UNIVERSITY OF ALBERTA
SCHOOL OF PUBLIC HEALTH

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

Yan Yuan, PhD
School of Public Health, University of Alberta
June 5, 2018 Montreal

Joint work with Michelle Zhou et al.

# Outline

- Motivation
- Measures for evaluating prediction performance of risk scores
- Estimator and simulation
- Data analysis
- Summary and future work

# Examples of Prevention and Early Detection in Clinical Practice

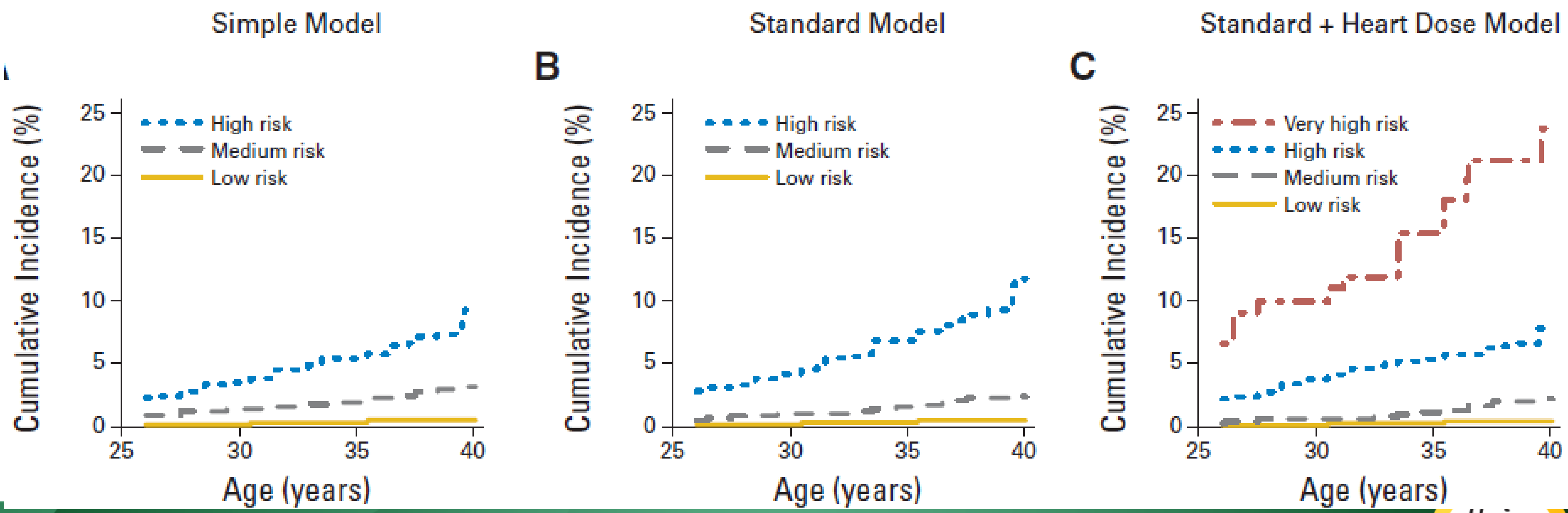
- The Prism risk tool (for re-hospitalization within a year)
- Risk charts for 182 countries to predict future risk of cardiovascular disease
- Multiple risk score systems (n>40) for diabetes risk in general population
- Risk prediction models for acute kidney injury in critically ill patients (2018)

