# Boosting Transfer Learning with Survival Data from Heterogeneous Domains

**Alexis Bellot**
University of Cambridge, UK
Alan Turing Institute, London, UK

**Mihaela van der Schaar**
University of California Los Angeles, USA
University of Cambridge, UK
Alan Turing Institute, London, UK

## Abstract

Survival models derived from health care data are an important support to inform critical screening and therapeutic decisions. Most models however, do not generalize to populations outside the marginal and conditional distribution assumptions for which they were derived. This presents a significant barrier to the deployment of machine learning techniques into wider clinical practice as most medical studies are data *scarce*, especially for the analysis of time-to-event outcomes. In this work we propose a survival prediction model that is able to improve predictions on a small data domain of interest - such as a local hospital - by leveraging related data from other domains - such as data from other hospitals. We construct an ensemble of weak survival predictors which iteratively adapt the marginal distributions of the source and target data such that similar source patients contribute to the fit and ultimately improve predictions on target patients of interest. This represents the first boosting-based transfer learning algorithm in the survival analysis literature. We demonstrate the performance and utility of our algorithm on synthetic and real healthcare data collected at various locations.

## 1 Introduction

Survival models characterize the probability of event occurrence over time. Understanding risk of disease or death, and how they vary with time, is important to assist clinicians in their treatment policies and disease diagnosis. Crucially to be useful in practice these models need to perform well on each intended hospital or patient population, but often individual centers lack the data to train complex models for their patients. As an example, NHS hospitals in the UK have on average 89 beds per hospital[1] - which for rare diseases can lead to fewer than 50 registered outcomes of interest per year. This problem is exacerbated in survival settings since the time of death or disease onset may be unobserved for many patients - these are called *censored* patients. Transferring patient data from similar domains is useful in such settings, but special care needs to be taken such as to ensure that it results on reliable predictions on the target domain (Wainberg et al., 2018).

Patient characteristics are governed by a joint distribution $p(X, T)$, where patient features $X$ take values in a space $\mathcal{X}$ and $T \in \mathbb{R}^+$ represents a patient's time to event. Using data from source domains different from a target domain presents major challenges because differences may occur (1) in the actual variables observed - $\mathcal{X}_{so} \neq \mathcal{X}_{ta}$ - (which we refer to as *heterogeneous domains*) and (2) in their joint distribution - $p(X_{so}, T_{so}) \neq p(X_{ta}, T_{ta})$. Hospitals for instance often have different standards of data recording, measurements taken for our target population of patients might not all be available in a different hospital population even though they might overlap substantially and be useful to use. Patients also tend to be highly heterogeneous, survival dynamics and covariate relationship cannot be expected to be alike across diseases or locations for example, that is joint distributions might differ. To the best of our knowledge, in this work we present the first transfer learning method for survival data that overcomes both these challenges.

Our proposed approach intends to provide a new charac-

Proceedings of the 22nd International Conference on Artificial Intelligence and Statistics (AISTATS) 2019, Naha, Okinawa, Japan. PMLR: Volume 89. 

[1] www.kingsfund.org.uk/publications/nhs-hospital-bed-numbers

terization of boosting algorithms in a transfer learning scenario (Dai et al., 2007; Pardoe and Stone, 2010) that effectively extends this family to time-to-event data from multiple sources that may have feature mismatch. To enable this, we define weak nonparametric survival trees that learn a shared representation between related domains that implicitly corrects for marginal distribution differences such as to improve predictions on the target population over successive iterations via instance boosting. To correct for feature mismatch the space of splitting rules of trees is expanded to include "missingness" of a variable itself as a valid splitting rule, in this way even source patients with only overlapping feature spaces can be used in the tree construction if they are considered to improve the fit. The proposed boosting process filters those source patients considered to favourably impact predictions and increase the predictive power over using the target data only. An important distinction of this approach to classification and regression algorithms relates to the error measure used to evaluate survival predictions of individual learners on an instance-basis; these need to accommodate for censoring and different prediction horizons.

### 1.1 Related work

The **problem** we consider is that of learning survival distributions from *censored data* by leveraging auxiliary data from *heterogeneous domains* with *no* assumptions on their joint distributional properties. Our **goal** is to improve prediction performance on the target data *only*.

Knowledge transfer is a rich field, our work touches on heterogeneous transfer learning, instance-based transfer learning and survival analysis. However, our problem differs in ways that make most of the existing approaches not directly applicable to our problem.

**Heterogeneous transfer learning** attempts to utilize domains with different feature spaces, but most often from a domain adaptation perspective, that is, they estimate a latent domain representation close in distribution to the target domain with the objective of augmenting the number of target examples. (Yoon, Jordon, and van der Schaar, 2018) is a recent example of this approach. Closer to our work is (Wiens, Guttag, and Horvitz, 2014). Like us, they attempt to improve predictions in a given hospital using potentially useful information from patients in different hospitals, but crucially do not account for distribution mismatch which will render the model biased if distributions differ. A key difference with the above is that none of them consider learning from censored data and acknowledge difficulties when the target data is small - which is precisely the strength of the proposed approach.

