Biostatistics (2013), **0**, 0, *pp.* 1–4
doi:10.1093/biostatistics/rsfCompetingRisks.supp.submissions

Supplementary Material to Random Survival Forests for Competing Risks

HEMANT ISHWARAN*,[1], THOMAS A. GERDS[2], UDAYA B. KOGALUR[3],

RICHARD D. MOORE[4], STEPHEN J. GANGE[5], BRYAN M. LAU[5]

[1]*Division of Biostatistics, University of Miami,*
[2]*Department of Biostatistics, University of Copenhagen,*
[3]*Department of Quantitative Health Sciences, Cleveland Clinic,*
[4]*Department of Medicine, Johns Hopkins University, School of Medicine,*
[5]*Department of Epidemiology, Johns Hopkins University, Bloomberg School of Public Health*

A: Benchmark performance of RSF

To further study performance of RSF, we used benchmark data obtained from Pintilie (2006). These included Table 1.3b involving hypoxia (hypox); Table 1.4b involving follicular cell lymphoma; and Table 1.6b involving Hodgkin's disease (hd). Also included is the well known PBC data set from Appendix D of Fleming and Harrington (1991). All datasets involved two events.

We fit the data using the same methods as in Section 6 of Ishwaran et al. (2013). For RSF, only event-specific models were considered (logrank-split forests are denoted by RSF:LR and Gray-split forests by RSF:CR). All parameter settings were kept the same with one exception. The event-specific prediction error for each dataset was calculated by using 1000 random splits of the data into independent training sets (90%) and test sets (10%). This method was called bootstrap cross-validation (Mogensen et al., 2012). Prediction error was estimated using the integrated Brier score, $IBS_j(\tau)$, and the C-index, $C_j(\tau)$, for each event $j = 1, 2$. Throughout we estimated the censoring distribution using the Kaplan-Meier estimator. The average prediction error and average C-index over the 1000 splits and the corresponding standard deviation were calculated by averaging over the runs. The results are displayed in Table 1.

Table 1. *Leave 10%-out bootstrap cross-validated prediction errors and C-index for benchmark data*[†] *averaged over 1000 independent replicates (averaged standard errors are given in parentheses). The prediction errors of the null model (NM) which ignores all covariates are used for comparison.*

		Event 1		Event 2	
		$IBS_1(\tau)$	$C_1(\tau)$	$IBS_2(\tau)$	$C_2(\tau)$
hypox	NM	19.8 (5.5)	–	11.5 (6.1)	–
	RSF.LR	11.6 (5.3)	73.1 (20.5)	10.4 (5.4)	72.8 (20.2)
	RSF.CR	11.7 (5.3)	72.7 (20.6)	10.4 (5.4)	73.6 (20.2)
	CoxBoost.CV	19.8 (5.5)	50.0 (0.0)	11.5 (6.1)	50.0 (0.0)
	Cox	17.5 (6.1)	71.1 (17.8)	11.8 (6.4)	55.9 (29.8)
	FineGray	18.3 (6.8)	69.9 (18.2)	12.1 (6.7)	57.8 (27.4)
	CRRstep	19.0 (6.8)	67.7 (18.9)	11.5 (6.1)	56.3 (24.2)
follic	NM	22.3 (1.3)	–	6.7 (2.4)	–
	RSF.LR	22.9 (2.1)	55.8 (5.5)	6.4 (2.2)	71.2 (10.2)
	RSF.CR	23.0 (2.1)	56.1 (5.5)	6.4 (2.2)	71.2 (10.2)
	CoxBoost.CV	22.3 (1.3)	50.0 (0.0)	6.7 (2.4)	50.0 (0.0)
	Cox	21.7 (1.7)	58.1 (5.8)	6.3 (2.2)	71.7 (9.7)
	FineGray	21.7 (1.7)	58.4 (5.9)	6.3 (2.2)	71.5 (9.6)
	CRRstep	21.7 (1.7)	58.6 (5.8)	6.3 (2.2)	72.4 (9.3)
hd	NM	20.1 (1.6)	–	5.8 (1.6)	–
	RSF.LR	21.0 (1.9)	54.7 (5.2)	5.3 (1.3)	74.7 (7.8)
	RSF.CR	20.9 (1.9)	54.8 (5.2)	5.3 (1.3)	74.6 (7.8)
	CoxBoost.CV	19.6 (1.7)	58.7 (5.4)	5.2 (1.4)	76.9 (7.1)
	Cox	19.6 (1.7)	58.6 (5.4)	5.1 (1.3)	76.2 (7.3)
	FineGray	19.6 (1.7)	58.7 (5.4)	5.2 (1.4)	76.1 (7.2)
	CRRstep	19.6 (1.7)	58.9 (5.4)	5.2 (1.4)	76.3 (7.2)
pbc	NM	3.1 (1.9)	–	16.1 (2.3)	–
	RSF.LR	3.0 (1.7)	74.3 (18.3)	10.3 (1.9)	79.3 (5.9)
	RSF.CR	2.9 (1.7)	76.6 (17.7)	10.3 (1.9)	79.6 (5.9)
	CoxBoost.CV	3.1 (1.7)	80.3 (11.5)	10.8 (2.0)	79.3 (5.8)
	Cox	3.2 (1.7)	80.0 (11.5)	10.9 (2.1)	77.3 (5.9)
	FineGray	3.3 (1.8)	76.6 (13.0)	10.9 (2.1)	78.9 (5.7)
	CRRstep	3.1 (1.8)	80.2 (12.7)	12.7 (2.5)	74.6 (6.3)

