Supplementary material for "Boosting Transfer Learning with Survival Data from Heterogeneous Domains"

1 Theoretical results

The desirable convergence properties of Adaboost (Freund and Schapire, 1995) on the prediction error on the target population can be shown to hold in our setting, albeit with a modification of our algorithm and a more careful interpretation of what it means to make errors in survival predictions. The discussion below extends the results in (Dai et al., 2007) in which the authors were able to simultaneously minimize the error on source and target populations using a boosting algorithm for classification transfer.

In what follows we analyze the binary error measure,

$$I\left(\frac{1}{\tau}\int_0^\tau \mathbb{E}_{(T,X)\sim p_{ta}}\left(\left(I(T>t)-\hat{h}(t;X)\right)^2\right)dt>\phi\right) \tag{1}$$

that maps the survival prediction error into $\{0,1\}$ - incorrect/correct outcomes - to be interpreted as to whether prediction agree within ϕ of the true outcome. We note also that the results below hold for the version of our algorithm using $\beta_S = 1/(1-\sqrt{2\log(n_s/M)})$ and using the full data in every iteration. However in experimental evaluations, our adjusted error scheme and subsampling (discussed also below) improved performance and thus believe the theoretical results hold more generally. The proofs follow directly from Theorem 6 in (Freund and Schapire, 1995) and Theorems 3 and 4 in (Dai et al., 2007).

Proposition 1 *Suppose hypotheses $\hat{h}^{(m)}$ produce errors $\epsilon^{(m)}$, $m = 1,...,M$ as defined in step 4 of Algorithm 1 in the main paper. Then, the probability that $\hat{h}_f$ agrees within ϕ for any given instance from the target data is bounded above by,*

$$2^{\lceil M/2\rceil}\prod_{m=\lceil M/2\rceil}^{M}\sqrt{\epsilon_m(1-\epsilon_m)} \leq \exp\left(-2\sum_{m=\lceil M/2\rceil}^{M}(1/2-\epsilon_m)^2\right) \tag{2}$$

Proposition 2 *Let d_{VC} be the VC-dimension of the hypothesis space, the generalization error on the target distribution data, with high probability, is bounded above by,*

$$\epsilon + \mathcal{O}\left(\sqrt{\frac{Md_{VC}}{n_T}}\right) \tag{3}$$

Here, n_T is the size of the target population, and ϵ is the empirical error on the target training data given by equation (1) with survival predictions from $\hat{h}_f$.

The bound on the target training population error suggests that the error decreases exponentially fast as the number and accuracy of predictors increases. In turn the generalization bound characterizes the expected error increment over the in-sample performance given in Proposition 1. This highlights the risk of overfitting to the training data if the hypothesis space is too complex or too many predictors are combined.

1.1 Motivation for incorporating stochasticity

We define the mean squared error as the expected integral over all time horizons of the squared difference between the estimated $\hat{h}_f$ and true survival function S. Here $\hat{h}_f$ is the predicted survival function of the ensemble of trees and S is the true underlying survival distribution.

