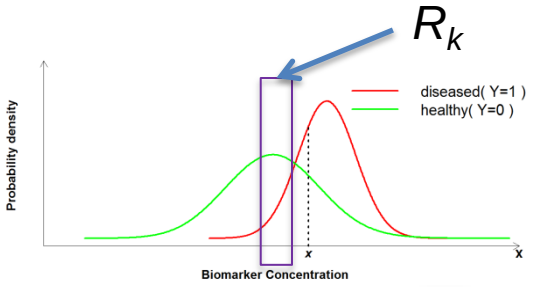
| Score | $x_1$ | $>$ | $x_2$ | $> \cdots >$ | $x_k$ | $>$ | $x_{k+1}$ | $> \cdots >$ | $x_K$ | |
|---|---|---|---|---|---|---|---|---|---|---|
| Partition | $R_1$ | | $R_2$ | $\cdots$ | $R_k$ | | $R_{k+1}$ | $\cdots$ | $R_K$ | Total |
| Class-1 | $Z_1$ | | $Z_2$ | $\cdots$ | $Z_k$ | | $Z_{k+1}$ | $\cdots$ | $Z_K$ | $n_1$ |
| Class-0 | $\bar{Z}_1$ | | $\bar{Z}_2$ | $\cdots$ | $\bar{Z}_k$ | | $\bar{Z}_{k+1}$ | $\cdots$ | $\bar{Z}_K$ | $n_0$ |
| Total | $S_1$ | | $S_2$ | $\cdots$ | $S_k$ | | $S_{k+1}$ | $\cdots$ | $S_K$ | $n$ |

Data in the above 2 X K table follow

$$
\begin{aligned}
(Z_1, Z_2, ..., Z_K)|n_1 &\sim \text{multinomial}(n_1; p_1, p_2, ..., p_K), \\
(\bar{Z}_1, \bar{Z}_2, ..., \bar{Z}_K)|n_1 &\sim \text{multinomial}(n - n_1; q_1, q_2, ..., q_K), \\
n_1 &\sim \text{binomial}(n, \pi),
\end{aligned}
$$

For continuous risk scores

$$p_k = \int_{R_k} f_1(x)dx, \quad q_k = \int_{R_k} f_0(x)dx,$$






*Yuan et al. (2015)*

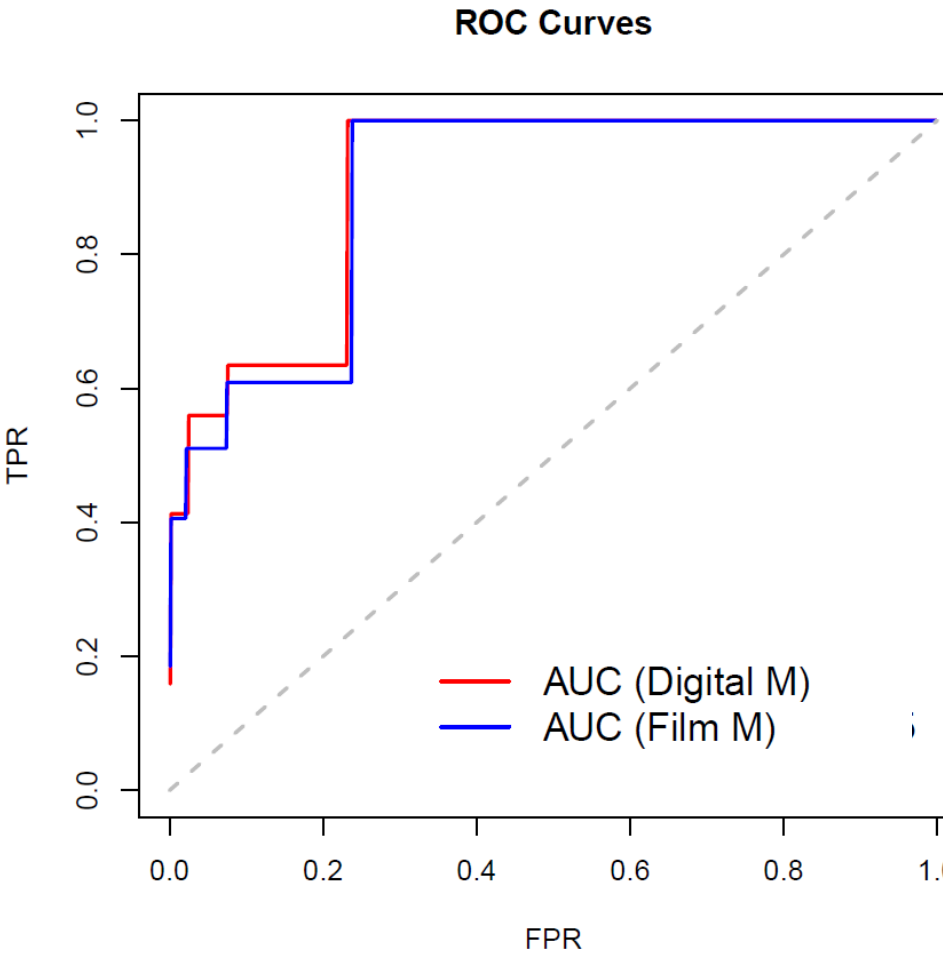
# Asymptotic Variance of AP

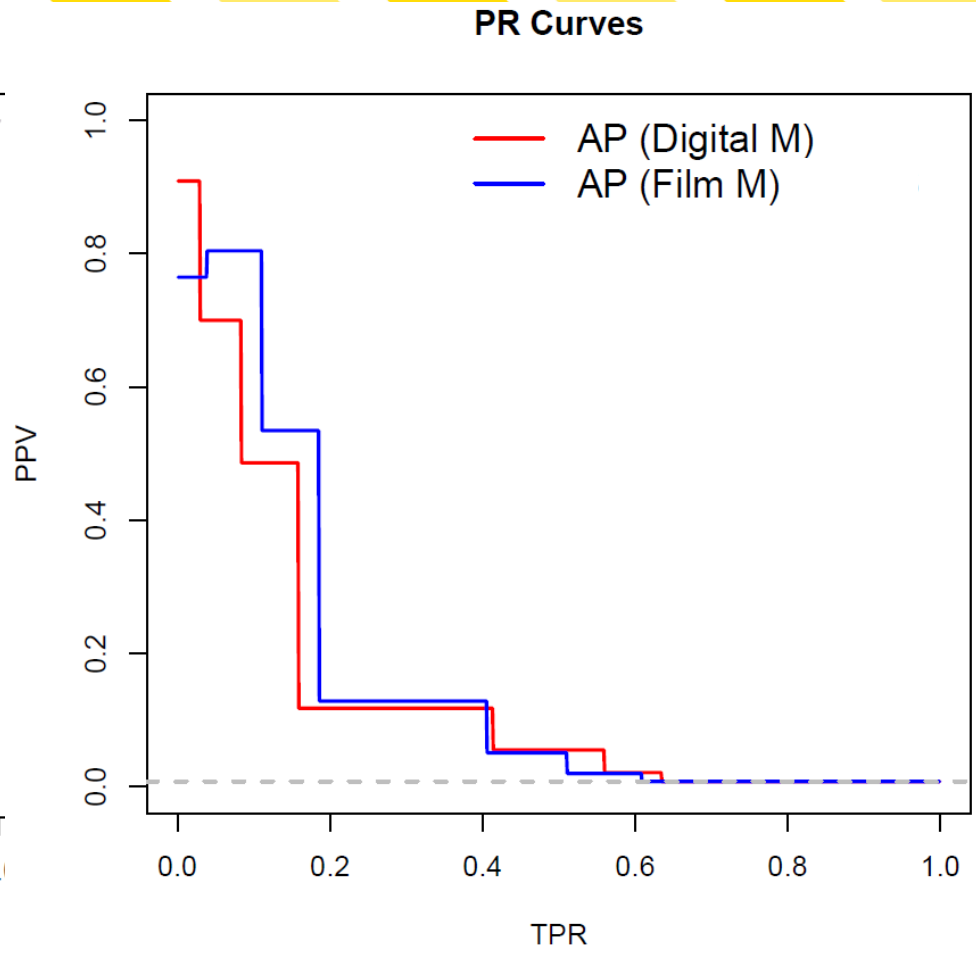
$$\widehat{\mathrm{AP}} = g(\widehat{p}_k, \widehat{q}_k, \widehat{\pi}) = \sum_{k=1}^{K} \left[ \widehat{p}_k \left( \frac{\widehat{\pi} \sum_{k' \le k} \widehat{p}_{k'}}{\widehat{\pi} \sum_{k' \le k} \widehat{p}_{k'} + (1 - \widehat{\pi}) \sum_{k' \le k} \widehat{q}_{k'}} \right) \right]$$

Applying the Delta method, we get the variance estimator

$$\widehat{var}\left(\widehat{AP}\right) \approx (\nabla g)^T \hat{J}^{-1} (\nabla g)$$