# Risk Score as a Screening Tool

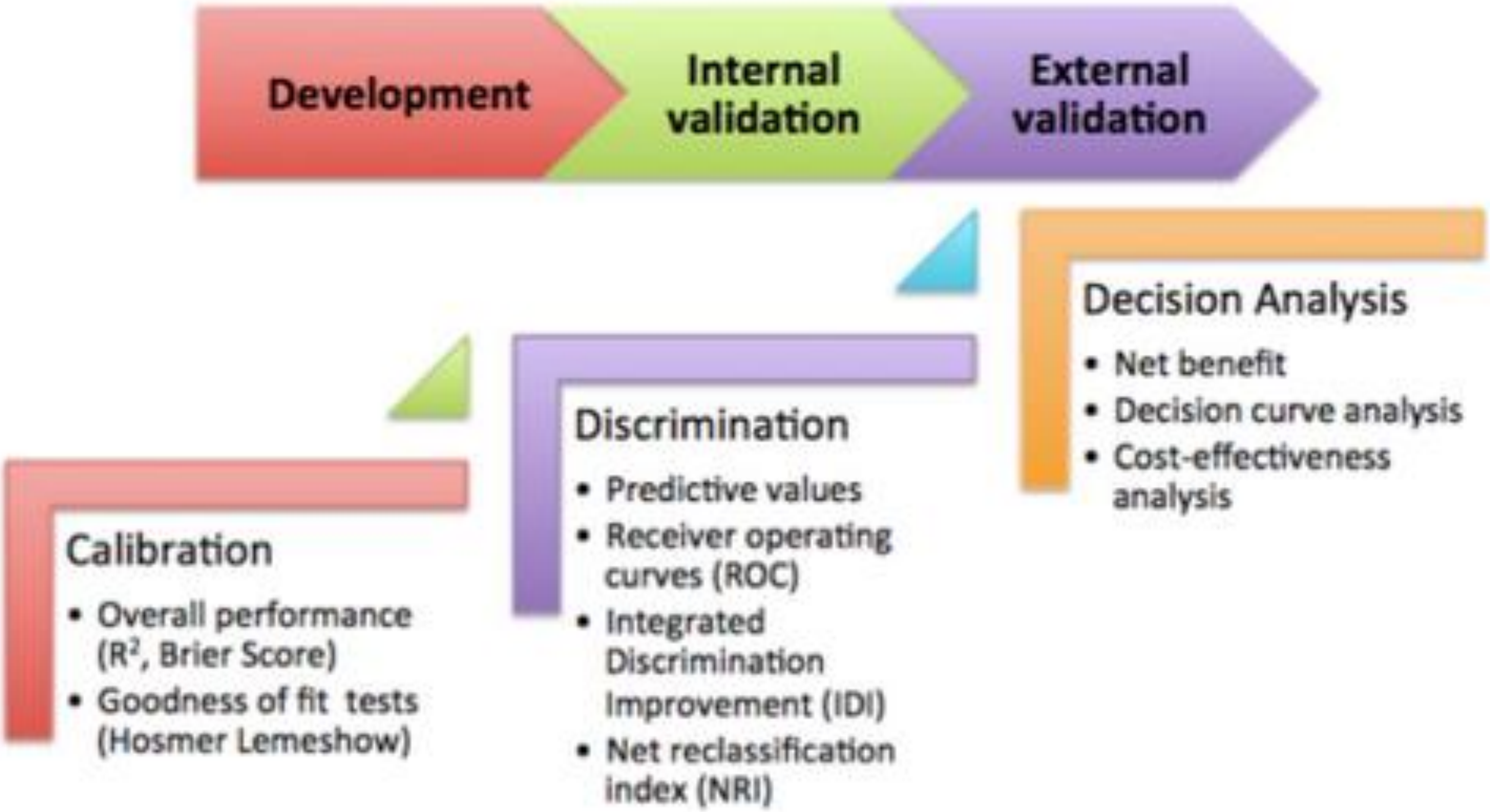
- Typical condition that risk scores are used/ developed for have the following characteristics
  - seriousness may result in a high risk of mortality or significantly affect the quality of life;
  - early detection/intervention can make a difference in disease prognosis;
  - the event rate is <u>low</u>