$$
\begin{aligned}
MSE(\hat{h}_f, S) &= \mathbb{E}\int \Big(\hat{h}_f(t;\boldsymbol{x}) - S(t;\boldsymbol{x})\Big)^2 dt = \mathbb{E}\int \Big(\hat{h}_f(t;\boldsymbol{x}) - \mathbb{E}\hat{h}_f(t;\boldsymbol{x})\Big)^2 + \Big(\mathbb{E}\hat{h}_f(t;\boldsymbol{x}) - S(t;\boldsymbol{x})\Big)^2 dt \\
&= \mathbb{E}\int \left(\sum_m \hat{\gamma}_m(\hat{h}_m(t;\boldsymbol{x}) - \mathbb{E}\hat{h}_m(t;\boldsymbol{x}))\right)^2 + \left(\sum_m \hat{\gamma}_m(\mathbb{E}\hat{h}_m(t;\boldsymbol{x}) - S(t;\boldsymbol{x}))\right)^2 dt \\
&= \int \sum_m \sum_k \hat{\gamma}_m\hat{\gamma}_k \Bigg(\mathbb{E}\Big((\hat{h}_m(t;\boldsymbol{x}) - \mathbb{E}\hat{h}_m(t;\boldsymbol{x})) \times (\hat{h}_k(t;\boldsymbol{x}) - \mathbb{E}\hat{h}_k(t;\boldsymbol{x}))\Big) + \\
&\quad \mathbb{E}\Big((\mathbb{E}\hat{h}_m(t;\boldsymbol{x}) - S(t;\boldsymbol{x}))(\mathbb{E}\hat{h}_k(t;\boldsymbol{x}) - S(t;\boldsymbol{x}))\Big) \Bigg) \\
&= \sum_m \hat{\gamma}_m^2 MSE(\hat{h}_m(\cdot;\boldsymbol{x})) + \\
&\quad \mathbb{E}\sum_m \sum_{k\neq m} \hat{\gamma}_m\hat{\gamma}_k \int (\mathbb{E}\hat{h}_m(t;\boldsymbol{x}) - S(t;\boldsymbol{x})) \times (\mathbb{E}\hat{h}_k(t;\boldsymbol{x}) - S(t;\boldsymbol{x})) + \mathrm{Cov}(\hat{h}_m(t;\boldsymbol{x}), \hat{h}_k(t;\boldsymbol{x}))dt
\end{aligned}
$$

where $\hat{h}_f(t;\boldsymbol{x}) = \sum_m \hat{\gamma}_m\hat{h}_m(t;\boldsymbol{x})$. Note the fact that $S(t;\boldsymbol{x}) = \sum_m \hat{\gamma}_m S(t;\boldsymbol{x})$ since $\sum_m \hat{\gamma}_m = 1$, used in line 3. Hence everything else being equal, lowering the correlation between successive weak learners reduces the mean squared error. This decomposition motivates combining trees trained on different samples of the data as their correlation tends to be lower.

Mean (Std. Dev.)	DIAMO	DIG	ECHOS	Euro	IN-CH
# of Patients	5486	7617	2880	8438	5499
Time to event (years)	1726 (1477)	1072 (446)	1078 (626)	91 (21)	325 (91)
Event occurrence	84%	33%	54%	7%	9%
Age (y)	71 (10)	63 (10)	73 (11)	70 (12)	62 (11)
Male	60%	75%	60%	53%	57%
Caucasian	99%	85%	97%	10%	34%
Body Mass Index	26 (5)	26 (5)	26 (5)	-	26 (4)
Sys. Blood Pres. (mmHg)	-	127 (20)	127 (20)	-	127 (20)
Dias. Blood Pres. (mmHg)	-	75 (11)	76 (12)	-	79 (10)
Ejection Fraction	31%	31%	43%	-	34%
Smoking	34%	-	28%	15%	73%
Creatinine	116 (60)	115 (56)	115 (62)	128 (104)	111 (59)
Stroke	-	-	10%	16%	1%
HF Duration	28 (46)	29 (36)	34 (48)	34 (48)	29 (42)
Beta Blocker	15%	36%	42%	38%	16%
ACE-inhibitor	50%	93%	55%	63%	84%
History Hypertension	24%	47%	26%	54%	13%
History MI	37%	69%	29%	35%	38%
History Diabetes	16%	28%	15%	27%	7%
History Atrial fibrillation	24%	0%	34%	23%	17%
Ischaemic aetiology	56%	69%	7%	59%	47%

Table 1: The main feature distribution of the 5 studies analyzed in the main body of this paper.

2 MAGGIC data

The Meta Analysis Global Group in Chronic Heart Failure (MAGGIC) (Pocock et al., 2012) performed a literature-based meta-analysis and extracted individual patient data from 30 studies regarding demographics, medical history, medical treatment, symptom status, clinical variables, laboratory variables and outcome. Table 1 gives mean and standard deviation summary statistics on the 5 studies considered in the main body of this paper.

