Multi-Source Survival Domain Adaptation

Ammar Shaker, Carolin Lawrence

NEC Laboratories Europe GmbH, Heidelberg, Germany
{Ammar.Shaker,Carolin.Lawrence}@neclab.eu

Abstract

Survival analysis is the branch of statistics that studies the relation between the characteristics of living entities and their respective survival times, taking into account the partial information held by censored cases. A good analysis can, for example, determine whether one medical treatment for a group of patients is better than another. With the rise of machine learning, survival analysis can be modeled as learning a function that maps studied patients to their survival times. To succeed with that, there are three crucial issues to be tackled. First, some patient data is censored: we do not know the true survival times for all patients. Second, data is scarce, which led past research to treat different illness types as domains in a multi-task setup. Third, there is the need for adaptation to new or extremely rare illness types, where little or no labels are available. In contrast to previous multi-task setups, we want to investigate how to efficiently adapt to a new survival target domain from multiple survival source domains. For this, we introduce a new survival metric and the corresponding discrepancy measure between survival distributions. These allow us to define domain adaptation for survival analysis while incorporating censored data, which would otherwise have to be dropped. Our experiments on two cancer data sets reveal a superb performance on target domains, a better treatment recommendation, and a weight matrix with a plausible explanation.

1 Introduction

The abundance of health records has massively increased in the last few decades, mainly due to the advancement of data collection methods and the increasing financial support for medical trials and research. To determine the effects of a specific environment or the success of a treatment, survival analysis can be used to study the relation between the characteristics of living entities and their respective survival times. This induced relation is often described by the survival function or the hazard function, which models the conditional propensity for the event of death to happen.

A crucial challenging characteristic of learning with health records is censoring, which is the case when only partial information about the patient's survival is known. This could happen either due to losing track of the patient or the termination of the study before observing the intended event on all patients. Simply discarding this data would lead to losing all the partial information carried by the censored cases, which would be particularly harmful when censoring is prevalent. This, for example, occurs in the messenger RNA data for breast adenocarcinoma, where censoring exceeds 87%[1].

While censoring makes the direct application of machine learning methods unfeasible, active research tries to tackle this challenge. One essential line of work to tackle this challenge adopts the proportional hazards assumption (PH) (Cox 1972), instead of attempting to fully model the survival function. More recently, traditional survival analysis methods (Cox 1972) have been complemented and then superseded by machine learning approaches; for a survey, see Wang, Li, and Reddy (2019). For example, with the increasing success of deep learning methods, DeepSurv (Katzman et al. 2018) has reported a significant increase in performance by employing a neural network with a loss function adapted to hold the assumed proportionality of hazards.

An additional challenge arises when there is insufficient data for a particular problem of interest. This scenario is quite relevant in the medical field, where some diseases are more common than others, such as the varying incidence rates of cancer types confirmed by The Cancer Genome Atlas (TCGA) data. The issue is also present for new illnesses that arise, such as a significantly changed variant of a previous disease. In such a setup, a fitting machine learning technique would be multi-source domain adaptation (Mansour, Mohri, and Rostamizadeh 2008), which tries to exploit the knowledge-transfer from multiple source domains into the target domain. To the best of our knowledge, there has not been yet any work that tackles domain or multi-source domain adaptation for survival domains.

In this work, we introduce a first attempt to transfer knowledge from multiple source survival domains to a target survival domain. Our main contributions are summarized as follows:

- We construct the symmetric discordance index (SDI) to measure the distance between risk functions. We show the utility of SDI in the survival domain adaptation in which multiple survival source tasks are observed

[1] https://www.cancer.gov/about-nci/organization/ccg/research/structural-genomics/tcga

(Section 3.1). Second, we introduce the survival domain discrepancy distance $D_{SDI-disc}(P_s, P_t)$ to measure the proximity between distributions (P_s, P_t) with respect to hypothesis space $\mathcal{H}$ (Section 3.2).