*Yuan et al. (2015)*



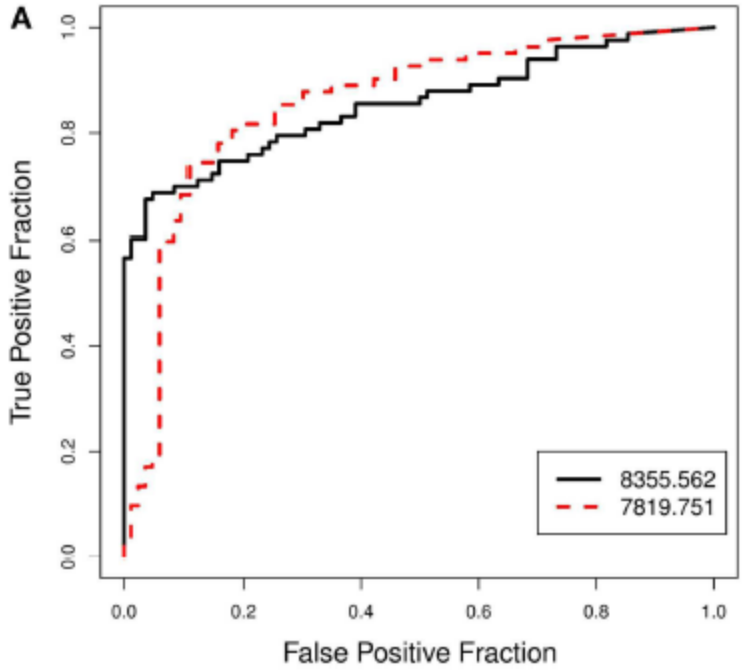

| Malignancy score | | 7 | 6 | 5 | 4 | 3 | 2 | 1 | Total |
|---|---|---|---|---|---|---|---|---|---|
| Digital M | Category Total | 11 | 29 | 69 | 1061 | 2224 | 6588 | 32588 | 42570 |
| | Cancers | 10 | 18 | 25 | 85 | 49 | 25 | 122 | 334 |
| Film M | Category Total | 17 | 29 | 70 | 942 | 2291 | 6910 | 32486 | 42745 |
| | Cancers | 13 | 24 | 25 | 74 | 35 | 33 | 131 | 335 |

Given that 335 breast cancer diagnosed in 42,760 screening participants at 15 months follow-up, the prevalence π is 0.78%.

| | **Seven-point Malignancy Scale** | |
|---|---|---|
| | $\widehat{AUC}$ (s.e.) | $\widehat{AP}$ (s.e.) |
| **Film mammography** | 0.735 (0.012) | 0.166 (0.022) |
| **Digital mammography** | 0.753 (0.012) | 0.144 (0.021) |

Remark: Resampling method can be used for the inference of the difference in AP if we have paired data.

*Yuan et al. (2015)*



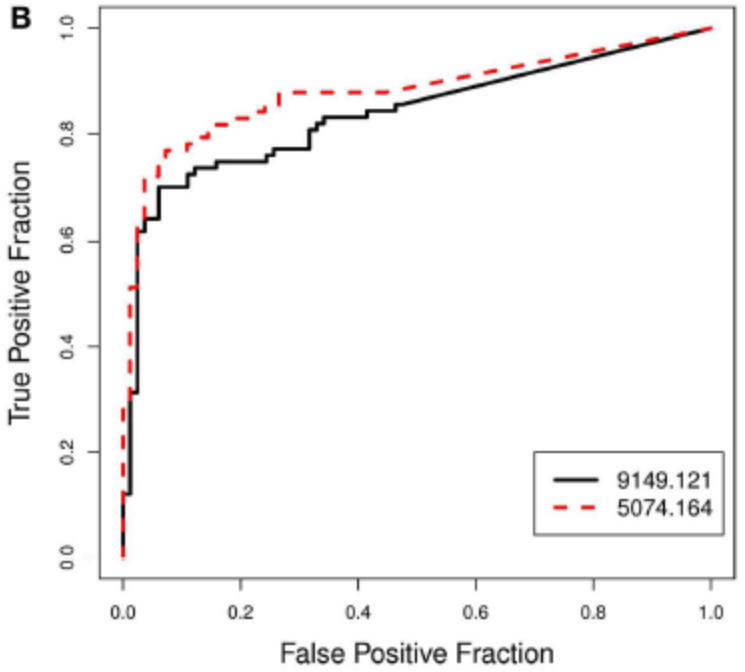




**Table 1 | Prostate cancer example.**

| Biomarkers | | AUC | | | AP | | |
|---|---|---|---|---|---|---|---|
| | | $n_0 \times 1$ ($\pi \approx 0.5$) | $n_0 \times 10$ ($\pi \approx 0.09$) | $n_0 \times 100$ ($\pi \approx 0.01$) | $n_0 \times 1$ ($\pi \approx 0.5$) | $n_0 \times 10$ ($\pi \approx 0.09$) | $n_0 \times 100$ ($\pi \approx 0.01$) |
| A | 8355.562 | 0.849 | 0.783 | 0.783 | 0.856 | 0.606 | 0.571 |
| | 7819.751 | 0.850 | 0.857 | 0.857 | 0.802 | 0.370 | 0.062 |
| B | 5074.164 | 0.886 | 0.869 | 0.869 | 0.833 | 0.306 | 0.043 |
| | 9149.121 | 0.832 | 0.793 | 0.793 | 0.822 | 0.512 | 0.225 |

*A simple thought experiment showing changes in the estimated AUC and AP as a result of artificially inflating the number of control subjects ($n_0$) to mimic real-life screening settings, where the prevalence ($\pi$) of disease is low.*

| Score | $x_1$ | $>$ | $x_2$ | $> \cdots >$ | $x_k$ | $>$ | $x_{k+1}$ | $> \cdots >$ | $x_K$ | |
|---|---|---|---|---|---|---|---|---|---|---|
| Partition | $R_1$ | | $R_2$ | $\cdots$ | $R_k$ | | $R_{k+1}$ | $\cdots$ | $R_K$ | Total |
| Class-1 | $Z_1$ | | $Z_2$ | $\cdots$ | $Z_k$ | | $Z_{k+1}$ | $\cdots$ | $Z_K$ | $n_1$ |
| Class-0 | $\bar{Z}_1$ | | $\bar{Z}_2$ | $\cdots$ | $\bar{Z}_k$ | | $\bar{Z}_{k+1}$ | $\cdots$ | $\bar{Z}_K$ | $n_0$ |
| Total | $S_1$ | | $S_2$ | $\cdots$ | $S_k$ | | $S_{k+1}$ | $\cdots$ | $S_K$ | $n$ |

$$\widehat{AP} = \underbrace{\left[\frac{Z_1}{S_1}\right]}_{w_1}\left[\frac{Z_1}{n_1}\right] + \underbrace{\left[\frac{Z_1+Z_2}{S_1+S_2}\right]}_{w_2}\left[\frac{Z_2}{n_1}\right] + \cdots + \underbrace{\left[\frac{Z_1+Z_2+\cdots+Z_K}{S_1+S_2+\cdots+S_K}\right]}_{w_K}\left[\frac{Z_K}{n_1}\right]$$

$$= \sum_{k=1} W_k \left[\frac{Z_k}{n_1}\right].$$

$$\widehat{AUC} = \frac{n}{n_0}\left\{\underbrace{\left[\frac{S_1+S_2+\ldots+S_K}{n}\right]}_{w_1'}\left[\frac{Z_1}{n_1}\right] + \underbrace{\left[\frac{S_2+\ldots+S_K}{n}\right]}_{w_2'}\left[\frac{Z_2}{n_1}\right] + \ldots + \underbrace{\left[\frac{S_K}{n}\right]}_{w_K'}\left[\frac{Z_K}{n_1}\right] - \frac{1}{2}\left(\frac{n_1}{n_0}\right)\right\} - \frac{1}{2}\left(\frac{n_1}{n_0}\right)$$

$$= \frac{n}{n_0}\sum_{k=1} w_k' \left[\frac{Z_k}{n_1}\right] - \frac{1}{2}\left(\frac{n_1}{n_0}\right)$$