**Instance-based transfer learning** methods reweight individual examples as a mechanism to correct for distributional differences. Because of the heterogeneity in modern datasets subgroups or even individuals in auxiliary datasets may be close in distribution to a target population while other source individuals are not. This idea motivated the design of our algorithm and has found particular success in boosting-based approaches. Our method is closest to the work of (Dai et al., 2007); the authors developed a classification transfer algorithm that iteratively combines weighted instances from auxiliary and target data to enable transfer learning. Boosting survival predictions however, requires a new definition of how to measure error on an individual basis (Bellot and van der Schaar, 2018) that appropriately considers censored patients. Moreover (Dai et al., 2007) (and also posterior extensions such as Pardoe and Stone (2010); Yao and Doretto (2010)) assume equal feature spaces. In this manner, we consider our work to be an extension to boosting methods for transfer learning for survival analysis from heterogeneous domains.

Within the **survival analysis** literature, to the best of our knowledge the only method that attempts transfer learning was recently proposed in (Li et al., 2016). They proposed an extension to Cox model; a linear model that assumes proportionality of hazards, which means that the relative rate of mortality between two patients stays constant over time. As a practical consequence, if with respect to a given patient survival probabilities of another patient are higher at one point in time it will be higher for all times, which is often not realistic to assume. Other important distinctions to our approach is that we allow for heterogeneous domains and make no assumptions on the data generating process. **Multitask learning** methods have been used for survival analysis and are often associated with transfer learning (Caruana, 1997), but their underlying objective is to discover a common representation among multiple tasks in order to improve generalization ability on *all* tasks. Critically this assumes a *common* data distribution for the whole population and requires a uniform feature space, which makes them biased in our context.

## 2 Problem Description

We use medicine as a running example but the method and analysis we introduce are more general and apply to fields of study such as reliability analysis in engineering and economics.

A domain $\mathcal{D}$ consists of a feature space $\mathcal{X}$ - such as for example $\mathbb{R}^d$ - and a marginal probability distribution $p(X)$, where $X \in \mathcal{X}$. $X_i$ describes an individual patient $i$ and $T_i \in \mathbb{R}^+$ defines the time to the event of

interest. We use subscripts $so$ and $ta$ to denote source and target domains respectively. Differing domains in realistic applications will often result in heterogeneous feature spaces $\mathcal{X}_{so} \neq \mathcal{X}_{ta}$; different marginal distributions $p_{so}(X) \neq p_{ta}(X)$ or different conditional distributions $p_{so}(T|X) \neq p_{ta}(T|X)$. In medicine the interpretation of time to event $T$ might coincide for different tasks, for example time from cancer diagnosis to death in different hospitals, or not, for example time to cancer onset versus time to cardiovascular disease onset. In addition, patients being followed in a medical study may drop-out resulting in a potential event being unobserved, $C$ represents this censoring time. We define a variable $\delta = I(T < C)$ that indicates the type of event observed and assumed independent of $X$. For a specific domain, a time-to-event dataset is assumed to be drawn $i.i.d$ from the random tuple $(X, \delta T + (1-\delta)C, \delta)$.

Our goal is to estimate the survival function $S$ for target patients which represents the probability of event occurrence after time $t$ as a function of patient covariates $X$ and time $t$,

$$S(t|X) = \mathbb{P}(T > t|X) \tag{1}$$

given labeled data from source and target domains. The objective is to improve survival predictions with the potentially useful information gained from a source domain.

## 3 Survival Transfer Algorithm

This section discusses our main contribution: a boosting algorithm for predicting time to event distributions in data scarce situations for which data from related tasks is available. We call our method TSB, short for Transfer Survival Boosting.

### 3.1 Weak Predictors

Weak predictors are trees composed of leaves and nodes. Leaves define a partition for the data and are responsible for making predictions and nodes guide examples towards appropriate leaves using binary splits based on boolean-valued rules.

**Splitting on different feature spaces** - We define a binary recursive partitioning scheme such that every node splits the whole population - both source and target - into homogeneous subsets with similar survival behaviour. Homogeneity is measured with the reduction in model deviance - a measure of goodness of fit - assuming an exponential model in each node (LeBlanc and Crowley, 1992). In other words, we choose the split that maximizes the likelihood ratio statistic that compares the likelihood of the resulting split with the likelihood of the original population. A variable proposed to split a given node might not be observed for all patients. For these patients three natural splitting procedures are proposed: (1) Patients with the considered variable missing are sent to the left child node, (2) Patients with the considered variable missing are sent to the right child node or (3) 'missingness' itself is used to partition the population.