- We derive an error generalization bound for survival target domains (Section 3.2) and we employ this bound in an adversarial min-max optimization problem objective (Section 3.3).
- We show empirically on two TCGA data sets the utility of our method in both the unsupervised and the partially supervised settings. We also show that our approach facilitates treatment recommendation that is in 66% of the cases better than the administered treatment. Additionally, we learn a weight matrix that discovers relations between the different cancer types (Section 4).

2 Background: Survival Analysis

2.1 Preliminaries

Survival analysis methods aim at learning the relation between features of individuals and their corresponding survival times (time-to-event). We use the term *instance* instead of *individual* since studied subjects could be humans, animals, or even mechanical parts. Typically, survival data take the form $D = \{(\boldsymbol{x}_i, t_i, \delta_i) | i \in \{1, \ldots, n\}\}$, where n is the number of instances, $\boldsymbol{x}_i \in \mathbb{R}^d$ is a vector of covariates, t_i is either the observation time of the event or the censoring time and δ_i is an event indicator that reveals the status of censoring, i.e., $\delta_i = 0$ for censored cases and $\delta_i = 1$ otherwise. Censoring occurs when the target event is not observed before the termination of the study; thus, we acquire only the partial information about surviving at least till t_i. We consider only right-censoring in which the actual survival time of a censored instance is after the time of the last observation, i.e., censoring time.

2.2 Survival Functions

The time-to-event t is a random variable that can be characterized by three functions: *(i)* the probability density function, *(ii)* the survival function, and *(iii)* the hazard function. Knowing any of these functions leads to deriving the other two. Given the random variable T, time-to-event, the density function models the probability for the event to occur in infinitesimal interval $[t, t + \Delta t]$, i.e., $f(t) = \lim_{\Delta t \to 0} \frac{\boldsymbol{P}\{t < T \leq t + \Delta t\}}{\Delta t}$. The survival function, $S(\cdot)$, models the probability of surviving till time t: $S(t) = \boldsymbol{P}\{T > t\} = 1 - F(t) = \int_t^\infty f(x)\, dx$, where $F(\cdot)$ is the cumulative distribution function. The conditional probability for the event to occur in the interval $[t, t+\Delta t]$, provided it has not occurred before t, is called the hazard function, $\lambda(\cdot)$; $\lambda(t) = \lim_{\Delta t \to 0} \frac{\boldsymbol{P}\{t < T \leq t + \Delta t | T > t\}}{\Delta t} = \frac{f(t)}{S(t)}$. Both $f(\cdot)$ and $\lambda(\cdot)$ can be derived from $S(\cdot)$ as $f(t) = \frac{d}{dt}[1 - S(t)] = -S'(t)$ and $\lambda(t) = \frac{f(t)}{S(t)} = \frac{-S'(t)}{S(t)}$.

Since the interest of the three survival functions is instance-wise, in the remaining of the paper, we extend the notation by adding the vector of the instance's covariates as a parameter, i.e., $f(t; \boldsymbol{x})$, $F(t; \boldsymbol{x})$, $S(t; \boldsymbol{x})$ and $\lambda(t; \boldsymbol{x})$.

The proportional hazards (PH) assumption, which is first introduced in the Cox proportional hazards model (Cox 1972), assumes constant proportionality of hazards between instances over time, i.e., the hazard ratio $HR = \lambda(t; \boldsymbol{x}_1)/\lambda(t; \boldsymbol{x}_2)$ between the instances $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$ is constant. Hence, for an instance $\boldsymbol{x}$, the hazard is the product of the baseline hazard $\lambda_0(t)$ and a time-independent function $r(\boldsymbol{x})$, i.e., $\lambda(t; \boldsymbol{x}) = \lambda_0(t) \cdot r(\boldsymbol{x})$. The Cox PH model assumes that $r(\boldsymbol{x})$ is a log-linear function of $\boldsymbol{x}$:

$$\lambda(t; \boldsymbol{x}) = \lambda_0(t) \cdot \exp\left(\sum_{i=1}^{n} \beta_i \cdot x_i\right) \quad , \tag{1}$$

where $\lambda_0(t)$ is the hazard when all covariates are set to zero (Lee 1992). The coefficients β_i are found by maximizing the log of the so-called partial likelihood (PL); this likelihood depends on ordering events instead of their joint probabilities. PL computes the event's conditional probability only for non-censored instances, given their risk set, which contains the surviving instances so far.