# AP – AUC Relationship

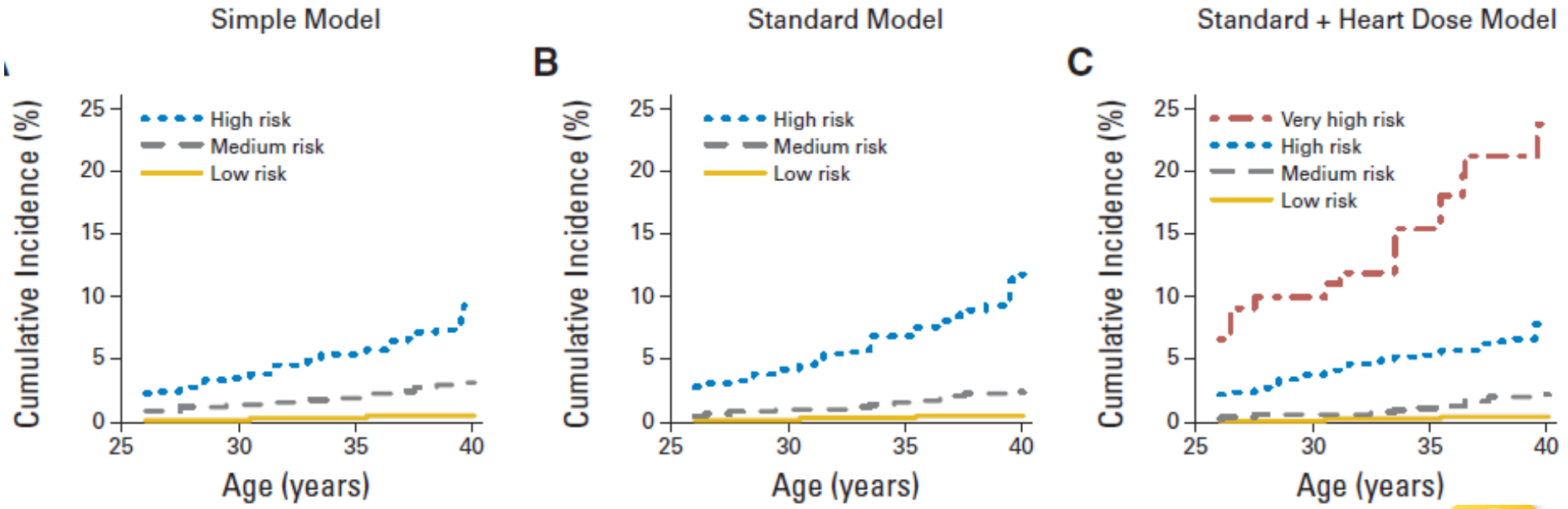
- When two risk scores $U_1$ and $U_2$ are compared
  - If ROC curve of $U_1$ dominates that of $U_2$ everywhere, then PR curve of $U_1$ dominates that of $U_2$ everywhere. $AUC_1 > AUC_2$ and $AP_1 > AP_2$
  - If ROC curves of $U_1$ and $U_2$ crosses, the ranking of $U_1$ and $U_2$ based on of AUC and AP may differ.
- Both AUC and AP are semi-proper scoring rule.



*Su et al. 2015*
*Yuan et al. 2018*

# Motivating Data – Time to Event outcome

- Late effects of cancer treatments in childhood cancer survivors – e.g. Congestive heart failure (Chow et al. 2015, *Journal of Clinical Oncology*)
- Cumulative risk of CHF is ~3% by 35 years post diagnosis






*Yuan et al. 2018*

# $AP_{t_0}$ for Time-to-Event Outcome

- Time-dependent Average Positive predictive value ($AP_{t_0}$)

$$\mathrm{AP}_{t_0} = \int_{\mathcal{R}} \mathrm{PPV}_{t_0}(z) d\mathrm{TPF}_{t_0}(z).$$

*Yuan et al. 2018*

# Nonparametric Estimator for Event Status

Let $(X, \delta, Z)$ be the standard time to event data notation,
*X*: the censored event time, $\delta$: the censoring indicator
*Z*: the risk score

$$\widehat{\mathrm{AP}}_{t_0} = \frac{\sum_{j=1}^{n} I(X_j \le t_0)\hat{w}_{t_0,j}\widehat{\mathrm{PPV}}_{t_0}(Z_j)}{\sum_{j=1}^{n} I(X_j \le t_0)\hat{w}_{t_0,j}}.$$

where

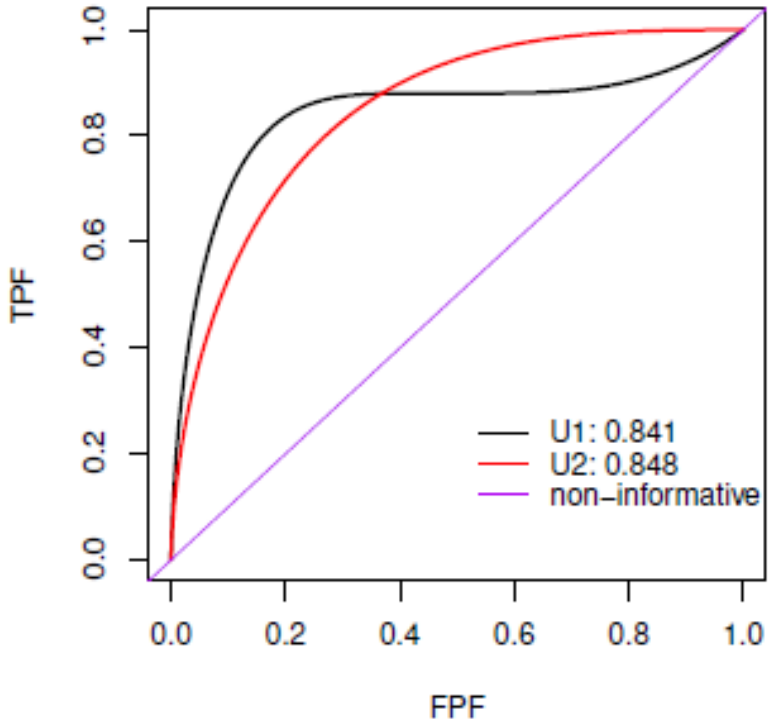
$$\hat{w}_{t_0,i} = \frac{I(X_i < t_0)\delta_i}{\widehat{\mathcal{G}}(X_i)} + \frac{I(X_i \ge t_0)}{\widehat{\mathcal{G}}(t_0)}$$

$$\widehat{\mathrm{PPV}}_{t_0}(z) = \frac{\sum_{i=1}^{n} \hat{w}_{t_0,i} I(Z_i \ge z) I(X_i < t_0)}{\sum_{i=1}^{n} \hat{w}_{t_0,i} I(Z_i \ge z)}$$