**Survival Predictions** - Predictions in each leaf of the resulting tree are made with the Kaplan-Meier estimator. Let $\mathcal{C}_j$ denote the index set of patients with terminal node $j$, we compute survival predictions in the terminal node $j$ with the Kaplan-Meier estimator,

$$\hat{h}_j(t) = \prod_{i \in \mathcal{C}_j : t_i \leq t} \left(1 - \frac{N_j(t_i)}{Y_j(t_i)}\right) \tag{2}$$

where $N_j(t_i)$ is the number of events at time $t_i$ in terminal node $j$ and $Y_j(t_i)$ is the total number of individuals at risk at time just before $t_i$ in terminal node $j$. Leaf nodes define the survival function for the tree,

$$\hat{h}(t; \boldsymbol{x}_i) = \sum_j I(i \in \mathcal{C}_j)\hat{h}_j(t) \tag{3}$$

We note that in none of the steps described above assumptions are being made on the data generating process. Most survival models in turn assume accelerated failure times or proportional hazards which can be limiting if patient populations are complex and heterogeneous.

### 3.2 Ensemble Construction

Many patient populations that we are interested in describing to improve day to day care are of small size. To enable transfer learning we would like part of the auxiliary populations - those that are most similar to our target patients - to play a role in the ensemble construction. The aim of our boosting architecture is to filter out those patients that are most dissimilar to our target population - and thus induce *negative transfer* - while incorporating those that are most similar and thus help target predictions. Figure 1 shows intuitively how this process can benefit the resulting overall fit.

**Individual errors** - Boosting architectures define a re-weighting scheme that encourages the algorithm to focus on those *instances of interest* that are being mis-predicted in the data to improve the fit overall. Mis-predictions in the survival setting are not well defined since a comparison has to be made between an estimated survival distribution and the observed event-time (or censoring time), in contrast to classification or regression approaches. We generalize existing implementations by considering a consistent error measure

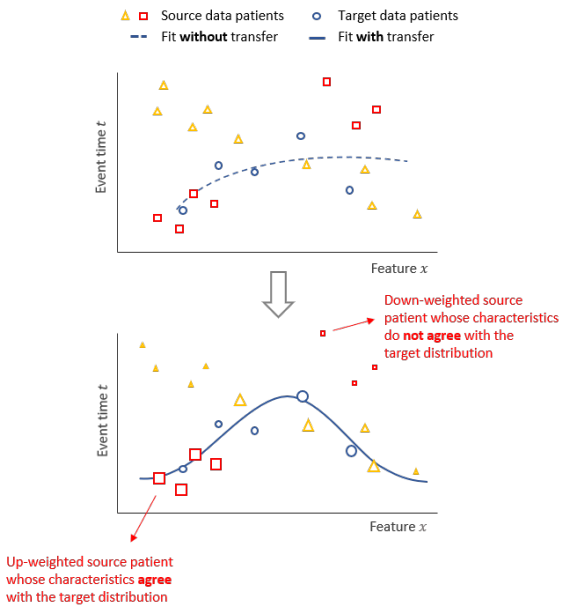


Figure 1: Transfer learning through instance boosting. The upper panel shows predictions of our algorithm without using auxiliary datasets while in the bottom panel our algorithm leverages those instances from related datasets that improve the fit on our target task (those are assigned higher weight - larger symbols), resulting in more accurate predictions. For simplicity of exposition we have ignored censoring and only show expected event times (rather than full predicted survival distributions).

for data subject to censoring and extend this to capture the full time horizon. On a per instance basis, we use the integrated brier score, a variant of the mean squared error adapted to the survival setting and integrated to capture all future time horizons.

$$\frac{1}{\tau}\int_0^{\tau} \mathbb{E}_{(T,X)\sim p_{ta}}\left[\left(I(T>t)-\hat{h}(t;X)\right)^2\right]dt \qquad (4)$$

$I$ stands for the indicator function and $\tau$ is defined as the maximum observed event time. Note that the expectation is taken with respect to the target population only. We approximate equation (4) by a consistent estimate and define individual errors of each weak predictor using a threshold $\phi$ (estimated by cross-validation) as follows,

$$O_i := \delta_i T_i + (1-\delta_i)C_i$$
$$e_i := I\left(\frac{1}{\tau}\int_0^{\tau}\hat{W}_i(t)\left(I(O_i>t)-\hat{h}(t;\boldsymbol{x}_i)\right)^2 dt > \phi\right) \qquad (5)$$

$\tau$ is the maximum observed event time and $O_i$ refers to the observed time of patient $i$. For censored patients ($\delta_i = 0$), the time to the event of interest will be unobserved and thus we approximate the integrand in equation (4) using inverse probability of censoring weights at each time $t$, $\hat{W}_i(t)$ (Mogensen, Ishwaran, and Gerds, 2012) which - assuming they are estimated consistently - ensure our estimator is consistent with the underlying Brier Score.

$$\hat{W}_i(t) = \frac{(1-I(T_i>t))\delta_i}{\hat{G}(T_i)} + \frac{I(T_i>t)}{\hat{G}(t)} \qquad (6)$$

$\hat{G}$ is an estimate of the censoring distribution.