# Motivating Data

- 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



# Prediction Performance Measure

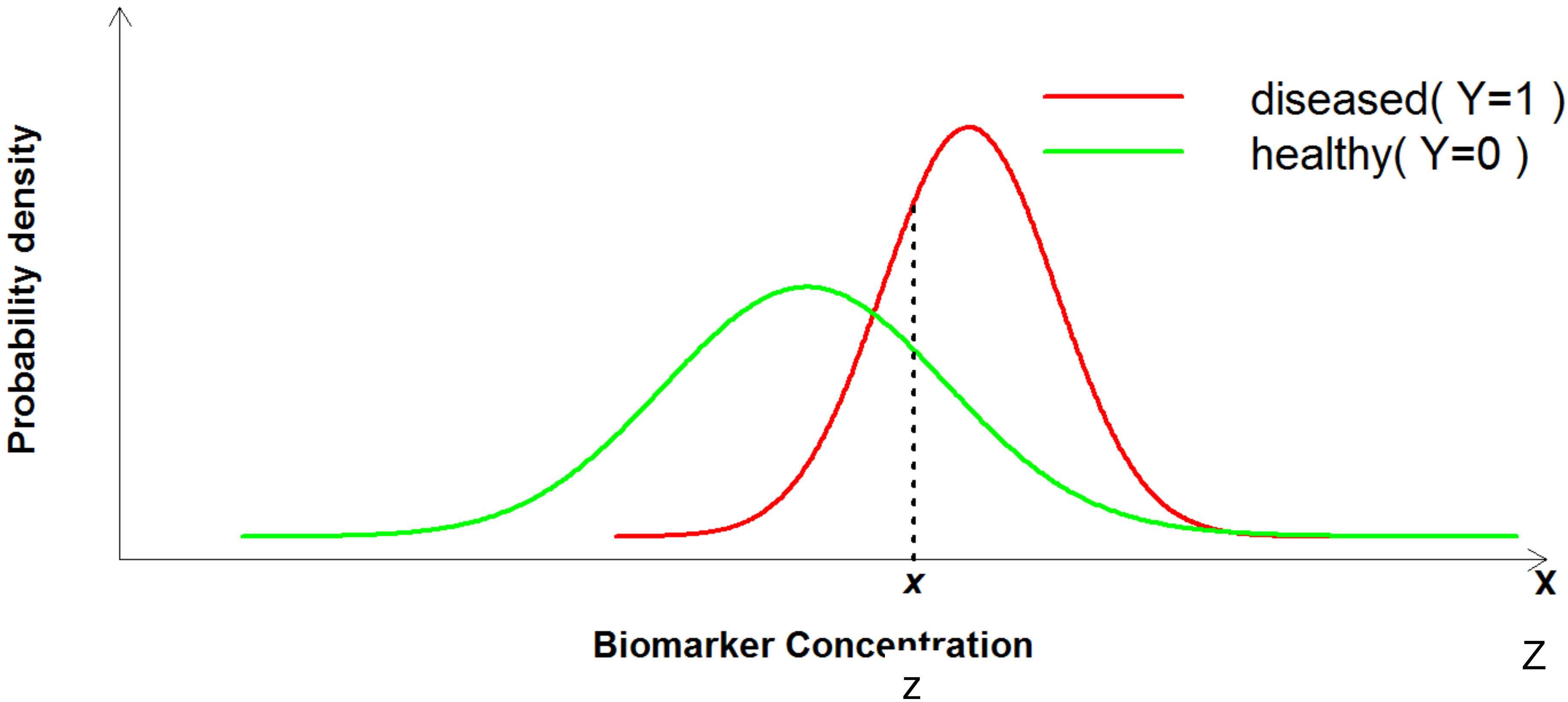



[Columbia University Mailman School of Public Health]

# Evaluating Model Performance when Predicting Low Prevalence Events

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

# When predictor is continuous or ordinal




Risk score

# Threshold-free Summary Measure

- Area Under the ROC* Curve (AUC, *aROC*)

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

- Extension to event status to accommodate censoring and time to event data -- $AUC_{t_0}$
- Criticisms of AUC as a measure for risk prediction
  - Retrospective measure
  - Insensitive
  - Over-optimistic

# A Threshold-free Alternative to AUC for Binary Outcome

- Average Positive predictive value (AP)