Summary Values[‡] **for Datasets**

	n	D_0	D_1	D_2	p	τ
hypox	109	59	33	17	6	8
follic	541	193	272	76	4	15
hd	865	439	291	135	6	20
pbc	418	223	25	161	17	3000

[†]Data available at http://www.uhnres.utoronto.ca/labs/hill/People_Pintilie.htm

[‡]n is the sample size; D_0 is the number of censored observations; D_j is the number of type j events, $j \geqslant 1$; p equals the number of variables; and τ is the time of evaluation.

Generally, we find the results to be fairly comparable. For the hypox data one variable was excluded from cause-specific Cox regression and Fine-Gray regression to achieve model convergence. This leads to significantly reduced performance for these models as compared to RSF and CoxBoost.

The two splitting rules did not have a significant effect on performance. Also, data adaptive selection of boosting steps did not significantly affect performance of CoxBoost without such adaptivity.

It should be noted that all the datasets contain only a small pre-selected list of covariates, where variable selection was often based on semiparametric modeling. Thus, it could not be expected that the RSF or CoxBoost could outperform the semiparametric models. However, there are some instances where RSF appears better. For example, type 2 events for the primary biliary cirrhosis (pbc) dataset. This may indicate that the semiparametric models are misspecified.

B: Transportability and interpretation of RSF analysis

Analytical methods with the goal of prediction require an ease of interpretation and ability to be transportable to other populations. However, given that a random survival forest provides an estimate for the cumulative incidence function, which is defined as the probability of event J occurring by time t, the interpretation is fairly straightforward. This is in contrast to parametric or semi-parametric approaches which often present cause-specific hazards ratios or subdistribution hazards ratios for survival time. The hazard function as a rate is in our opinion less intuitive than a probability. In terms of transportability, as remarked in the Discussion of Ishwaran et al. (2013), it is possible to create software to permit a RF analysis to be restored at a later time for prediction on new data, thus for example, making it possible to apply RSF to clinical medical settings.

However, we point out that the use of a parametric or non-parametric approach may not always be the pertinent issue. Rather the transportability of the prediction model relies on the populations it is applied to. That is, if the population to which the model is to be applied is selected in a manner in which the attributes of some unmeasured factor is a modifier of the relationship between the covariates in the model (paramet-

ric or random forest) and the outcome, then it is unlikely for the model to be generalizable regardless of the analytical method (of course, if no such unmeasured factor exists, but the parametric method fails to correctly model the relationship between variables and the outcome, then the nonparametric method is at an advantage). Furthermore in competing risks, for prediction estimates to be transportable, the competing risk events must have the same distribution as the original study sample. Consider for the moment the non-competing risk situation, the upper bound for CIF for the event of interest is 1.0 such that by time infinity all individuals have the event (for a proper distribution). However, in the competing risk situation, by time infinity a certain proportion of individuals will have experienced the competing event preventing the event of interest to occur. This creates a boundary for the event of interest that is less than one. Now transporting the model to another population in which the competing event is even more likely to occur creates an even lower boundary than the original population. Thus even if the cause-specific hazards driving the event of interest does not change between populations, the CIF is not transportable (Lau et al., 2009). Rather a re-calibration of the CIF would be necessary.

REFERENCES

Fleming T. and Harrington D. *Counting Processes and Survival Analysis*. Wiley, New York, 1991.

Ishwaran H., Gerds T.A., Kogalur U.B., Moore R.D., Gange S.J., and Lau B.M. Random survival forests for competing risks. 2013.

Lau B., Cole S.R., Gange S.J. Competing Risk Regression Models for Epidemiologic Data. *American Journal of Epidemiology*, 170,244–256, 2009.

Mogensen U.B. and Ishwaran H. and Gerds T.A. Evaluating random forests for survival analysis using prediction error curves. *Journal of Statistical Software*, 50(11):1–23, 2012.

Pintilie M. *Competing Risks: A Practical Perspective*. John Wiley and Sons, West Sussex, 2006.