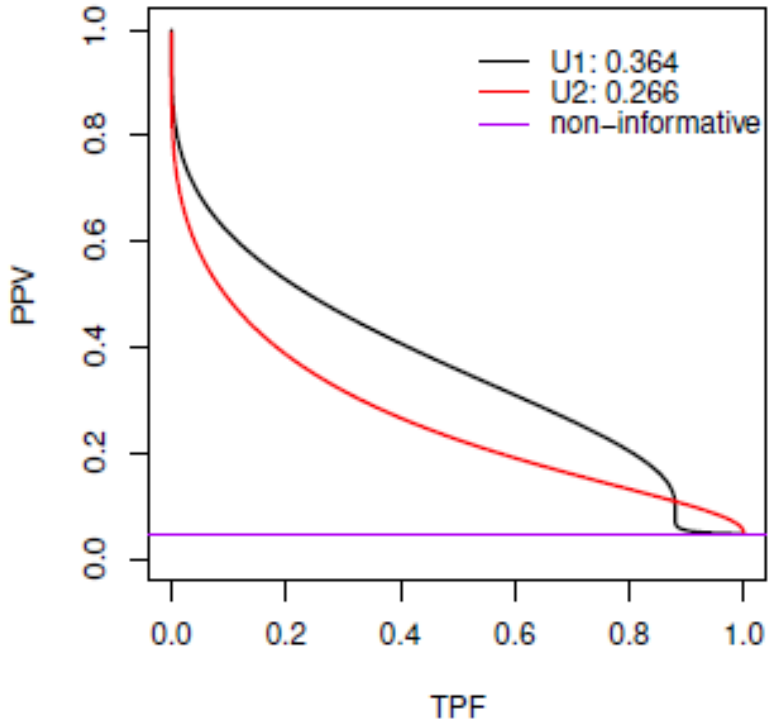
# Simulation Study

$$\log(T_i) = 7.2 - 1.1U_{i1} - 2.5U_{i2} - 1.5log(U_{i1}^2) + \epsilon_T,$$

$ROC_{t_0=8}$




$PR_{t_0=8}$



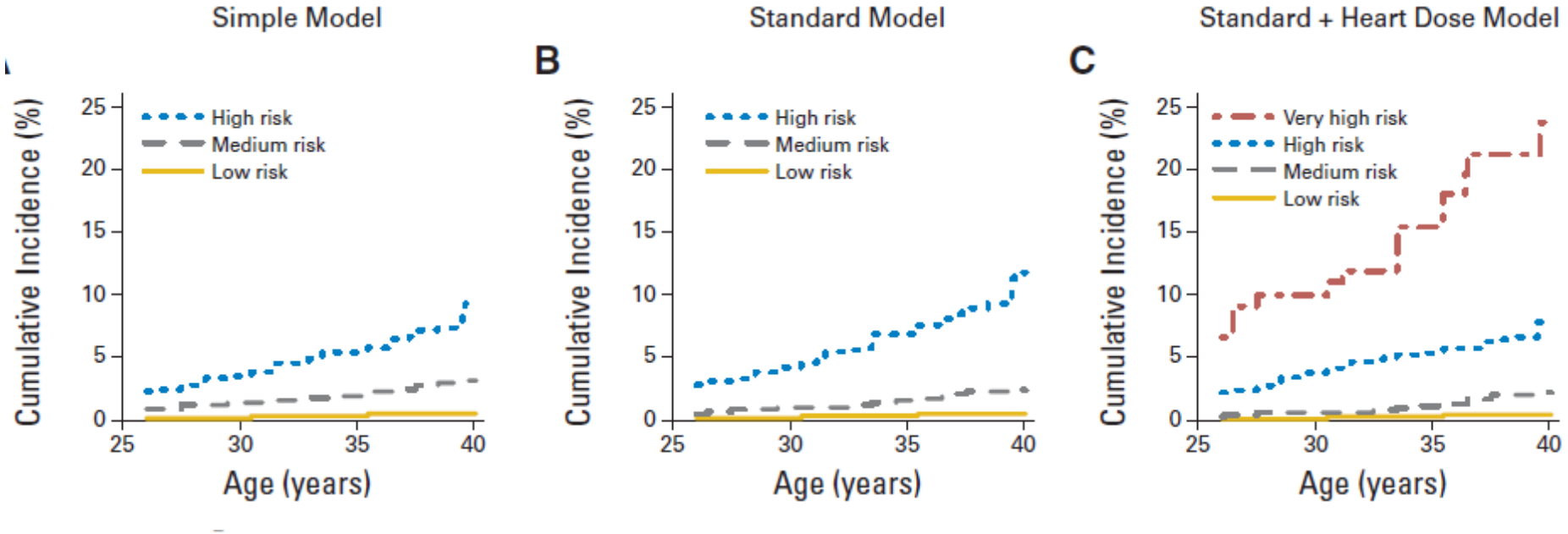
# Results (n=2000)

| $t_0$ | Event rate | Risk score | AP | | | | | AUC |
|---|---|---|---|---|---|---|---|---|
| | | | TRUE | BIAS | ESE | $ASE^b$ | $ECOVP^b(\%)$ | TRUE |
| 0.5 | 0.0101 | $U_1$ | 0.182 | 0.0361 | 0.0806 | 0.0794 | 92.2 | 0.920 |
| | | $U_2$ | 0.124 | 0.0339 | 0.0687 | 0.0679 | 94.1 | 0.904 |
| | | $\Delta$ | 0.058 | 0.0251 | 0.102 | 0.116 | 96.1 | 0.016 |
| | | Ratio | 1.47 | 0.4820 | 1.470 | 1.740 | 92.4 | 1.02 |
| 8 | 0.0495 | $U_1$ | 0.364 | 0.0085 | 0.0508 | 0.0499 | 94.4 | 0.841 |
| | | $U_2$ | 0.266 | 0.0121 | 0.0435 | 0.0439 | 94.8 | 0.848 |
| | | $\Delta$ | 0.098 | -0.0028 | 0.0707 | 0.072 | 96.3 | -0.007 |
| | | Ratio | 1.37 | 0.0123 | 0.310 | 0.322 | 95.8 | 0.99 |
| 36 | 0.0991 | $U_1$ | 0.462 | 0.0060 | 0.0416 | 0.0431 | 94.2 | 0.786 |
| | | $U_2$ | 0.375 | 0.0074 | 0.0387 | 0.0393 | 96.3 | 0.824 |
| | | $\Delta$ | 0.087 | -0.0045 | 0.0655 | 0.0633 | 95.7 | -0.038 |
| | | Ratio | 1.23 | -0.0010 | 0.189 | 0.187 | 94.5 | 0.95 |

# Results (n=5000)

| $t_0$ | Event rate | Risk score | AP | | | | | AUC |
|---|---|---|---|---|---|---|---|---|
| | | | TRUE | BIAS | ESE | $ASE^b$ | $ECOVP^b(\%)$ | TRUE |
| 0.5 | 0.0101 | $U_1$ | 0.182 | 0.0185 | 0.0498 | 0.0503 | 93.6 | 0.920 |
| | | $U_2$ | 0.124 | 0.0154 | 0.0415 | 0.0415 | 93.6 | 0.904 |
| | | $\Delta$ | 0.058 | 0.0056 | 0.0696 | 0.0712 | 94.2 | 0.016 |
| | | Ratio | 1.47 | 0.1490 | 0.709 | 0.756 | 92.9 | 1.02 |
| 8 | 0.0495 | $U_1$ | 0.364 | 0.0041 | 0.0327 | 0.0324 | 94.0 | 0.841 |
| | | $U_2$ | 0.266 | 0.0043 | 0.0285 | 0.0280 | 95.5 | 0.848 |
| | | $\Delta$ | 0.098 | -0.0005 | 0.0473 | 0.0460 | 96.3 | -0.007 |
| | | Ratio | 1.37 | 0.0099 | 0.209 | 0.204 | 94.5 | 0.99 |
| 36 | 0.0991 | $U_1$ | 0.462 | 0.0023 | 0.0273 | 0.0275 | 95.0 | 0.786 |
| | | $U_2$ | 0.375 | 0.0015 | 0.0247 | 0.0251 | 95.5 | 0.824 |
| | | $\Delta$ | 0.087 | 0.0003 | 0.0398 | 0.0402 | 95.1 | -0.038 |
| | | Ratio | 1.23 | 0.0058 | 0.117 | 0.120 | 95.0 | 0.95 |

# Time to event outcome: CCSS CHF Risk




$$\text{PPV}_{t_0}^{\text{CHF}}(z) = Pr\{T < t_0, \Delta = 1 \mid Z \geq z\} \quad \text{and} \quad \text{TPF}_{t_0}^{\text{CHF}}(z) = Pr\{Z \geq z \mid T < t_0, \Delta = 1\}.$$

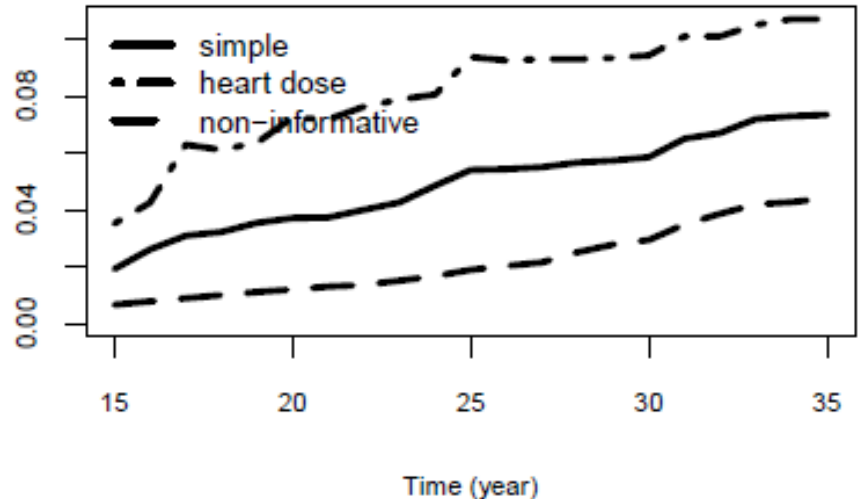
$$\widehat{\text{PPV}}_{t_0}^{\text{CHF}}(z) = \frac{\sum_{i=1}^{n} \hat{w}_{t_0,i} I(Z_i \geq z) I(X_i < t_0) I(\Delta_i = 1)}{\sum_{i=1}^{n} I(Z_i \geq z)}$$

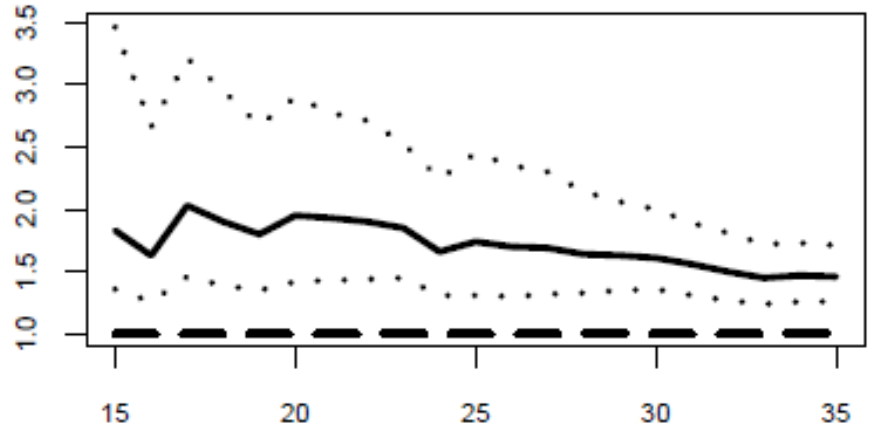
$$\widehat{\text{TPF}}_{t_0}^{\text{CHF}}(z) = \frac{\sum_{i=1}^{n} \hat{w}_{t_0,i} I(Z_i \geq z) I(X_i < t_0) I(\Delta_i = 1)}{\sum_{i=1}^{n} \hat{w}_{t_0,i} I(X_i < t_0) I(\Delta_i = 1)}$$