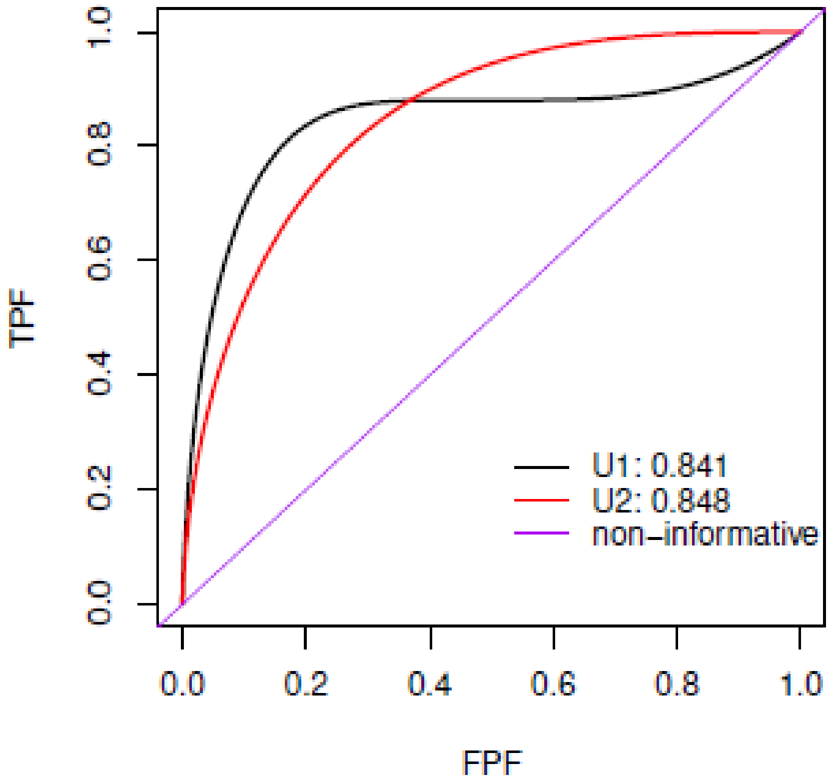
$$\mathrm{A}P \equiv \int_R \mathrm{PPV}(z) d\mathrm{TPF}(z)$$

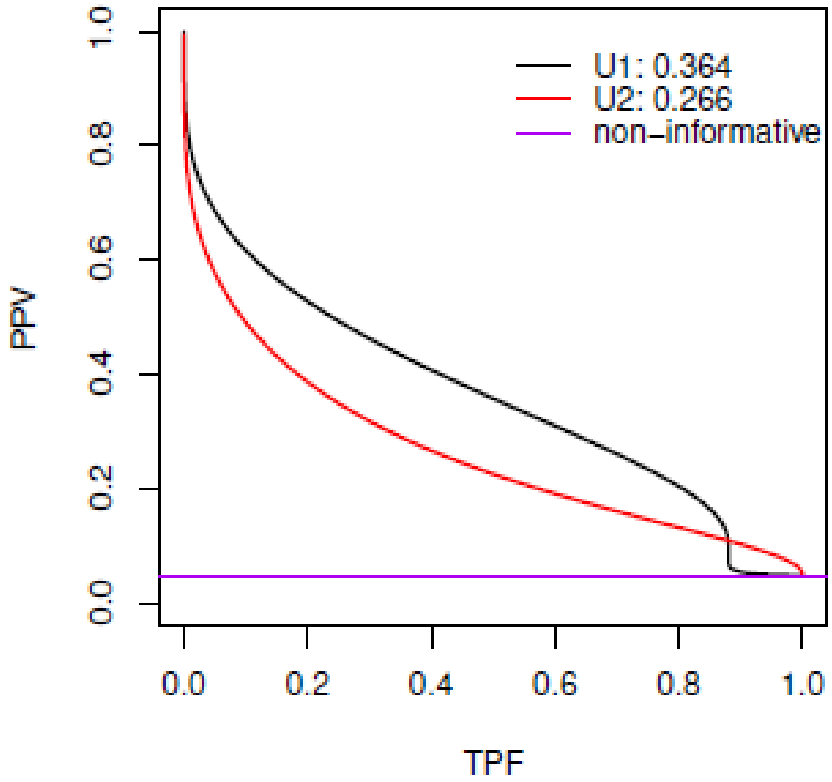
Remark:

– Range: [π, 1] where π is the prevalence rate and corresponds to a random risk score

*Yuan et al. (2015) Frontiers in Public Health 3:57.*

# ROC curve PvR curve



# Relationship to AUC

- When two risk scores $U_1$ and $U_2$ are compared
  - If ROC curve of $U_1$ dominates that of $U_2$ everywhere, the $AUC_1 > AUC_2$ 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 can differ.