2.3 Performance Measures

To estimate the performance of a fitted survival model, evaluation measures compute the agreement between the rank of the predicted survivals and the actual survival times. The concordance, also known as the C-index, (Harrell et al. 1982; Harrell Jr et al. 1984) measures how well a risk model ranks instances according to their estimated hazards, survivals, or predicted death times. To this end, it considers each pair of instances and checks if the model's prediction ranks the two instances in accordance with their true order of events. Each non-censored instance is compared against all instances that outlive it (having a larger event or censoring time). Each correct ranking of pairs is counted as 1, and the final score is normalized over the total number of valid pairs. For example, if the Cox PH model, Eq. (1), is used, the C-index takes the form

$$\text{C-index}(r; D) = \frac{1}{Z} \sum_{\substack{(\boldsymbol{x}_i, t_i, \delta_i) \in D \\ \wedge \delta_i = 1}} \sum_{\substack{(\boldsymbol{x}_j, t_j, \delta_j) \in D \\ \wedge t_j > t_i}} I[r(\boldsymbol{x}_i) > r(\boldsymbol{x}_j)], \tag{2}$$

where $Z = \sum_{\substack{(\boldsymbol{x}_i, t_i, \delta_i) \in D \\ \wedge \delta_i = 1}} \sum_{\substack{(\boldsymbol{x}_j, t_j, \delta_j) \in D \\ \wedge t_j > t_i}} 1$, $\{(\boldsymbol{x}_j, t_j, \delta_j) \in D | t_j > t_i\}$ is the risk set of the instance $(\boldsymbol{x}_i, t_i, \delta_i)$, $r(\cdot)$ is the time-independent risk function, and $I[\cdot]$ is the indicator function. Eq. (2) becomes the area under the curve (AUC) when the event times are replaced with the binary problem (event, no event) with no censoring cases; see Haider et al. (2020). Alternatively, the loss based on discordance can be computed as in D-index$(r; D) = 1 -$ C-index$(r; D)$.

3 Multi-Source Survival Domain Adaptation

For survival instances $(\boldsymbol{x}_i, t_i, \delta_i)$, let $\mathcal{X} = \mathbb{R}^d$ and $\mathcal{Y} = \mathbb{R}^+$ be the spaces of the input's covariates and the event time, as described earlier. Let $\{D_i\}_{i=1}^K$ be K survival source domains characterized by the distributions P_i, and let $\{(\boldsymbol{x}_i^j, y_i^j, \delta_i^j)\}_{j=1}^{N_i}$ be the acquired instances for each domain D_i. Let D_t be the target survival domain for which samples

are described only by their covariates $\boldsymbol{x}_t^j$ and the event indicators δ_t^j, whereas the survival times remain missing, i.e., the instances $\{(\boldsymbol{x}_t^j, ?, \delta_t^j)\}_{j=1}^{N_t}$ are observed from P_t.

Typically, multi-source domain adaptation aims at adapting a model fitted on the source domains to the target domain while minimizing the expected loss on D_t. This work considers multi-source survival domain adaptation (MSSDA) where the true survival times t_t^j are unknown.

3.1 Discordance Loss for Survival Data

The D-index could serve as a loss for risk functions on survival domains; however, it does not enjoy symmetry when the roles of the ground truth and the risk function are exchanged due to censoring cases. We aim to enforce symmetry because it is an essential property for bounding the generalization loss on the target domain. To impose symmetry, we define the symmetric discordance index (SDI) for two risk functions r_1 and r_2, and prove that it is a metric satisfying the triangular inequality.