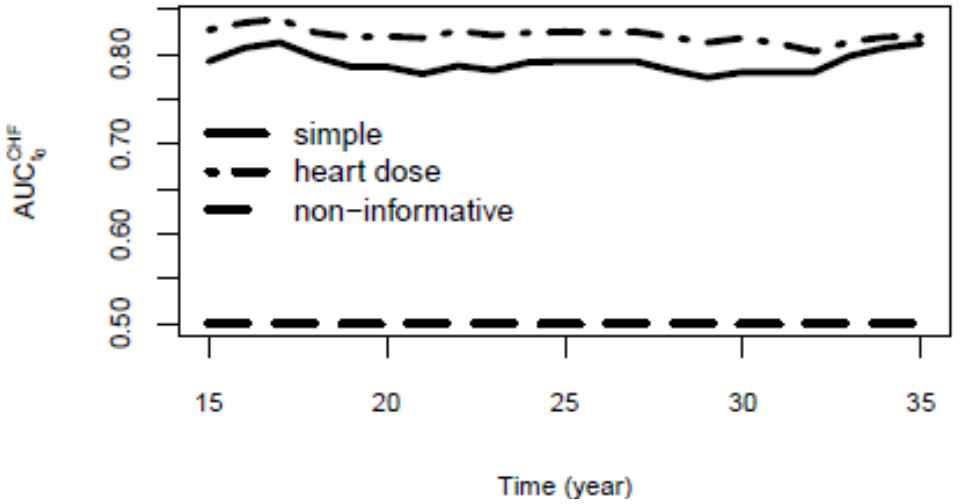
*Yuan et al. 2018*

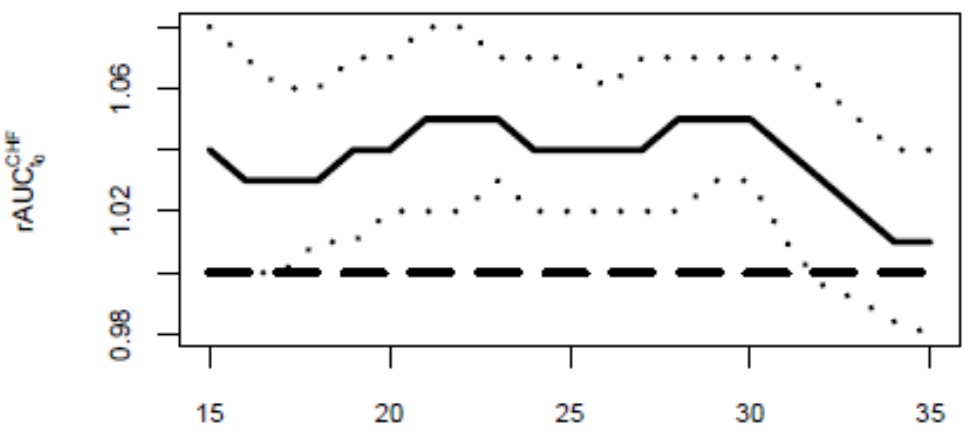
# $AP_{t_0}\ vs.\ t_0$




# $AUC_{t_0}\ vs.\ t_0$






*Yuan et al. 2018*

# Incremental Value

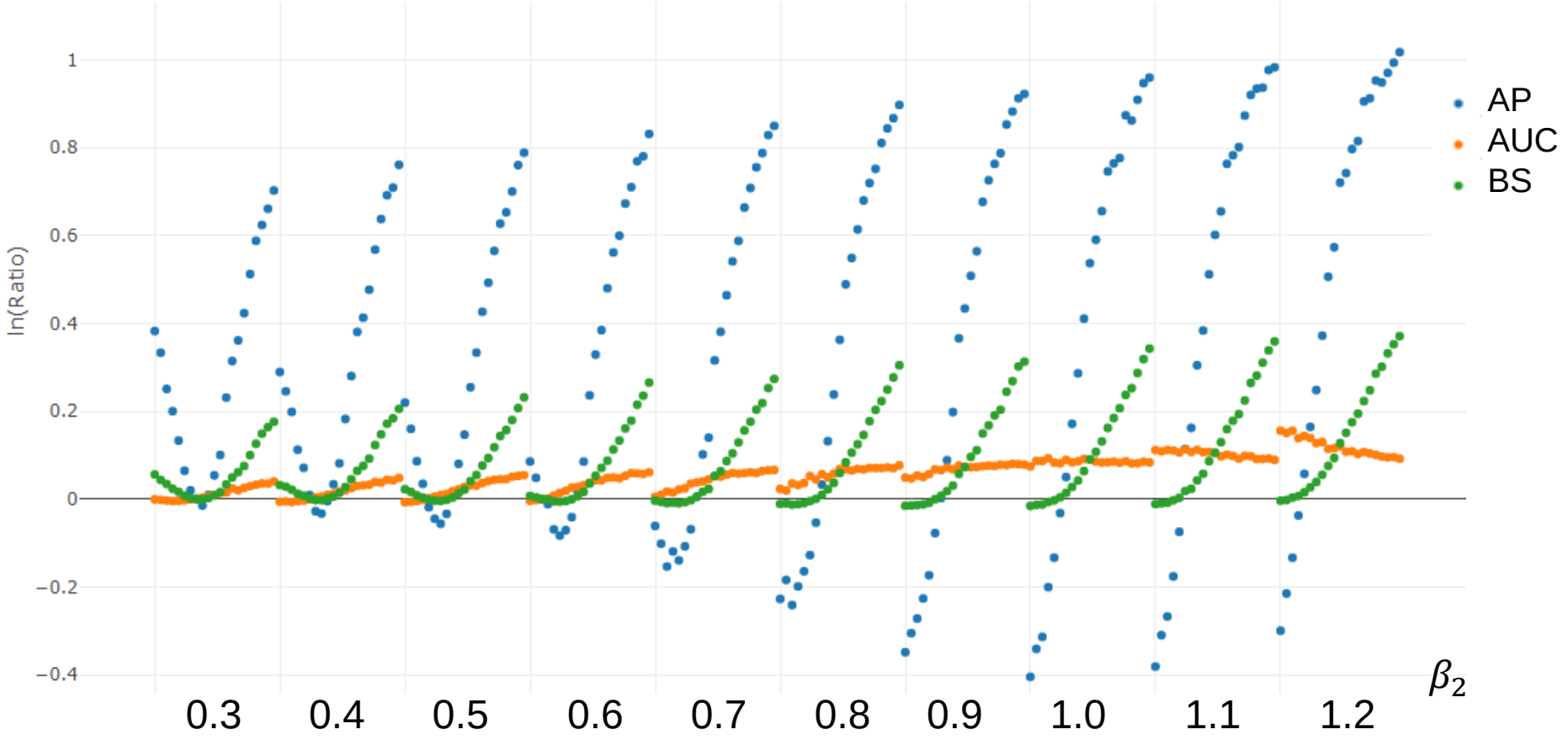
- Risk factor & outcome association vs. information/calibration gain in prediction
- Existing metrics
  - Changes in AUC and Brier scores (BS)
  - NRI (net reclassication improvement)
  - IDI (integrated discrimination improvement)

How does AP changes, in comparison to changes in AUC and BS?