*Su et al. (2015) Proceedings of the 2015 International Conference on Theory of Information Retrieval. pp.349-352.*

# An Alternative to $AUC_{t_0}$ for Time-to-event Outcome

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

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

# Nonparametric Estimator for Survival Status

Let $(X, \delta, Z)$ be the standard survival time 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

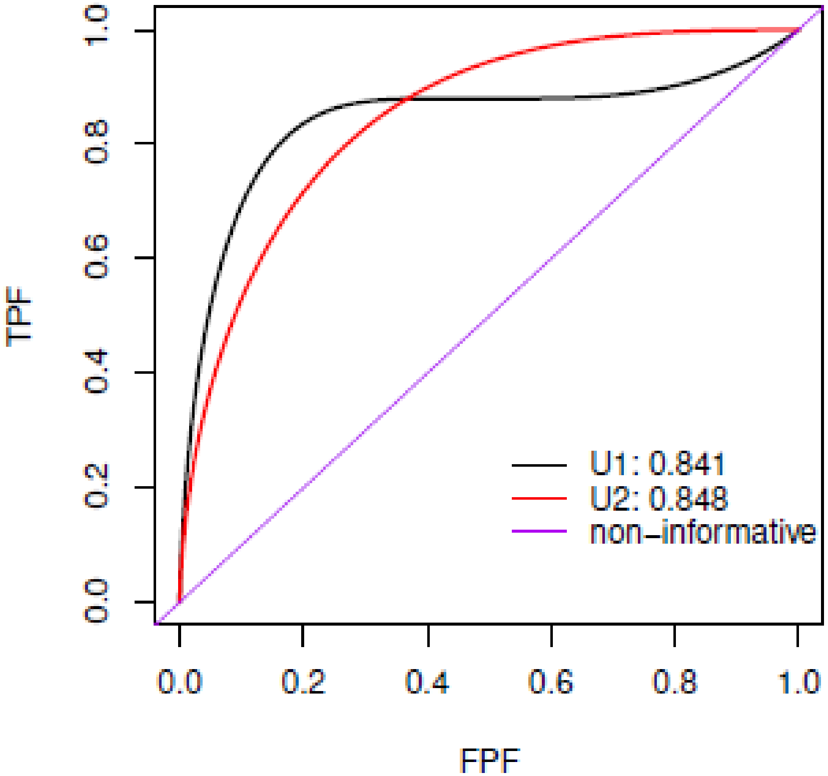
$$\widehat{w}_{t_0,i} = \frac{I(X_i < t_0)\delta_i}{\widehat{\mathcal{G}}(X_i)} + \frac{I(X_i \ge t_0)}{\widehat{\mathcal{G}}(t_0)}$$

