A Threshold-free Time-dependent Prospective Prediction Accuracy Measure for Censored Time to Event Data

Yan Yuan[a], Bingying Li[b], Qian Zhou[b]

a. School of Public Health, University of Alberta, Edmonton, AB, Canada T6G 1C9, b Department of Statistics and Actuarial Science, Simon Fraser University, Burnaby, B.C. Canada V5A 1S6

Introduction

Clinical decisions on disease management and disease prevention have been increasingly guided by risk scoring systems. One successful example of clinical adoption of risk prediction is the Framingham risk score (FRS). Developed in the 1990s, it has been adopted by primary care physicians for cardiovascular disease prevention in the general population[1,2]. Patients in the high risk group (FRS>20%) were recommended with statin therapy and radical behaviour modification, whereas the patients in the low risk group (FRS<10%) are primarily recommended with healthy behaviour modification and regular monitoring at 1 to 3 years interval[3].

In risk prediction, the goal is to estimate the probability of an adverse event within a specific time frame for each individual patient. This is a different concept from estimating relative risk which measures effects of risk factors in the form of relative risk, odds ratio, hazard ratio, or rate ratio, and does not indicate how likely an event of interest will develop. Before a risk scoring system being adopted into clinical practice, it is critical to evaluate its accuracy. The receiver operating characteristic (ROC) curve is the most popular accuracy measures, which provides a summary of two **retrospective** accuracy metrics, the true positive fraction (TPF) and false positive fraction (FPF). The **prospective** accuracy measures, such as positive/negative predictive values (PPV/NPV), provide more appropriate assessment of the prediction performance of the risk score[4].

The PPV is threshold dependent and different risk score systems could outperform at different cut points[5]. In this project, we developed a threshold-free summary measure based on the PPV to evaluate risk score systems for time-dependent binary outcome (event status), i.e. censored time-to-event data.

Definition

Yuan et al. (2015)[6] defined the average positive predictive values, for the binary outcome D[7].

$$\mathrm{AP} = \int_{\infty}^{\infty} \mathrm{PPV}(z) d\mathrm{TPF}(z),$$

where PPV(z) = Pr{D = 1 | $Z \geq z$}, TPF(z) = Pr{$Z \geq z$ | D = 1}, Z is a continuous risk score, and D is a disease indicator.
Let the disease status D depend on time t and T be the time to the event of interest, i.e. $D(t_0)$ = I($T < t_0$). Time-dependent PPV, TPF and AP are given by

PPV(t_0, z) = Pr{$T < t_0$ | $Z \geq z$} and TPF(t_0, z) = Pr{$Z \geq z$ | $T < t_0$}

$$\mathrm{AP}_{t_0} = \int_{-\infty}^{\infty} \mathrm{PPV}_{t_0}(z) d\mathrm{TPF}_{t_0}(z).$$

Note that TPF is the distribution function of risk score Z in "cases". It can be shown that AP(t_0) = E_{Z1}{PPV (t_0, $Z1$)}, where $Z1$ denotes the risk score Z in "cases".
A perfect risk score system would always assign higher values to "cases", individuals who experience the event before t_0, compared to those "controls", individuals who do not experience the event before t_0. This leads to AP(t_0) = 1. A useless risk score system would randomly assign risk scores to both cases and controls. This leads to AP(t_0) = π(t_0), the event rate by time t_0 in the target population.

The theoretical range of AP(t_0) is [π(t_0) , 1].

Estimator

Due to censoring, one can only observe X = min{T, C} where C is the censoring time, and δ = I($T < C$).
Let {(X_i, δ_i, Z_i), i = 1, · · · , n} be n independent realizations of (X, δ, Z). We use the inverse probability weighting to account for censoring. The nonparametric estimators of time-dependent PPV and TPF are:

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

where

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

$\widetilde{\mathcal{G}}(c)$ is a consistent estimator of the survival function of the censoring time, Pr($C \geq c$).

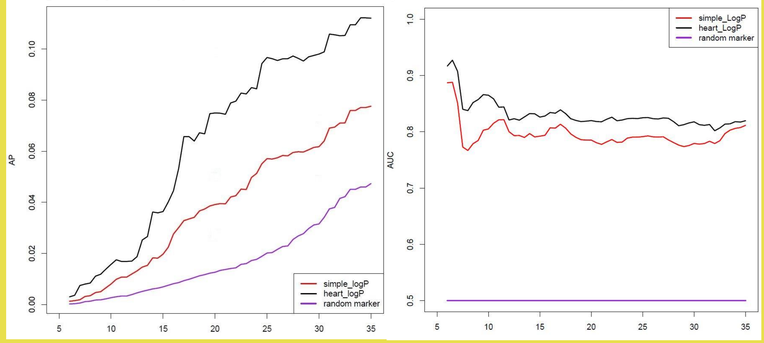
The density function of the estimator AP(t_0) is unknown, making the variance difficult to estimate. We use the standard nonparametric bootstrap for obtaining the standard error and confidence interval.

