STUDY TITLE

Evaluation of risk prediction models of late effects among childhood cancer survivors – A case study of the risk scoring systems for Congestive Heart Failure

WORKING GROUP

Primary: Epidemiology & Biostatistics

Secondary:

INVESTIGATORS

Name	*Institution*	*Specialty*	*Email*
Yan Yuan	University of Alberta	Biostatistics	yyuan@ualberta.ca
Qian Zhou	Simon Fraser University	Biostatistics	qmzhou@sfu.ca
Eric Chow	Fred Hutchinson CRC	Pediatric Oncology	ericchow@u.washington.edu
Greg Armstrong	St Jude CRH	Pediatric Oncology	Greg.Armstrong@stjude.org
Dan Mulrooney	St Jude CRH, University of Tennessee	Pediatric Oncology	daniel.mulrooney@stjude.org
Les Robison	St Jude CRH	Epidemiology	les.robison@stjude.org
Wendy Leisenring	Fred Hutchinson CRC	Biostatistics	wleisenr@fredhutch.org
Kevin Oeffinger	Memorial Sloan Cancer Center	Primary care	oeffingk@mskcc.org
Yutaka Yasui	University of Alberta	Biostatistics	yyasui@ualberta.ca

INTRODUCTION AND RATIONALE

Clinical decisions on disease management and early intervention are increasingly guided by risk scoring systems. The evaluation and comparison of risk scores depend on suitable performance measures. We propose a new risk prediction accuracy measure that is intended to be an especially

useful measure when the incidence rate is low. In addition to two methodological papers addressed to biostatisticians, we propose a knowledge translation paper that uses risk score systems derived from the Childhood Cancer Survivor Study (CCSS) (Chow *et al.,* 2015) to illustrate different performance measures with special attention to the clinical utility of these measures.

Risk prediction models have been widely used in medical research to predict the absolute risk (probability) of an adverse event of interest by a pre-determined set of time t_0, or to stratify apparently healthy individuals into different risk categories, given current known covariate values. For example, several risk scoring systems were developed for predicting congestive heart failure (CHF) among childhood cancer survivors by Chow *et al.* (2015). To evaluate and compare these risk scoring systems at various time points t_0, we need time-dependent accuracy measures, typically extensions of performance measures for predicting binary disease status. Time-dependent ROC curve and its threshold-free numeric summary index AUC(t) have been the most widely used measures in the literature to evaluate risk scores. While the AUC captures certain aspect of the performance of the classifiers such as discriminatory ability, it relies on how well the distribution of risk scores for cases and controls are separated, but it does not capture the actual predicted risks. Thus, AUC may not be optimal in assessing models that predict future risk or stratify individuals into risk categories. Cook (2007) pointed out that including a biomarker with a risk ratio of 3.0 may show little improvement on the AUC while it could shift the predicted 10-year cardiovascular risk for an individual patient from 8% to 24%, resulting in different recommendations on follow-up/intervention strategies. Using simulated examples, Wald and Bestwick (2014) showed that AUC can be an unreliable performance measure in the medical screening settings.

An alternative measure to the AUC, the positive predictive value (PPV), has been presented as more relevant to clinical utility and prediction accuracy (Zheng *et al.,* 2008 & 2010, Geoffrey *et al.,* 1994). Given a risk score, the PPV gives the absolute probability that a subject has the disease of interest. However, a disadvantage of the PPV is that it depends on a subjective threshold. To mend this threshold-dependency, we (Yuan *et al*., 2015) proposed a **threshold-free** numeric summary index of PPV, namely the average precision (AP) for prediction of binary disease status. In Yuan *et al*. (2015), clinical advantages of the AP were demonstrated and contrasted to the AUC in medical screening settings where the incidence of the event of interest is low. This is a relevant setting for risk prediction among childhood cancer survivors. For example, among childhood cancer survivors in CCSS, the cumulative incidence of selected serious cardiovascular events is 1.5~4% at 30 years post diagnosis

(Mulrooney *et al.* 2009). In our proposed investigation, we first sought to develop the time-dependent average precision (AP(t)) for the assessment of a single risk score.

In clinical settings, to be cost and time efficient, we would like to collect and incorporate minimum covariates into the risk score systems without losing significant prediction accuracy. Therefore, a second relevant methodological question is to evaluate whether incorporating additional covariates into an existing risk score would significantly improve prediction accuracy (both clinically and statistically). We propose to develop a method that assesses incremental values (IncV) in the AP(t) when new covariates/information is incorporated on top of the existing risk profile. That is, comparing the AP(t) of two possibly nested models/risk scores, e.g. one model with the standard covariates and the other model with standard covariates plus additional covariate(s).