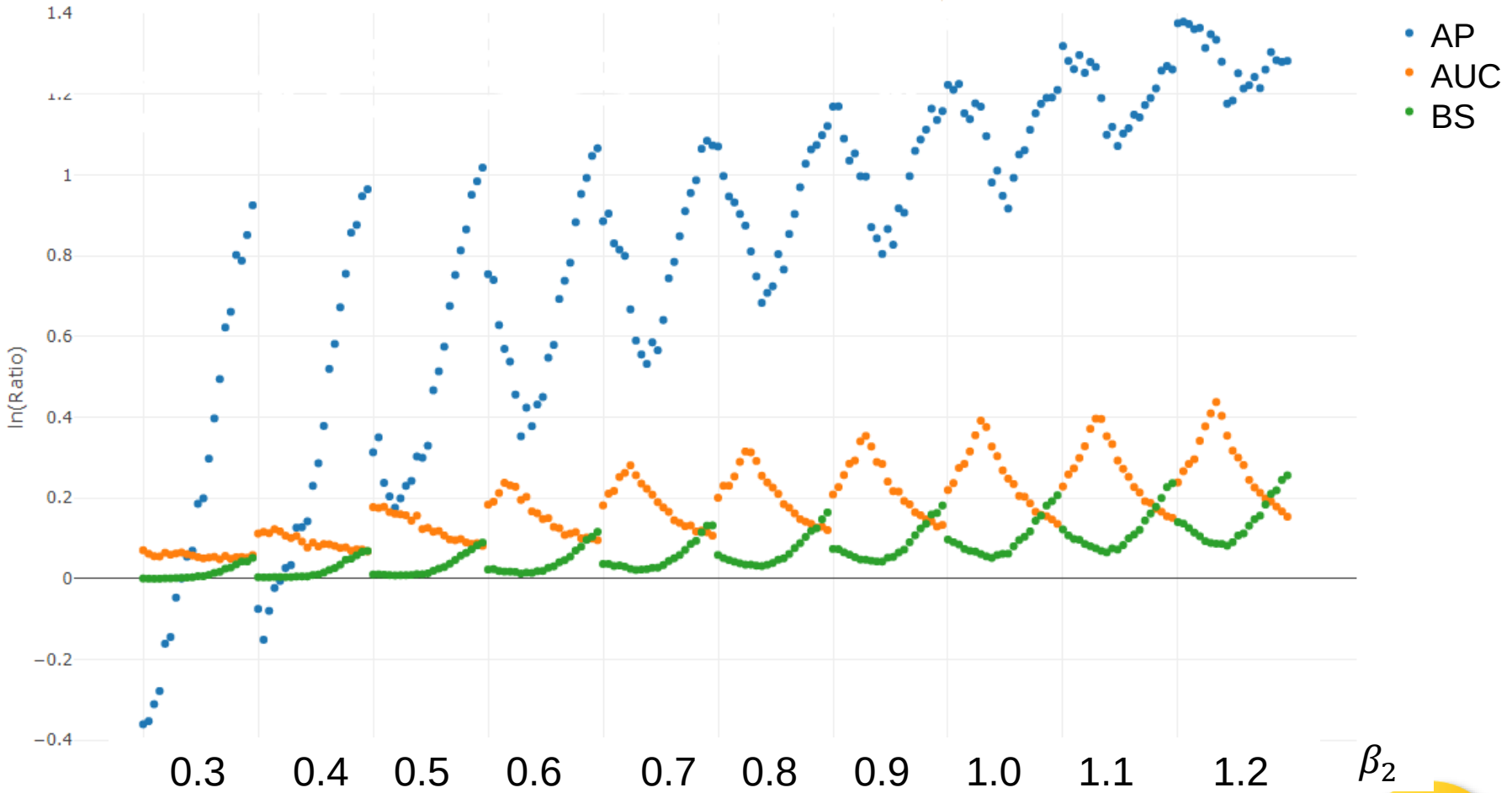
# Simulation Study

- True model: $logit(\pi) = \beta_0 + \beta_1 U_1 + \beta_2 U_2 + \beta_3 U_1 U_2,$
  - $\beta_1$ and $\beta_2$ range: [0.3, 1.2]
  - $\beta_3$ range: [-1,1]
  - Independent U1 & U2 ~ iid N(0,1)
  - Event rate: ~5%
- Working model
  - Model 1: $logit(\pi) = \beta_0 + \beta_1 U_1$
  - Model 2: $logit(\pi) = \beta_0 + \beta_1 U_1 + \beta_2 U_2$
- Metrics
  - rAUC, rAP and rBS

# $\beta_1 = 1.2$

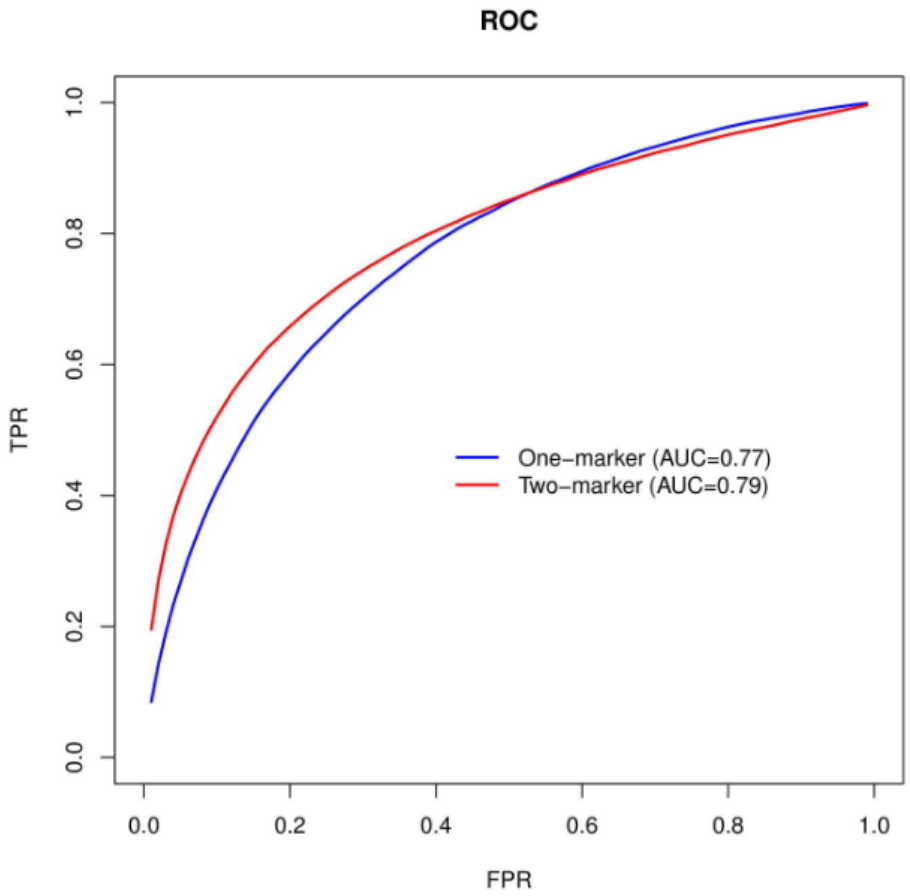

ln(Ratio)
1
0.8
0.6
0.4
0.2
0
−0.2
−0.4
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1.0
1.1
1.2
$\beta_2$
AP
AUC
BS

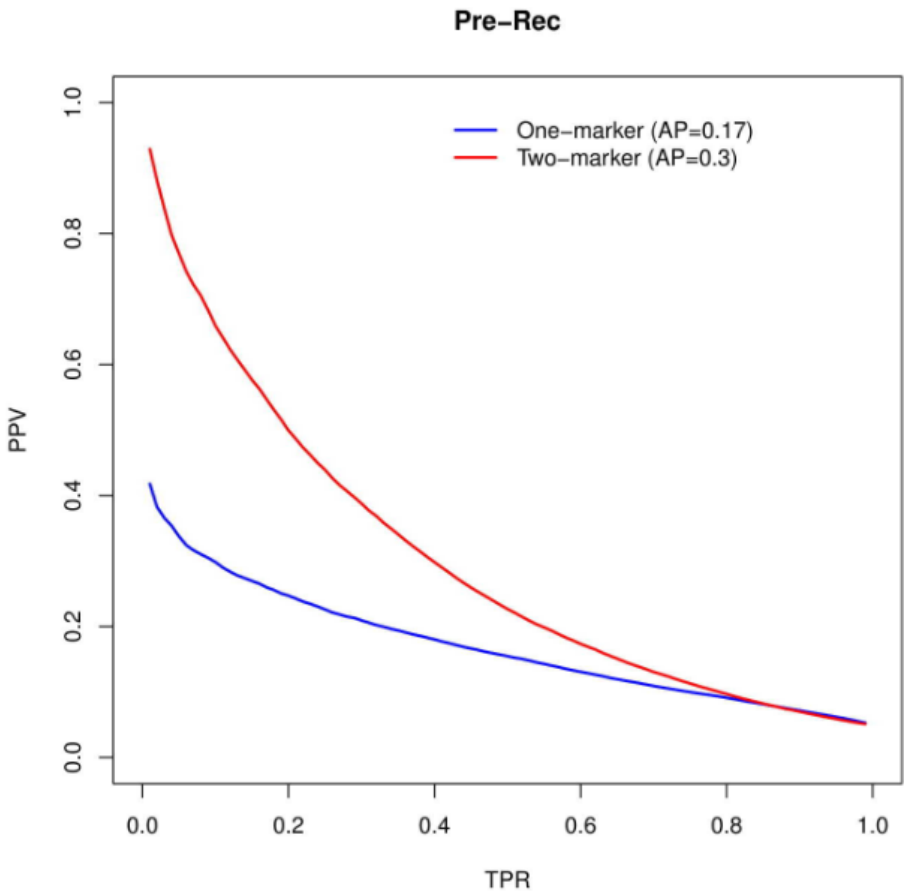
# $\beta_1 = 0.3$


1.4
1
0.8
0.6
0.4
0.2
0
−0.2
−0.4
ln(Ratio)
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1.0
1.1
1.2
$\beta_2$
AP
AUC
BS