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

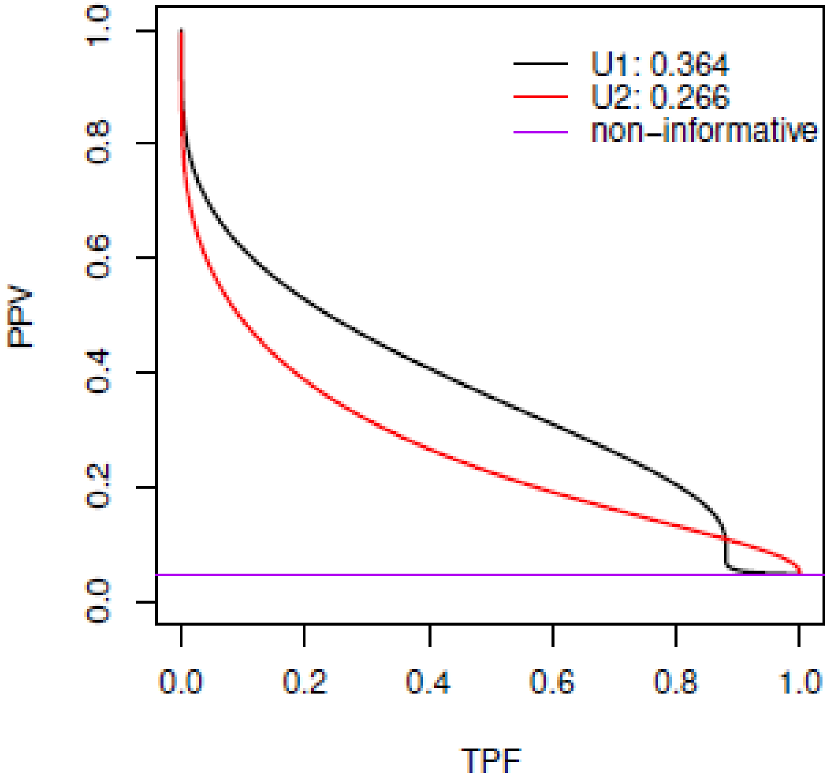
# Simulation Study

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

$ROC_{t_0=8}$




$PR_{t_0=8}$



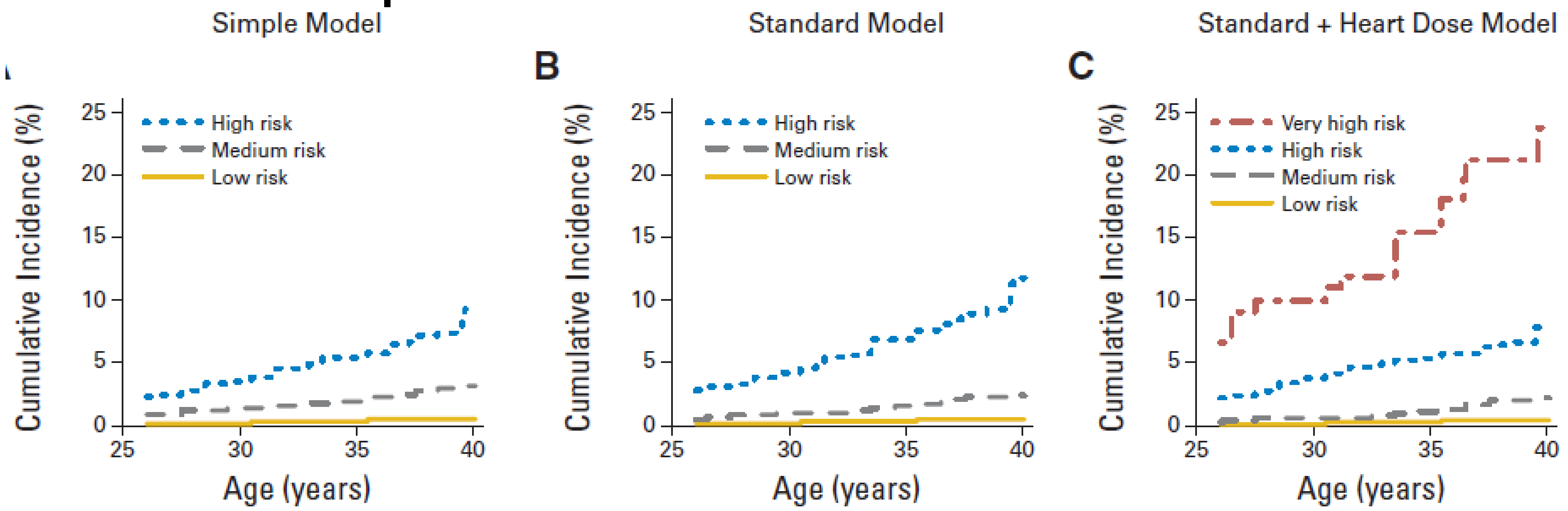
# Results (n=2000)