The clinical setting and risk scores developed in Chow *et al*. (2015) is ideal for the illustration of the proposed new measure AP(t) and the incremental values in the AP(t). Chow *et al*. focused on the long term risk of CHF post cancer treatments in childhood cancer survivors and developed three risk score systems. We propose to examine these three risk scores using the AP(t): a simple risk score where chemotherapy agents and radiotherapy to the chest a yes/no binary variable, a standard risk score where the cumulative dose of certain chemotherapy agents and cumulative chest and heart doses of radiotherapy were incorporated and a heart dose risk score where the heart-specific average radiation dose were derived and used with the cumulative dose of certain chemotherapy agents. There is a clear need of assessing the incremental values of AP(t) for standard risk score vs. simple risk score and heart dose risk score vs. simple risk score. The clinical significance of the incremental value will aid clinical decision making as well as future data collection decisions.

Lastly, we propose a knowledge translation paper on risk prediction measures for the clinical research community in collaboration with clinicians and epidemiologists (such as Drs. Chow, Armstrong, Mulrooney, Oeffinger, and Robison). A number of measures have been used in the clinical research and publications for the evaluation of risk prediction models. Besides AUC and PPV, some of these other measures include the net reclassification index (NRI), integrated discriminatory improvement index (IDI), C statistic, Brier score, and the new proposed AP. Different measures could rank competing risk score systems in different orders. Table 1 gives an overview of pros and cons of the aforementioned measures. In Yuan *et al*. (2015), two risk scores for detecting the same condition can have similar AUC values but drastically different AP values. It could be confusing to navigate the potentially conflicting rankings given by various performance measures and choose the best risk score system that suits

specific needs of clinical research and/or practice. Thus, there is an urgent need to investigate methodologically how to align the choice of measures with the clinical context and questions. We will illustrate these various measures with a comprehensive analysis of the risk score systems, such as the CHF risk score systems developed with the CCSS data. For example, there are a total of 9 risk score systems to rank/group subjects in Chow *et al*. (2015) with the three underlying regression models. Each model gives three score systems to group subjects, e.g. ordinal risk score (range 0-11), categorical risk group (low, moderate and high) and numerical linear predictor from the underlying regression model. We will evaluate these nine possible score systems with the measures mentioned above, and discuss the clinical implications of the rankings by each measure for the end users, i.e. clinicians.

SPECIFIC OBJECTIVES:

1. Develop time-dependent average precision (AP(t)) with single risk score for the assessment of risk prediction scores; contrasting the AP(t) and AUC(t) using the three CHF risk scoring systems as an illustrative example; examining the properties of AP(t) estimator through simulation studies to investigate the validity of the proposed estimation and inference procedures, including consistency and empirical coverage probabilities of confidence intervals.

2. Incremental value: ΔAP(t) as a measure for the improved prediction accuracy from new variable/information – simulation studies to examine 1) nominal type I error rate when the null hypothesis of no improvement in accuracy by an addition of a covariate is true; 2) the ability to detect true incremental values of new variable/information when true improvement in prediction accuracy exists, e.g. the adjusted HR of the new marker is 3 or higher.

3. Knowledge translation: A comprehensive review of risk prediction measures and their application to the risk score systems developed by Chow *et al*. (2015)

APPROACHES:

<u>For Objective 1</u>

- Construct a robust consistent estimator of AP(t) for censored time-to-event data where a single risk score is the predictor.
- Derive large-sample inference procedures of AP(t)
- Examples – Illustrating and discussing the AP(t) and AUC(t) on
  - Three risk scoring systems derived for CHF in childhood cancer survivors – the simple, standard and heart dose scores.
- Simulation study for finite sample and large sample behavior
  - Simulation Settings
- Discussion and recommendations

Table 1: A list of typically used performance measures for risk prediction.

Performance Measure	**Numerical Range**	**Desirable features**	**Criticisms**
AUC(t)/ C-index (Discrimination measure)	0.5^{R} to 1^{P}	Can be estimated from case-control study; Invariant to cumulative incidence rate; Threshold independent; Has a (conditional) probability interpretation; Can be interpreted for one model or used to compare multiple models.	Increments in AUC insensitive to clinically important risk factor[3]; Unreliable measure in the medical screening setting[4]; Retrospective accuracy, which is less relevant to the end users (clinicians and patients) than the prospective accuracy is.
PPV(t, *v*) (Prediction measure)	Cumulative incident rate $\pi_t{}^{R}$ to 1^{P}	Prospective accuracy, simple yet meaningful, thus end-user (clinicians and patients) friendly; Can be interpreted for one model or used to compare multiple models.	Threshold dependent, i.e. need to specify quantile *v*.
AP(t) (Prediction measure)	$\pi_t{}^{R}$ to 1^{P}	Prospective accuracy, thus end-user friendly; Threshold independent; Can be interpreted for one model or used to compare multiple models.	To be identified.
$NRI_t(>0)$ (Incremental impact measure)	0 to 1*	Sensitive to clinically important changes in risk[9, 10]; Threshold independent.	Can only be used when **two** models are compared; Improper scoring rule; Misleading p-value[11,12].
IDI(t) (Incremental impact measure)	0 to $2\pi_t(1-\pi_t)$*	Sensitive to clinically important changes in risk[10]; Threshold independent.	Can only be used when **two** models are compared; Improper scoring rule; 95% CI estimation not valid when models are nested[11,13].
Brier score (BS(t)) (Residual variation measure)	0^{P} to $\pi_t(1-\pi_t)^{R}$	Overall measure for both model discrimination and calibration; Threshold independent; Can be interpreted for one model or used to compare multiple models.	Difficult to interpret; Lack clinical relevance.