RESULTS

Table 1 (Simulation): Results for the estimator of AP(t_0) at three event rates and two sample sizes n, based on 1000 replications. "ESD" is the empirical standard deviation of the estimates; "ASE" is the average of the standard error obtained from the bootstrap resamples. "Cov" is the coverage probability.

Event rate	Sample Size	TRUE	BIAS	ESD	ASE	Cov
0.01	5000	0.095	-0.0063	0.030	0.027	0.87
	10000		-0.0024	0.022	0.021	0.91
0.05	5000	0.23	-0.0021	0.025	0.024	0.94
	10000		-0.0017	0.019	0.018	0.94
0.1	5000	0.33	-0.0021	0.022	0.021	0.94
	10000		-0.0017	0.015	0.015	0.95

Figure 1 (Example): The AP(t_0) (left) and AUC(t_0) (right) of risk scores predicting CHF in CCS.



References

1) Wilson PW, et al. Prediction of coronary heart disease using risk factor categories. Circulation 1998; 97:1837-1847.
2) Wolf PA, et al. Probability of stroke: a risk profile from the Framingham Study. Stroke 1991; 22:312-318.
3) World Health Organization, UNAIDS. Prevention of cardiovascular disease. World Health Organization; 2007.
4) Moskowitz CS el al. Quantifying and comparing the predictive accuracy of continuous prognostic factors for binary outcomes. Biostatistics, 2004;5:113–127
5) Zheng Y. et al. Semiparametric models of time-dependent predictive values of prognostic biomarkers. Biometrics, 2010; 66:50–60
6) Yuan Y et al. Threshold-free measures for assessing the performance of medical screening tests. Front. Public Health 2015; 3:57.
7) Chow et al. Individual prediction of heart failure among childhood cancer survivors. J Clin Oncol, 2015; 33:394-399.

Simulation Study

To examine the performance of the proposed estimator and inference procedure in finite samples, a simulation study is conducted.

- Generate the risk scores $Z_i \sim$ N(0, 0.5)
- Generate the event times using simulation model log(Ti) = β Z_i + ε, where β = −2 and $\varepsilon \sim$ N(0, 1.5).
- Generate the censoring time C_i from a gamma distribution with shape=1.7 and rate=1.6 to give an overall censoring rate of 50%.
- Obtain the observed event time X = min{T, C} and the censoring indicator δ = I($T < C$).
- Consider three prediction time points t_0 where the corresponding event rates are 0.01, 0.05 and 0.1, respectively.

Results are shown in Table 1.

Example

Chow et al. (2015)[7] developed and validated several risk score systems for predicting congestive heart failure (CHF) in childhood cancer survivors (CCS). For the purpose of illustration, we chose two risk score systems, a simple model vs. a heart dose model. Compared with the simple model, the heart dose model includes detailed clinical information on the average radiation dose to the heart and the cumulative dose of the specific chemotherapy agent used. The estimated linear predictors, denoted by logP, were treated as the continuous risk scores.

We include in our analysis 11,457 subjects from the Childhood Cancer Survivor Study who met the original study inclusion criteria and had the both risk scores. The time-dependent prediction accuracy were assessed with AP(t_0) and AUC(t_0) in Figure 1.

Discussion

The estimator and inference procedure for AP(t_0) works well, except for the under-coverage when AP(t_0) is small. Further investigation may be warranted to improve its coverage probability.

Between 15 and 35 years post diagnosis, the AUC(t_0) didn't change much within each risk score, or differentiate between the two risk scores. These observations are typical with the time-dependent AUC(t_0) which agrees with the criticism of AUC(t_0) being insensitive for comparing models.

The proposed AP(t_0) provides summaries of both PPV(t_0) and TPF(t_0) over the entire ranges of risk scores. It meets the need for a summary measure of prospective prediction accuracy.

AP is event rate dependent. Thus, it is possible that AP selects different risk score systems for different study populations.

Acknowledgment: QZ's research is supported by NSERC discovery grant, YY's research is supported by University of Alberta's start up grant.