| $t_0$ | 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 |

# Example: CCSS CHF Risk Prediction

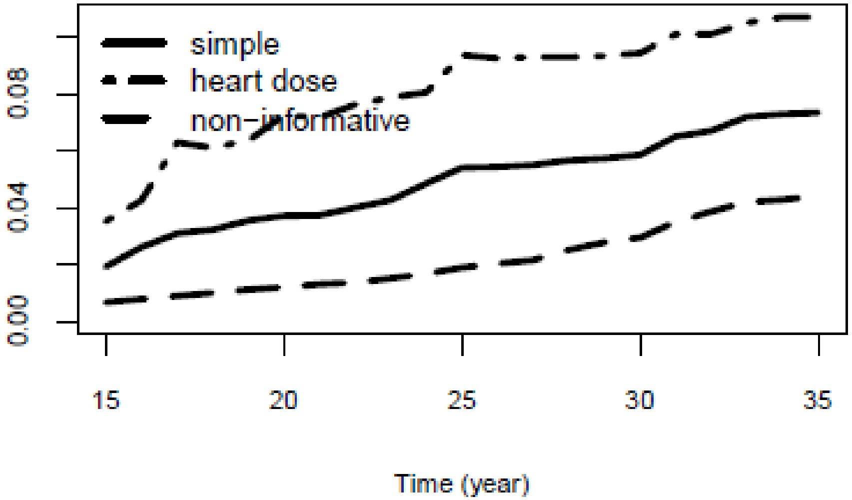


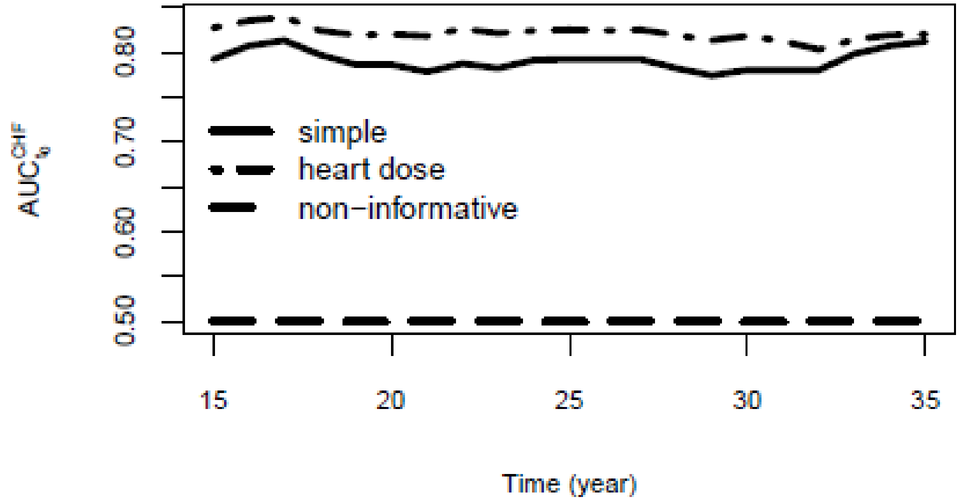

$$\mathrm{PPV}_{t_0}^{\mathrm{CHF}}(z) = Pr\{T < t_0, \Delta = 1 \mid Z \geq z\} \quad \text{and} \quad \mathrm{TPF}_{t_0}^{\mathrm{CHF}}(z) = Pr\{Z \geq z \mid T < t_0, \Delta = 1\}.$$

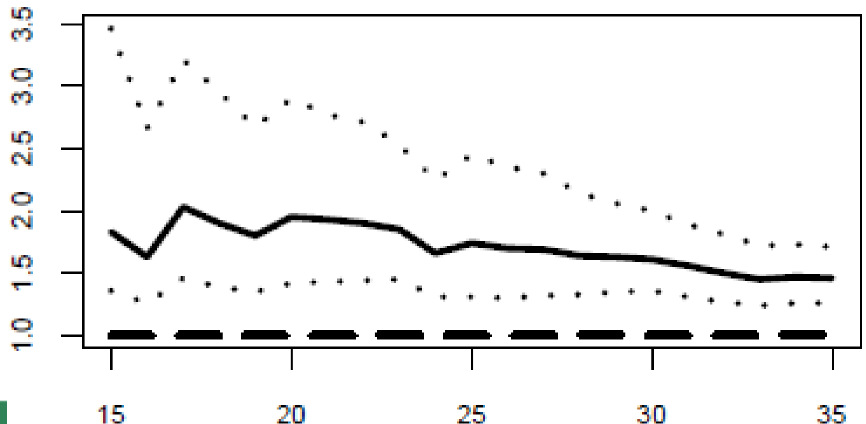
$$\widehat{\mathrm{PPV}}_{t_0}^{\mathrm{CHF}}(z) = \frac{\sum_{i=1}^{n} \widehat{w}_{t_0,i} I(Z_i \geq z) I(X_i < t_0) I(\Delta_i = 1)}{\sum_{i=1}^{n} I(Z_i \geq z)}$$