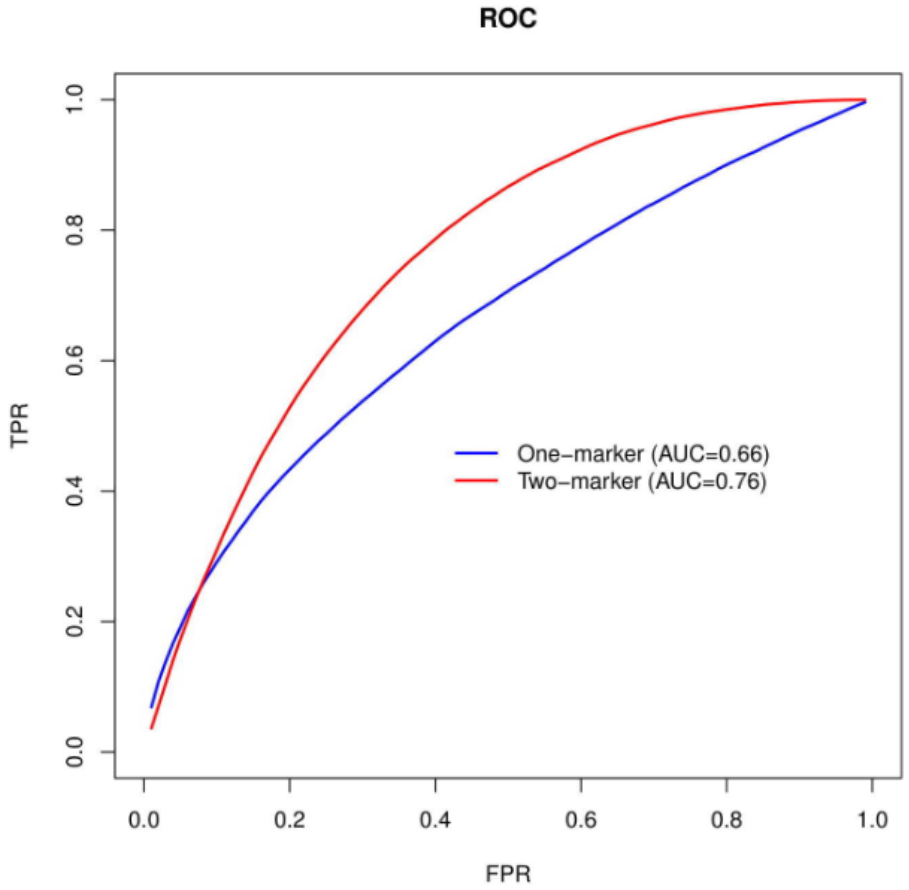
| Metrics | Correlation | |
|---|---|---|
| Log(ratio of metrics: M2/M1) | Pearson | Spearman |
| -ln(rBS) and ln(rAUC) | 0.083 | 0.30 |
| -ln(rBS) and ln(rAP) | 0.76 | 0.89 |
| ln(rAUC) and ln(rAP) | 0.48 | 0.51 |

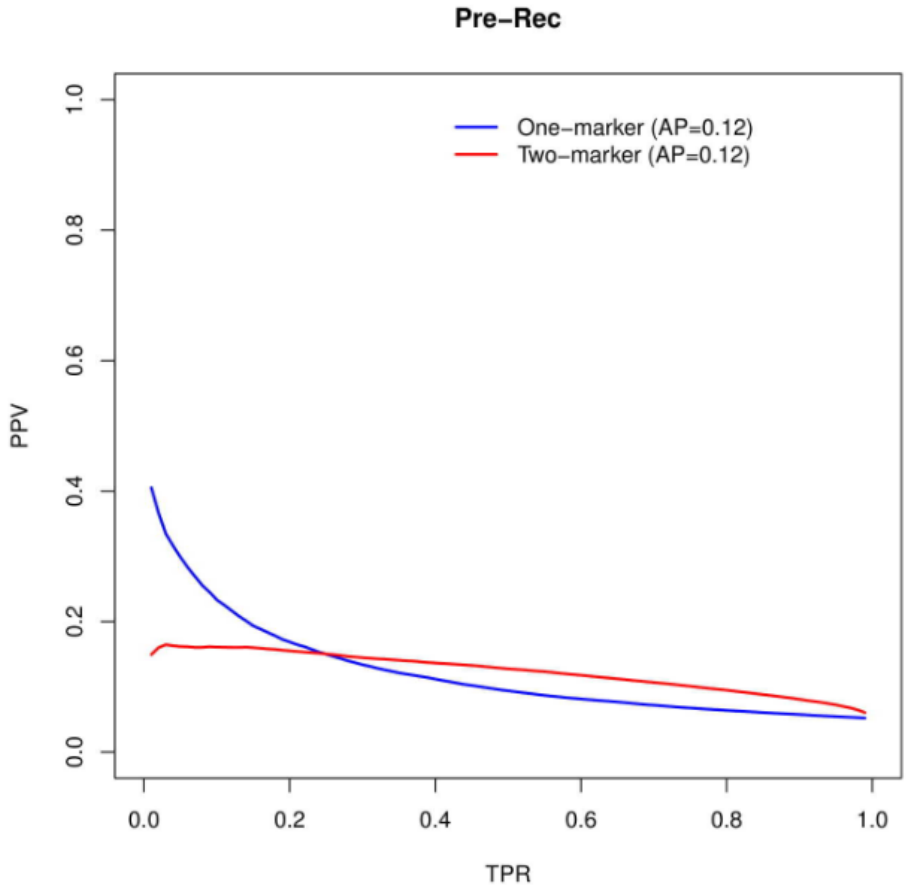
$$\beta_1 = 0.9, \beta_2 = 0.3, \beta_3 = 0.6$$



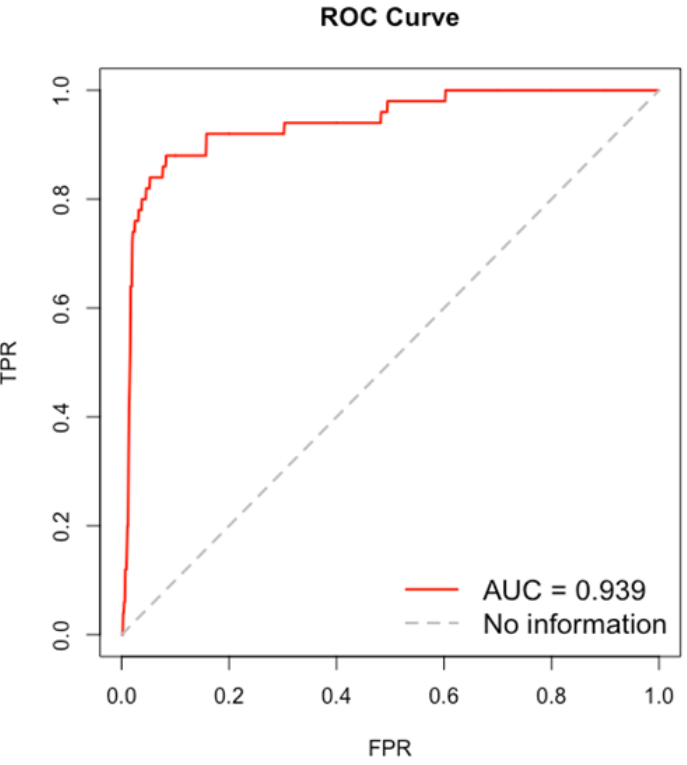
$$\beta_1 = 1, \beta_2 = 1, \beta_3 = -0.6$$