SDI is composed of two parts *(i)* the disagreements in ranking each pair of non-censored instances, and (ii) the disagreements in ranking pairs of censored and non-censored instances. The weights α_1 and α_2 transform SDI into a convex combination of these two parts. Moreover, SDI's symmetry follows from counting the discordance with censored cases twice, once for each of the risk functions while considering the other as the ground truth.

$$SDI(r_1, r_2; D) = \frac{\alpha_1}{\alpha_1+\alpha_2}\frac{1}{Z}\sum_{\substack{(\boldsymbol{x}_i,t_i,\delta_i),\\(\boldsymbol{x}_j,t_j,\delta_j)\in D_{ev}\\ i<j}} I\Bigg[\Big(\big(r_1(\boldsymbol{x}_i) < r_1(\boldsymbol{x}_j)\big)\wedge \big(r_2(\boldsymbol{x}_i) > r_2(\boldsymbol{x}_j)\big)\Big) \vee \Big(\big(r_1(\boldsymbol{x}_i) > r_1(\boldsymbol{x}_j)\big)\wedge \big(r_2(\boldsymbol{x}_i) < r_2(\boldsymbol{x}_j)\big)\Big)\Bigg] + \frac{\alpha_2}{\alpha_1+\alpha_2}\frac{1}{|D_{ce}|}\sum_{(\boldsymbol{x}_i,t_i,\delta_i)\in D_{ce}}\frac{|C_{r_1,\boldsymbol{x}_i}\triangle C_{r_2,\boldsymbol{x}_i}|}{|C_{r_1,\boldsymbol{x}_i}\cup C_{r_2,\boldsymbol{x}_i}|} \tag{3}$$

$$\text{s.t. } C_{r,\boldsymbol{x}} = \Big\{(\boldsymbol{x}_j,t_j,\delta_j)\in D_{ev} | r(\boldsymbol{x}_j) > r(\boldsymbol{x})\Big\}$$

$$\alpha_1 = \binom{|D_{ev}|}{2},\ \alpha_2 = |D_{ev}|.|D_{ce}|/2,\ Z = \sum_{\substack{(\boldsymbol{x}_i,t_i,\delta_i),\\(\boldsymbol{x}_j,t_j,\delta_j)\in D_{ev}\\ i<j}} 1$$

$$D_{ev} = \{(\boldsymbol{x}_j,t_j,\delta_j)\in D | \delta_j = 1\}, D_{ce} = \{(\boldsymbol{x}_j,t_j,\delta_j)\in D | \delta_j \neq 1\}\ ,$$

where $D_{ev} \subseteq D$ and $D_{ce} \subseteq D$ are the sets of non-censored and censored instances, respectively. $C_{r,\boldsymbol{x}}$ is the set of instances (from D) that are assumed to outlive $\boldsymbol{x}$, according to the risk function r. $\triangle$ is the set symmetric difference (disjunctive union).

Notice that the SDI is equivalent to Kendall's tau distance between two rankings when *(i)* counting 0.5 as a score for ties on the survival times, and *(ii)* no censoring.

Next, we present the formal definition of Kendall tau as a rank distance; thereafter, Theorem 2 proves that SDI is a metric by presenting it as a weighted sum of the Kendall tau and the Jaccard metric.

Definition 1. *Kendall tau (Cicirello 2019): Let $S = \{s_1, \ldots, s_n\}$ be the set of n ordered instances, and let τ_1 and τ_2 be two different permutations of instances in S, such that for each $s_i \in S$, $\tau(s_i)$ gives the ranking of s_i in the permutation τ. Kendall tau distance (Kendall 1948) measures the number of pair-wise interventions needed to make two permutations become the same. Kendall tau between the permutations τ_1 and τ_2 is defined as:*

$$\kappa(\tau_1,\tau_2) = \frac{2K_d}{n*(n-1)} \tag{4}$$

$$K_d = |\{(s_i,s_j)\in S\times S | \, i<j \wedge (((\tau_1(s_i)<\tau_1(s_j))\wedge(\tau_2(s_i)<\tau_2(s_j)))\vee ((\tau_1(s_i)>\tau_1(s_j))\wedge(\tau_2(s_i)>\tau_2(s_j))))\}|\ , \tag{5}$$

where K_d is the number of discordance pairs.