**Algorithm 1**

**Input:** Target survival data $\mathcal{D}_{ta}$ of size $n_{ta}$, source survival data $\mathcal{D}_{so}$ of size $n_{so}$, number of iterations $M$, initial weights $w_i^{(1)} \propto 1$, sampling fraction $s$ (threshold $\phi$ is selected by cross-validation).

**for** $m = 1$ **to** $M$ **do**

1. Let $\mathcal{D}$ be a randomly sampled fraction $s$ of the combined data $\{\mathcal{D}_{ta}, \mathcal{D}_{so}\}$ with distribution $w^{(m)}$ (a vector of size $n_{ta}+n_{so}$).
2. Learn hypothesis $h^{(m)} : \mathcal{X}\times T \to [0,1]$ on $\mathcal{D}$.
3. Calculate prediction error $e_i^{(m)}$ for each instance $i = 1,...,n_{ta}+n_{so}$ with equation (5).
4. Calculate adjusted error of $h^{(m)}$ on the target population, $\epsilon^{(m)} = \sum_{i=1}^{n_{ta}} e_i^{(m)} w_i^{(m)}$.
5. Calculate confidence in individual hypothesis $\beta_T^{(m)} = \frac{\epsilon^{(m)}}{1-\epsilon^{(m)}}$:
6. Update data distribution

$$w_i^{(m+1)} \propto \begin{cases} w_i^{(m)}(\beta_T^{(m)})^{-e_i^{(m)}}, & i = 1,...,n_{ta} \\ w_i^{(m)}(\beta_S^{(m)})^{e_i^{(m)}}, & i = n_{ta}+1,...,n_{ta}+n_{so} \end{cases}.$$

where $\beta_S^{(m)}$ is chosen such that

$$\sum_{i=n_{ta}+1}^{n_{ta}+n_{so}} w_i^{(m)} / \sum_{i=1}^{n_{ta}+n_{so}} w_i^{(m)} = \frac{n_{so}}{n_{so}+n_{ta}}\left(1-\frac{mn_{ta}}{2Mn_{so}}\right)$$

**end for**

**Output:** Final hypothesis $h_f$, the weighted median of $h^{(m)}$ for $\lceil M/2\rceil \le m \le M$ using $\log(1/\beta_T^{(m)})$ as the weight of hypothesis $h^{(m)}$.

**Learning to Transfer** - A sequence of successive weak survival predictors form the building blocks of our learning algorithm. We give a formal description of our method in Algorithm 1. In a given iteration, a weak survival predictor is trained on a weighted sample of a combination of source and target data; such as (1) to improve survival predictions of correctly predicted patients from the target data - similarly to conventional

boosting algorithms - while also; (2) increasing the influence of correctly predicted source data patients - intuitively those are most similar to our target patients - and decreasing the influence of incorrectly predicted source data patients - those are most dis-similar to our target patients. The magnitude of these updates is given by $\beta_S$ and $\beta_T$ for the source and target data respectively.

This idea was also used in classification and regression transfer algorithms in (Dai et al., 2007) and (Pardoe and Stone, 2010). However as discussed in (Al-Stouhi and Reddy, 2011), previous boosting algorithms such as (Dai et al., 2007) although provably optimal, in practice even source instances that are representative of the target population tend to have their weights reduced quickly and erratically such that they become irrelevant and no longer influence the combined output. The error scheme of (Dai et al., 2007) fixes $\beta_S = 1/(1 - \sqrt{2\log(n_s/M)})$. We opt instead for an *adjusted* error scheme based on experimental approximation similarly to (Pardoe and Stone, 2010). $\beta_S$ is selected such that the sum of source weights smoothly decreases to $1/2 \times n_{ta}/n_{so}$ over successive iterations, we found this to balance target and source influence well in experimental settings.

A decomposition of the mean squared error suggests improved performance with uncorrelated weak predictors, which we encourage by randomly sub-sampling the data before training each weak predictor (see step 1. in Algorithm 1 and derivation in the Supplementary material).