2.1 Additional experiments

We have implemented all algorithms on the remaining 20 studies with more than 200 patients (anything below does not give a large enough test set for reliable performance computation). As can be seen in Table 2, our algorithm, TSB, outperforms in 10 of those studies and has competitive performance on the rest – giving a similar conclusion to the results in the main body of the paper. In all experiments we have also included results for all benchmarks trained on source data only – we found these slightly under-perform algorithms using both source and target, as consequentially for the former class there is a larger shift between training and testing data.

We explored in addition sub-sampling the training data prior to learning in an attempt to select those instances that are most closely related to the target. Survival Boosting (SB) is then trained on the target and the sampled subset of the auxiliary training data. We built a classifier (logistic regression) to determine the probability p of belonging to the target domain and selected those with $p > 0.75$ to train all benchmarks (in addition to target data). This strategy leads to performance of SurvBoost (a conventional survival model) which we denote SB (Samp) similar to Multitask RSF but below TSB on average. This suggests that learning predictions based on relevant instances in an integrated way, rather using a two-stage approach, is more efficient.

Studies	**Cox (T)**	**Cox (All)**	**SB (T)**	**SB (All)**	**SB (Samp)**	**M RSF**	**TSB**
BATTL	0.689	0.699	0.736	0.744	0.755	0.732	0.753
Berry	0.647	0.661	0.628	0.633	0.630	0.632	0.644
Gotsm	0.549	0.619	0.615	0.652	0.654	0.668	0.709
Grigo	0.509	0.539	0.535	0.543	0.560	0.559	0.579
Guazz	0.609	0.631	0.645	0.653	0.666	0.660	0.679
HFC E	0.600	0.615	0.622	0.603	0.625	0.629	0.627
HJL A	0.610	0.645	0.631	0.640	0.637	0.640	0.651
HOLA	0.699	0.677	0.710	0.703	0.704	0.689	0.691
IN-CH	0.699	0.698	0.705	0.721	0.730	0.709	0.741
Kirk	0.655	0.663	0.645	0.670	0.688	0.698	0.709
Macin	0.648	0.659	0.667	0.668	0.665	0.669	0.671
Mim B	0.587	0.600	0.601	0.633	0.618	0.600	0.579
Music	0.702	0.755	0.702	0.732	0.780	0.769	0.790
Newto	0.700	0.783	0.754	0.777	0.750	0.779	0.769
Rich	0.698	0.764	0.717	0.713	0.732	0.744	0.732
Richa	0.755	0.731	0.801	0.824	0.821	0.821	0.832
SRC A	0.683	0.690	0.708	0.703	0.710	0.715	0.721
Tribo	0.649	0.656	0.633	0.650	0.652	0.650	0.653
Tsuts	0.678	0.689	0.645	0.665	0.644	0.653	0.678
Varel	0.732	0.730	0.729	0.703	0.720	0.713	0.712

Table 2: C-index figures (higher better) and standard deviations on MAGGIC data studies.

References

Dai, W.; Yang, Q.; Xue, G.-R.; and Yu, Y. 2007. Boosting for transfer learning. In *Proceedings of the 24th international conference on Machine learning*, 193–200. ACM.

Freund, Y., and Schapire, R. E. 1995. A desicion-theoretic generalization of on-line learning and an application to boosting. In *European conference on computational learning theory*, 23–37. Springer.

Pocock, S. J.; Ariti, C. A.; McMurray, J. J.; Maggioni, A.; Køber, L.; Squire, I. B.; Swedberg, K.; Dobson, J.; Poppe, K. K.; Whalley, G. A.; et al. 2012. Predicting survival in heart failure: a risk score based on 39 372 patients from 30 studies. *European heart journal* 34(19):1404–1413.