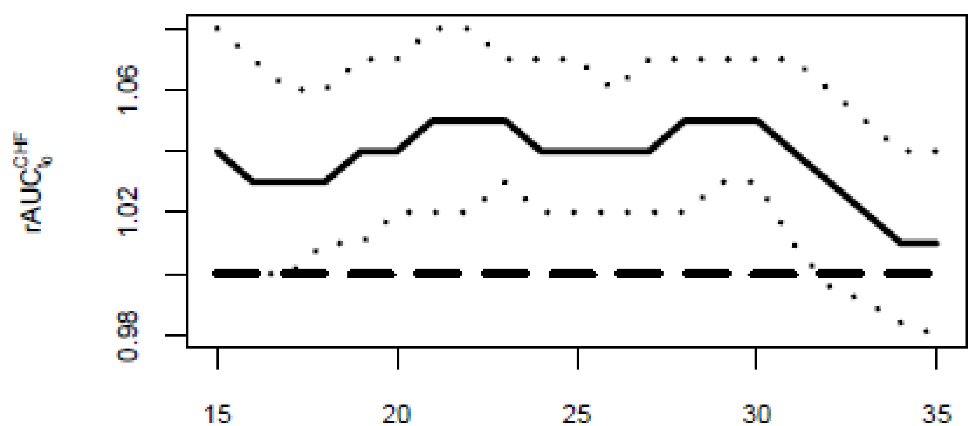
$$\widehat{\mathrm{TPF}}_{t_0}^{\mathrm{CHF}}(z) = \frac{\sum_{i=1}^{n} \widehat{w}_{t_0,i} I(Z_i \geq z) I(X_i < t_0) I(\Delta_i = 1)}{\sum_{i=1}^{n} \widehat{w}_{t_0,i} I(X_i < t_0) I(\Delta_i = 1)}$$

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



# Comparison

| $t_0$ | Event rate | Risk score system | $AP^{CHF}$ | $AUC^{CHF}$ |
|---|---|---|---|---|
| 20 years | 0.0120 | Simple | 0.037 (0.028, 0.051) | 0.786 (0.746, 0.824) |
| | | Heart dose | 0.072 (0.047, 0.120) | 0.820 (0.780, 0.859) |
| | | $\Delta$ | 0.035 (0.015, 0.077) | 0.035(0.013, 0.056) |
| | | Ratio | 1.95 (1.42, 2.90) | 1.04 (1.02, 1.07) |
| 35 years | 0.0440 | Simple | 0.073 (0.062, 0.088) | 0.812 (0.778, 0.846) |
| | | Heart dose | 0.107 (0.088, 0.135) | 0.820 (0.784, 0.856) |
| | | $\Delta$ | 0.034(0.020, 0.055) | 0.008 (-0.016, 0.029) |
| | | Ratio | 1.46 (1.26, 1.71) | 1.01 (0.98, 1.04) |

# Summary

- Point and interval estimators of AP for binary outcome (ordinal risk score);
- Nonparametric estimator of $AP_{t_0}$ for censored event status and in the presence of competing risks (continuous risk score);
- Inference procedure to compare $AP_{t_0}$ for two risk scores;
- APtools: an R package for binary and survival time data.

# Discussion

- AP is a single numerical measure, in this respect it is similar to AUC.
- A summary measure of positive predictive value, better suited in comparing prospective prediction performance of competing risk scores
- More sensitive than AUC as illustrated by the data analysis
- Event rate dependent, AP should be estimated in a prospective cohort or population-based study

# Future Work

- To evaluate how sensitive and robust the AP is as a measure of prediction accuracy Partial AP
- To extend the AP for evaluation of multicategory outcomes
- Partial AP

# Acknowledgement

Collaborators

- Dr. Qian Michelle Zhou
- Dr. Eric Chow
- Dr. Greg Armstrong

Students

- Doris Li
- Hengrui Cai