Finally, survival predictions on the target data $\hat{h}_f$ result from a weighted median of the predictions of individual weak survival predictors $\hat{h}^{(m)}$, weighted by $log(1/\beta^{(m)})$ which can be thought of as a measure of confidence in $\hat{h}^{(m)}$ since higher $log(1/\beta^{(m)})$ implies higher overall predictive performance of $\hat{h}^{(m)}$ on the target population.

$$\hat{h}_f(t;\boldsymbol{x}_i) := \\ median(\{\log(1/\beta_T^{(m)})\hat{h}^{(m)}(t;\boldsymbol{x}_i)\}_{m=\lceil M/2 \rceil}^{M}) \quad (7)$$

### 3.3 Discussion on Convergence Guarantees

The desirable convergence properties of Adaboost (Freund and Schapire, 1995; Dai et al., 2007) 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 binary error measure in equation (5) maps the survival prediction error into $\{0, 1\}$ - incorrect/correct outcomes - to be interpreted as to whether prediction agree within $\phi$ of the true outcome where agreement is measured with equation (4). The key intuition to ensure the error bound of Adaboost in the survival setting is that our final hypothesis $\hat{h}_f$ (in the form of a weighted median) "$\phi$-disagrees" on a patient $i$ with its true outcome only if *more* than (weighted) half of the learned weak predictors "$\phi$-disagree", then combining a large number of predictors (assumed weak learners) exponentially decreases the training error. This observation is used analogously in the original derivation of the error bound in Freund and Schapire (1995) for classification with the concept of majority voting.

(Dai et al., 2007) leveraged this result in the context of transfer learning for classification to simultaneously minimize the error on source and target populations. By instead using the error scheme of (Dai et al., 2007) that fixes $\beta_S = 1/(1-\sqrt{2\log(n_s/M)})$ and using the full data in every iteration; a straightforward modification of the convergence results in (Dai et al., 2007) apply to our setting and are given in the Supplementary material.

## 4 Experiments

### 4.1 Prediction Performance Evaluation

We measure performance with a common metric used in the literature: the (time-dependent) concordance index ($C$-index) defined as follows (Wolbers et al., 2014):

$$C(t) := \mathbb{P}(\hat{S}_i(t) > \hat{S}_j(t) | \delta_i = 1, t \leq T_j, T_i > T_j) \quad (8)$$

where $\hat{S}_i(t)$ is the predicted survival probability beyond time $t$ for a test patient $i$. The time-dependent $C$-index as defined above corresponds to the probability that predicted survival probabilities are ranked in *accordance* to the actual observed survival times given the occurrence of an event. The $C$-index thus serves as a measure of the discriminative power of a model and can be interpreted as an extension of the AUROC for censored data. Random guessing corresponds to a $C$-index of 0.5 and perfect prediction to a $C$-index of 1. On all experiments the $C$-index is computed *only* for patients in the target population.

### 4.2 Synthetic Experiments

In this paper, we address the distributional differences take place in the marginal $p(x)$ and in the conditional distributions $p(t|x)$. In the following subsections we analyze the extent to which transfer learning is beneficial by varying the similarity in the source and target domain distributions.

**Scenario description -** 10 covariates are drawn from a uniform distribution $\mathcal{U}(-1.5.2.5)$ with the first 5 covariates influencing survival through the following asso-

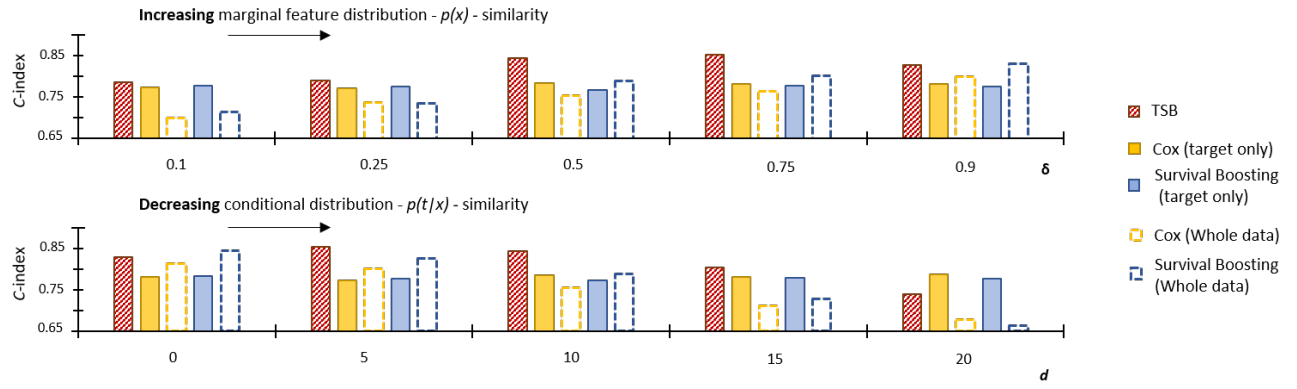


Figure 2: Performance on synthetic data experiments as a function of marginal and conditional distribution similarities.

ciation rule:

$$\Lambda(\boldsymbol{x};\boldsymbol{a}) := 10a_1x_1(x_1-1)(x_1+1)(x_1-2) + 5a_2x_2^2 - 2a_3\log(x_3+3) - a_4(x_4-1)^2$$

then $T|\boldsymbol{a} \sim \mathcal{N}(100+\Lambda(\boldsymbol{X};\boldsymbol{a}),10)$, and censoring is imposed on 20% of the population by drawing $C \sim \mathcal{U}(0,T)$. By varying the proportion of samples with covariates in a specific interval $L \subset (-1.5,2.5)$ we control the amount of overlap between source and target marginal feature distributions while the $a_i$'s govern the underlying relationships between features and time-to-event conditional distribution. In all experiments a target population of 100 instances is generated by restricting covariates to lie in the interval $L=(-1,1)$ and each $a_i$ to be equal to 1; we set the number of source instances to $1,000$ but vary their marginal and conditional distributions as detailed below. Performance is computed on a held out target dataset of $10,000$ patients.