^{R}: Random non-informative model (marker)

^{P}: Perfect model (marker)

*: The maximum values of IDI(t) and $NRI_t(>0)$ are achieved when a perfect model (marker) is compared to a random model (marker).

For Objective 2

- Propose time-dependent AP with multiple risk factors; here we will consider two types of models for estimating t_0-year risk: (1) Cox proportional hazards models, which assume constant marker effects over time; (2) time-dependent generalized linear models which allow marker effects vary with time t_0.
- Evaluate the incremental values (IncV) in AP(t) by adding specific treatment information (dose) on top of the existing risk factors age, sex and generic treatment information (illustrate with the risk scores developed by Chow *et al.* on CCSS data). We will compare the AP(t) of two models: one model with the existing risk factors and the other with both the existing risk factors and the specific treatment information.
- Propose the estimation and inference procedures for the IncV in AP(t) and conduct simulation study to assess objectives 2.1 and 2.2, i.e. type I error and power.
- Generalize the IncV in AP to compare any two risk scores (not necessarily nested).

For Objective 3

- Summarize statistical literature review to identify measures being proposed, clinical literature review to identify measures used, and identify pros and cons of these measures
- Illustrate commonly used measures on the CCSS data sets
- Discuss the interpretation and clinical utility of various performance measures, including assessment and preference for clinicians
- Make recommendations on how to choose performance measures based on the study objectives, clinical needs, population characteristics (such as incidence rate) etc. to select risk score system and report the performance.

REFERENCES:

1. Chow E, Chen Y., Kremer LC et al. (2015) Individual prediction of heart failure among childhood cancer survivors. JCO. 33:394-399.
2. Yuan Y, Su W and Zhu M (2015) Threshold-free measures for assessing the performance of medical screening tests. Front. Public Health 3:57. PMID 25941668 doi: 10.3389/fpubh.2015.00057
3. Cook NR (2007) Use and Misuse of the Receiver Operating Characteristic Curve in Risk Prediction. Circulation. 115(7):928-35.
4. Wald NJ and Bestwick JP (2014) Is the area under an ROC curve a valid measure of the performance of a screening or diagnostic test? J Med Screen. 21:51–6. doi:10.1177/0969141313517497
5. Zheng Y, Cai T, Pepe MS, Levy WC. (2008) Time-dependent predictive values of prognostic biomarkers with failure time outcome. Journal of the American Statistical Association, 103(481):362–368.
6. Zheng Y, Cai T, Stanford J, Feng Z. (2010) Semiparametric Models of Time-Dependent Predictive Values of Prognostic Biomarkers. Biometrics. 66:50-60.

7. Geoffrey A, John SA, David A, Robert B, Renaldo NB, Beaulieu M,*et al.* Canadian Task Force on the Periodic Health Examination. (1994) The Canadian Guide to Clinical Preventive Health Care. Canadian Government PubCentre.
8. Mulrooney DA, Yeazel MW, Kawashima T, *et al*.(2009) Cardiac outcomes in a cohort of adult survivors of childhood and adolescent cancer: retrospective analysis of the Childhood Cancer Survivor Study cohort. BMJ. 339:b4606.
9. Pencina MJ, Steyerberg EW, D'Agostino RB Sr (2011) Extensions of net reclassification improvement calculations to measure usefulness of new biomarkers. Stat Med. 2011 Jan 15; 30(1): 11–21.
10. Uno H, Tian L, Cai T, Kohane IS, Wei LJ. (2013) A unified inference procedure for a class of measures to assess improvement in risk prediction systems with survival data. Stat Med. 32(14): 2430–2442. doi:10.1002/sim.5647.
11. Hilden J, Gerds TA. (2013) A note on the evaluation of novel biomarkers: donot rely on integrated discrimination improvement and net reclassification index. Stat Med. doi:10.1002/sim.5804.
12. Pepe MS, Janes H, Li CI (2014) Net Risk Reclassification P Values: Valid or Misleading? J Natl Cancer Inst 106(4): dju041 doi:10.1093/jnci/dju041
13. Kerr KF, McClelland RL, Brown ER, Lumley T. (2011) Evaluating the incremental value of new biomarkers with integrated discrimination improvement. Am J Epidemiol. 174(3):364-74.