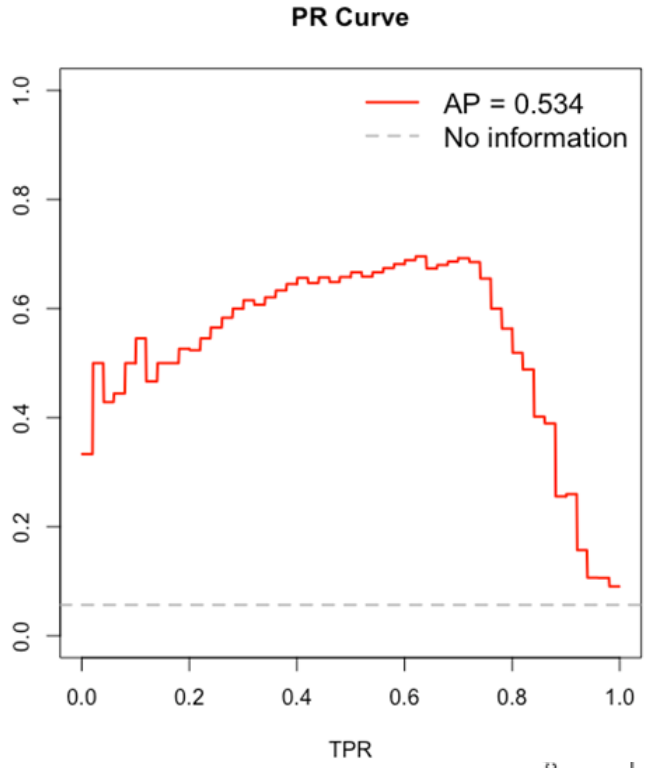


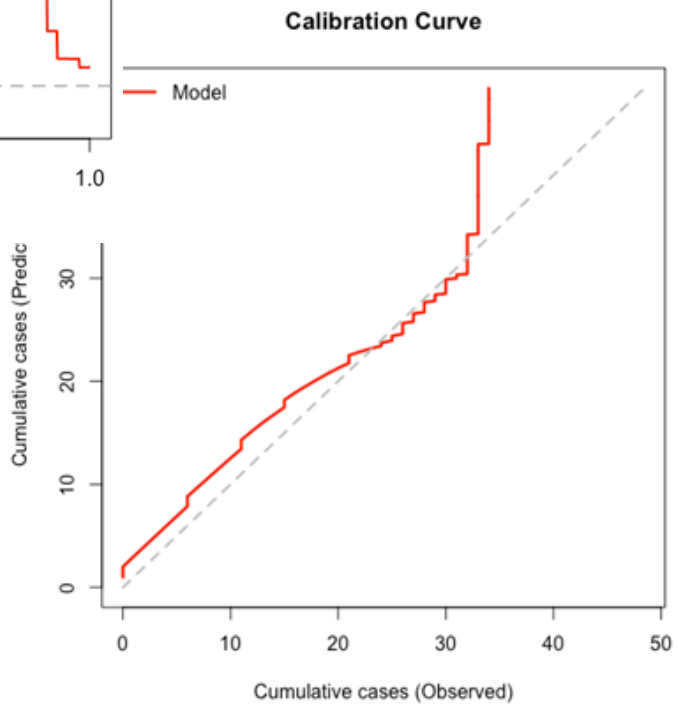



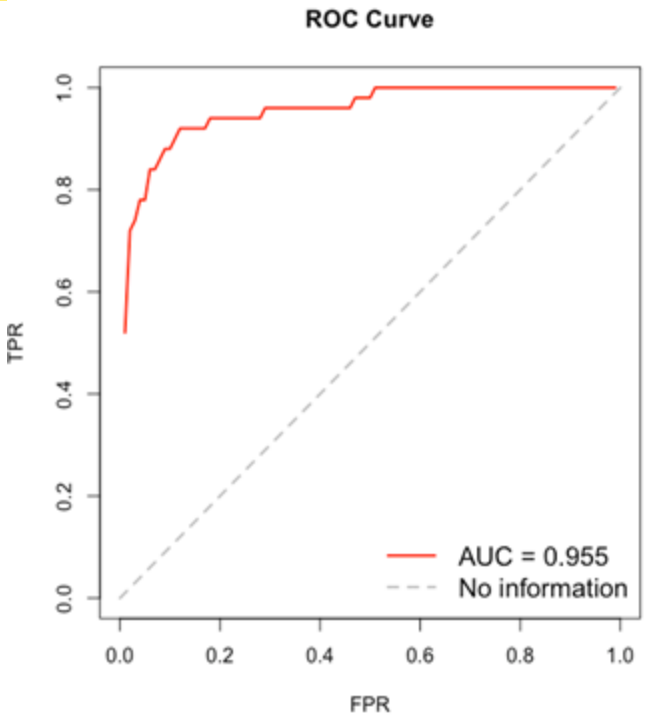
# Risk Prediction for Ovarian Failure

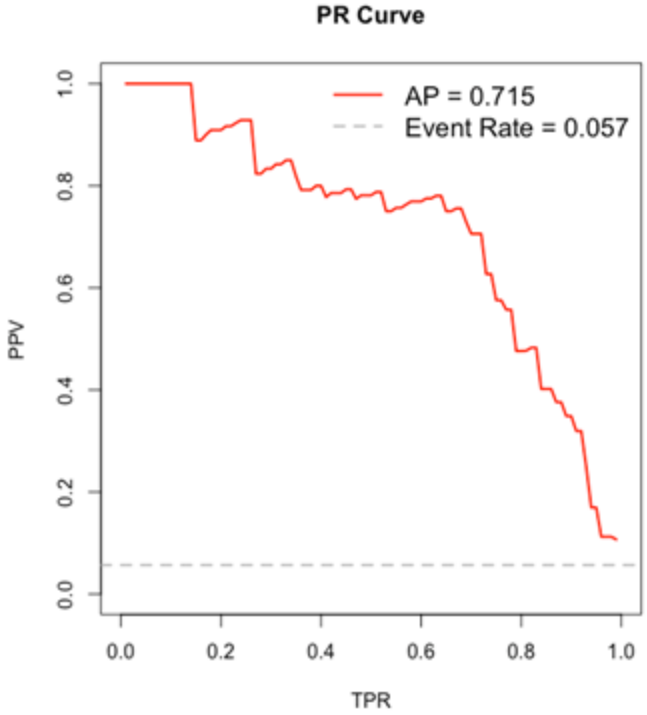
- Goal
  - Developing risk prediction model for ovarian failure (OF) in childhood cancer survivors (CCS)
- Data
  - About 6000 female CCS (dx 1970-1999)
- Methods
  - Logistic regression; Random Forest; and Support Vector Machines
- Results
  - AUC 0.82 and AP 0.50 for Acute OF (Internal validation)

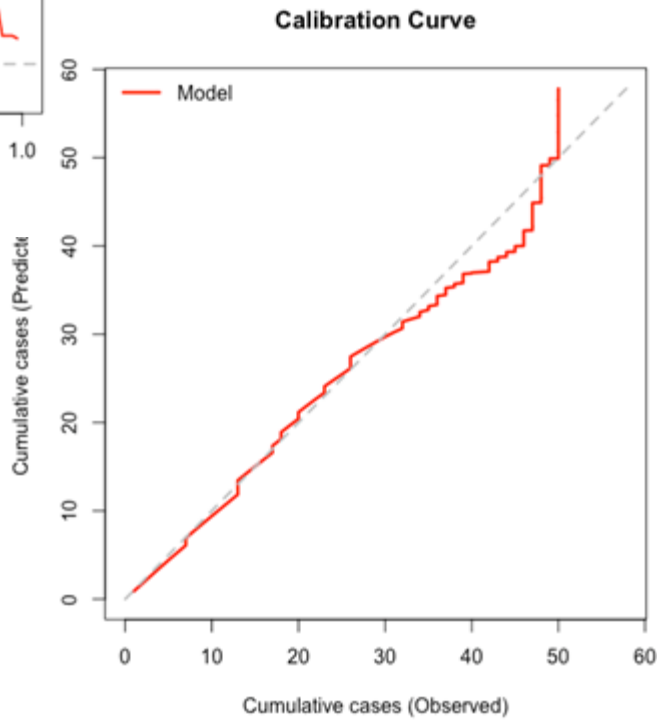

ROC Curve
TPR
FPR
AUC = 0.939
No information
0.0
0.2
0.4
0.6
0.8
1.0


PR Curve
AP = 0.534
No information
TPR
0.0
0.2
0.4
0.6
0.8
1.0


Calibration Curve
Model
Cumulative cases (Predic
Cumulative cases (Observed)
0
10
20
30
40
50


ROC Curve
TPR
FPR
AUC = 0.955
No information


PR Curve
PPV
TPR
AP = 0.715
Event Rate = 0.057

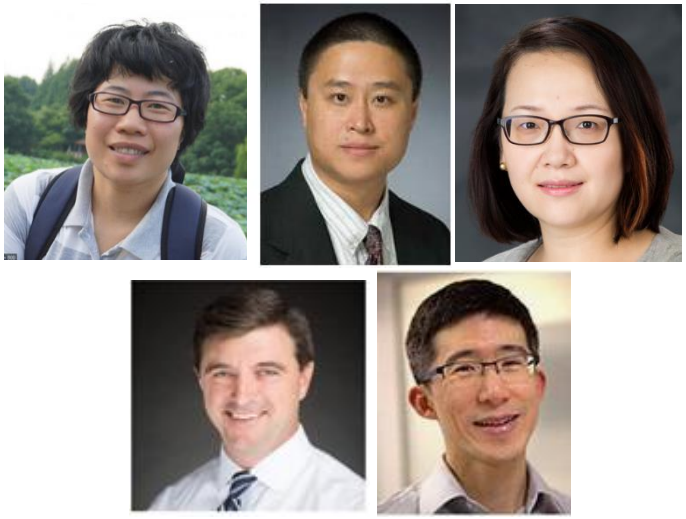

Calibration Curve
Model
Cumulative cases (Predicte
Cumulative cases (Observed)

# Discussion

- AP is a single numerical measure, in this respect it is similar to AUC.
- A summary measure of positive predictive value, useful for evaluating and comparing prospective prediction performance of risk scores.
- More sensitive than AUC.
- Better aligned with the strict proper scoring rule Brier score than AUC (under misspecificed working models)
- Event rate dependent, AP should be estimated in a prospective cohort or population-based study
- R package <APtools> and SAS macro for binary and survival time data https://sites.ualberta.ca/~yyuan/software.html

# Acknowledgement

**Students and staff**
**Maoji Li, Dr. Khanh Vu**
**Doris Li, Hengrui Cai, Zorina Han, Rebecca Clark, Michael Lu**



M.S.I. Foundation
Lois Hole Hospital for Women
women & children's health research institute
CIHR IRSC
Canadian Institutes of Health Research
Instituts de recherche en santé du Canada

# Thank you!

Questions???