**Baselines** - We compare TSB with conventional survival models: (1) trained using both the source and target data, and (2) trained solely on the target data. We evaluate the original Cox proportional hazard model (Cox, 1972) which serves as a semi-parametric alternative that is expected to perform well when the underlying variable interactions are linear and the hazard is unspecified. Next, we implement the boosting algorithm for survival data in (Bellot and van der Schaar, 2018). This method is relevant because like ours they do not make assumptions on the data generating process and use a boosting scheme to improve survival predictions over time; the comparison thus shows the extent to which transfer learning influences performance. In all experiments the tree-depth of TSB is set to 3 and 250 boosting iterations, while hyperparameter settings of all competing algorithms are set to default specifications.

**Performance as $p(x)$ varies -** We investigate the influence of marginal feature differences as follows. A proportion $\delta$ of source patients have their covariates generated uniformly in the interval $(-1,1)$ - just as target patients - while the remaining $(1-\delta)$ portion of source patients have their covariates generated uniformly in $(-1.5,-1)\cup(1,2.5)$. $\delta$ in this way controls the overlap in these patients populations. Performance results computed with the $C$-index are given in the upper panel of Figure 2. TSB's main contribution occurs on data with a moderate degree of feature distribution overlap, in the more extreme cases of little overlap, algorithms using only the target slightly outperform TSB, as would be expected since very little knowledge is to be transferred from auxiliary patients. The same logic holds for the case of large overlap.

**Performance as $p(t|x)$ varies -** Transfer learning should be possible even if the underlying survival relationship mildly differs among different populations. For example, such a situation is to be expected in patient cohorts at risk of related diseases. We generate synthetic data to mimic this behaviour by drawing the $a_i$'s from $\mathcal{N}(1,0.1d)$ while the marginal feature distributions are set to be equal for both source and target. Larger values of $d$ will result in larger survival difference between source and target populations - recall that for the target population the $a_i$'s are equal to 1. The lower panel of Figure 2 gives performance as a function of $d$. For $d=0$ source and target populations are equal in distributions, it is natural thus for conventional survival models to outperform. For increasing values of $d$ the gain of TSB becomes apparent as the competing methods are not designed for transfer. $d=20$ corresponds to the extreme case of no relationship between source and target survival patterns, using the target data only gives better performance.

Notice that on both panels on Figure 2 performance of "target only" methods stays the same in all experiments as only the source population is modified in each

| Models | DIAMO | DIG | ECHOS | Euro | IN-CH |
|---|---|---|---|---|---|
| Cox (Target) | $0.609 \pm 0.01$ | $0.585 \pm 0.02$ | $0.603 \pm 0.02$ | $0.500 \pm 0.02$ | $0.579 \pm 0.01$ |
| SurvBoost (Target) | $0.593 \pm 0.01$ | $0.609 \pm 0.02$ | $0.621 \pm 0.02$ | $0.562 \pm 0.02$ | $0.619 \pm 0.02$ |
| Cox (All) | $0.619 \pm 0.01$ | $\mathbf{0.621 \pm 0.01}$ | $0.617 \pm 0.01$ | $0.640 \pm 0.01$ | $0.649 \pm 0.01$ |
| SurvBoost (All) | $0.610 \pm 0.01$ | $0.600 \pm 0.01$ | $0.639 \pm 0.01$ | $0.637 \pm 0.02$ | $0.643 \pm 0.01$ |
| DiscoGAN | $0.552 \pm 0.02$ | $0.550 \pm 0.03$ | $0.590 \pm 0.02$ | $0.577 \pm 0.02$ | $0.587 \pm 0.02$ |
| Multitask RSF | $0.620 \pm 0.01$ | $0.618 \pm 0.01$ | $0.648 \pm 0.02$ | $0.648 \pm 0.01$ | $0.661 \pm 0.01$ |
| TSB | $\mathbf{0.638 \pm 0.01}$ | $0.612 \pm 0.03$ | $\mathbf{0.666 \pm 0.01}$ | $\mathbf{0.677 \pm 0.01}$ | $\mathbf{0.686 \pm 0.01}$ |

Table 1: $C$-index figures (higher better) and standard deviations on MAGGIC data studies.

experiment.

### 4.3 Real Data Experiments

This section investigates MAGGIC (Pocock et al., 2012), a collection of real datasets from 30 different medical studies containing patients that experienced heart failure. The question of how to transfer medically relevant information across patients in different studies is important as many of them contain fewer than 100 patients.

**Data description -** Each study contains a number of patients ranging from as little as 66 to 8438. We consider the 5 studies with highest amount of patients which we subsample to 100 patients for training and the rest for testing, we repeat this process 5 times to give a measure of uncertainty around mean performance. For patients in each one of the selected studies our aim is to improve survival predictions using data from the remaining 4 studies. On average 22 variables overlap in the considered datasets, with the highest number of observed variables in a single dataset being 30 variables - these correspond to demographics, lab tests, co-morbidities and medications. 32% of all patients died while the rest were censored before death time. Study-specific data statistics and an analysis of all remaining studies within MAGGIC are given in the Supplementary material.

**Algorithms for performance comparison** - In order to evaluate competing algorithms that do not account for heterogeneous feature spaces we restrict the input for these methods to overlapping features. We compared performance with the multitask algorithm Random Survival Forest (Ishwaran et al., 2008), originally designed for competing risks. A relevant comparison can be also made with the unsupervised domain translation GAN in (Kim et al., 2017). We proceed by adapting the feature distribution of source data to the target domain and predict with a Cox model on the augmented data. Target only and combined source and target algorithms implemented in the synthetic data experiments are also included for comparison on real data. In all experiments the tree-depth of TSB is set to 3 and 250 boosting iterations, while hyperparameter settings of all competing algorithms are set to default specifications.

#### 4.3.1 Discussion of Performance

As can be seen in Table 1 mean performance of TSB is highest in comparison to all other methods in all but one study. Methods using source data in combination with target data outperform target only methods, in this case even indiscriminately including auxiliary patients improves the fit as the low number of target patients does not allow for reliable predictions using only this set. Among these methods we observe that the more flexible boosting algorithm provides little (if any) benefit over Cox which suggests that the underlying feature interactions are likely close to linear, this is in contrast to the performance pattern observed in the synthetic experiment. Multitask Random Forest is able to successfully leverage the shared relationship between the studies which is effective is most cases but its performance lags because it does not directly consider shifts in distributions.

Because of the low number of target examples the unsupervised GAN model - DiscoGAN - is unable to effectively translate the source distribution and underperforms in all cases.

#### 4.3.2 Source of Gain

**Number of target patients -** We selected the study DIAMO to illustrate the influence of the number of target patients on predictive performance. We randomly subsample DIAMO increasing the number of training instances from 50 to 1000, with the remaining patients being used for testing. As can be seen in Table 2 the main contribution of TSB occurs when fewer than 200 target patients are available. This performance gain becomes marginal beyond that point as conventional supervised learning are appropriately powered for the

complexity of the data.

| # Target patients | **SurvBoost (Target)** | **TSB** |
|---|---|---|
| 50 | $0.541 \pm 0.08$ | $0.610 \pm 0.06$ |
| 100 | $0.582 \pm 0.08$ | $0.635 \pm 0.07$ |
| 200 | $0.628 \pm 0.06$ | $0.649 \pm 0.05$ |
| 300 | $0.540 \pm 0.06$ | $0.652 \pm 0.05$ |
| 500 | $0.663 \pm 0.04$ | $0.650 \pm 0.04$ |
| 1000 | $0.668 \pm 0.04$ | $0.653 \pm 0.04$ |

Table 2: $C$-index performance on DIAMO as a function of the number of target patients.

**Heterogeneous domains -** In Table 3 we compared across all 5 studies the performance of TSB using its full capabilities to TSB fit using only variables recorded in all studies - 22 patient variables. The performance gain can be attributed to the information present in additional variables.

| Study | **All variables** | **Overlapping variables** |
|---|---|---|
| DIAMO | $0.638 \pm 0.01$ | $0.625 \pm 0.01$ |
| DIG | $0.612 \pm 0.03$ | $0.614 \pm 0.02$ |
| ECHOS | $0.666 \pm 0.01$ | $0.660 \pm 0.01$ |
| Euro | $0.677 \pm 0.01$ | $0.678 \pm 0.01$ |
| IN-CH | $0.686 \pm 0.01$ | $0.677 \pm 0.01$ |

Table 3: $C$-index performance gain due to heterogeneous domains.

#### 4.3.3 Deeper Analysis of TSB fit

TSB filters out those patients from auxiliary hospitals that do not conform to the survival behaviour observed in the target hospital. We analyze in this section the behaviour of TSB fit on ECHOS as target patient population with the remaining 4 studies as auxiliary source data. We explicitly analyze patients by looking at the average weight $w_i$ for all source patients over all boosting iterations. Those with high weight (above median source weight) are considered to contribute to the overall fit and therefore benefit target predictions while those with low weight (below median source weight) are discarded, since they exhibit the most different survival behaviour.

**What patients are "transferred"? -** As can be seen in Figure 3 we observe a large difference in the empirical marginal survival distributions of auxiliary patients incorporated to contribute to ECHOS patient predictions and of those discarded. TSB selects those auxiliary patients with similar marginal survival distribution while rejects those that exhibit an obviously different survival pattern. Interesting differences are also observed in the respective feature distributions of the three groups: for example, the average heart failure duration in discarded patients was $34s$, much longer than target patients - $26s$ - and included patients - $28s$. Ethnicity was also discovered to play a prominent role in survival behaviour, 97% of ECHOS patients happen to be Caucasian, 88% of included source patients are also Caucasian while only 6% of discarded patients are Caucasian.

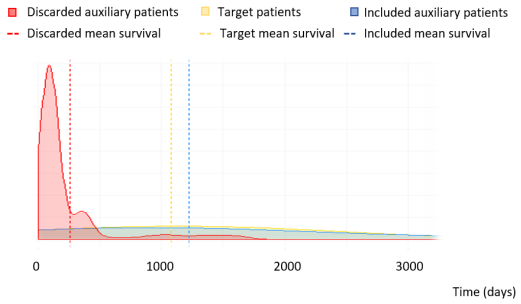


Figure 3: Empirical survival distributions of ECHOS target patients, included auxiliary patients and excluded auxiliary patients.

This experiment shows that TSB is able to discover fine-grained similarities between target and source patients by individually weighting patients and thus identifying relevant source subgroups close in distribution without a priori knowledge of the underlying survival behaviour.

## 5 Conclusion

Developing accurate survival prediction models with only scarce data available is an important challenge to overcome for wider use of machine learning techniques - especially in health care. In this paper we proposed the first survival analysis method that is able to utilize data from multiple heterogeneous auxiliary domains to improve the prediction performance on a target population of interest.

From a clinical perspective this means that clinicians can benefit from accurate decision support mechanisms even when the target population at the local hospital is small. Future work will investigate how to avoid *negative transfer* - a situation where the addition of auxiliary data harms the performance on a target population - in an effort to make transfer learning more reliable.

**Acknowledgements**

This research is supported by the Office of Naval Re-

search (ONR), the NSF (Grant number: ECCS1462245, ECCS1533983, and ECCS1407712) and the Alan Turing Institute under the EPSRC grant EP/N510129/1.

## References

Al-Stouhi, S., and Reddy, C. K. 2011. Adaptive boosting for transfer learning using dynamic updates. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, 60–75. Springer.

Bellot, A., and van der Schaar, M. 2018. Boosted trees for risk prognosis. In *Machine Learning for Healthcare Conference*.

Caruana, R. 1997. Multitask learning. *Machine learning* 28(1):41–75.

Cox, D. R. 1972. Regression models and life tables (with discussion). *Journal of the Royal Statistical Society. Series B* 34:187–220.

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.

Ishwaran, H.; Kogalur, U. B.; Blackstone, E. H.; and Lauer, M. S. 2008. Random survival forests. *The annals of applied statistics* 841–860.

Kim, T.; Cha, M.; Kim, H.; Lee, J. K.; and Kim, J. 2017. Learning to discover cross-domain relations with generative adversarial networks. *arXiv preprint arXiv:1703.05192*.

LeBlanc, M., and Crowley, J. 1992. Relative risk trees for censored survival data. *Biometrics* 411–425.

Li, Y.; Wang, L.; Wang, J.; Ye, J.; and Reddy, C. K. 2016. Transfer learning for survival analysis via efficient l2, 1-norm regularized cox regression. In *Data Mining (ICDM), 2016 IEEE 16th International Conference on*, 231–240. IEEE.

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

Pardoe, D., and Stone, P. 2010. Boosting for regression transfer. In *Proceedings of the 27th International Conference on International Conference on Machine Learning*, 863–870. Omnipress.

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.

Wainberg, M.; Merico, D.; Delong, A.; and Frey, B. J. 2018. Deep learning in biomedicine. *Nature biotechnology* 36(9):829.

Wiens, J.; Guttag, J.; and Horvitz, E. 2014. A study in transfer learning: leveraging data from multiple hospitals to enhance hospital-specific predictions. *Journal of the American Medical Informatics Association* 21(4):699–706.

Wolbers, M.; Blanche, P.; Koller, M. T.; Witteman, J. C.; and Gerds, T. A. 2014. Concordance for prognostic models with competing risks. *Biostatistics* 15(3):526–539.

Yao, Y., and Doretto, G. 2010. Boosting for transfer learning with multiple sources. In *Computer vision and pattern recognition (CVPR), 2010 IEEE conference on*, 1855–1862. IEEE.

Yoon, J.; Jordon, J.; and van der Schaar, M. 2018. Radialgan: Leveraging multiple datasets to improve target-specific predictive models using generative adversarial networks. In *International Conference on Machine Learning*.