Theorem 2. *Given the survival data $D = \{(\boldsymbol{x}_i, t_i, \delta_i) | i \in \{1,\ldots,n\}\}$ and the risk estimators $r_1, r_2 : \mathbb{R}^d \to \mathbb{R}^+$, the symmetric discordance index SDI (Eq. 3) is a metric.*

The proof of Theorem 2 follows from demonstrating that the SDI is a weighted average of two metrics, Kendall tau, and the Jaccard index. The first term, Kendall tau, is measured over the set of non-censored instances. The second term is the sum of the Jaccard index, for each censored instance, on the two risk sets induced by the ranking function r_1 and r_2, see the proof in the supplementary material. Establishing SDI as a metric implies that it enjoys the triangular inequality. This in turn is a necessary criterion that will allow us to derive a generalization bound for the target domain.

3.2 Generalization Bound for Target Domain

To derive the bound of the loss on the target domain by that of the source domains, we follow Cortes and Mohri (2014). We first define a discordance-based distance $D_{SDI-disc}$ to quantify the discrepancy of two distributions P_s and P_t, over sets from $\mathcal{X}$, based on the loss $SDI : \mathcal{H}\times\mathcal{H}\times\mathcal{X}^N \to [0,1]$, where N is the size of the sets over which the distance between two rankings is measured, and $\mathcal{H}$ is the hypothesis space.

Definition 3. *$D_{SDI-disc}$: The discordance-based distance ($D_{SDI-disc}$) is the largest distance between two domains (concerning the hypothesis space $\mathcal{H}$) in a metric space equipped with the metric SDI as a distance function. Let D_s and D_t be two survival domains with their corresponding distributions P_s and P_t. In survival domains, some samples undergo censoring independent of their features, where the censoring time is bound by the survival time. Each hypothesis in $\mathcal{H}$ is a scoring function that acts as a ranking or a risk function. For the distributions P_s and P_t, and $N \in \mathbb{N}$, $D_{SDI-disc}$ takes the form:*

$$D_{SDI-disc}(P_s,P_t) = \max_{h,h'\in\mathcal{H}} \mathbb{E}_{\substack{M_s=\{x_1,\ldots,x_N\sim P_s\}\\ M_t=\{x_1,\ldots,x_N\sim P_t\}}} \big|SDI(h,h';M_s) - SDI(h,h';M_t)\big|\ , \tag{6}$$

where M_s and M_t are the sets of size N from the source and target domains, respectively.

The discordance distance, $D_{SDI-disc}$ reaches its maximum when two ranking functions $h, h' \in \mathcal{H}$ rank the instances of the survival source domain similarly (high concordance) and differently rank the samples of the target domain (high discordance), or vice-versa. Theorem 4 utilizes

ID	Cancer name	Acr.	Instances #	Instances $\delta = 1$
1	Breast Adenocarcinoma	BRCA	707	90
2	Glioblastoma Multiforme	GBM	275	176
3	Head-Neck Squa. Cell Carci.	HNSC	298	119
4	Kidney Renal Clear Cell Carci.	KIRC	415	136
5	Acute Myeloid Leukaemia	LAML	172	105
6	Lung Adenocarcinoma	LUAD	148	49
7	Lung Squamous Cell Carci.	LUSC	163	68
8	Ovarian Serous Carcinoma	OV	315	181

Table 1: Properties of the mRNA data.

$D_{SDI-disc}$ as a distance between distributions to bound the discordance loss on the target survival domain.

Theorem 4. *Let S be a set of K source survival domains $S = \{D_{s_1}, \ldots, D_{s_K}\}$ with distributions P_{s_i}, and denote the ground truth mapping (risk) function in D_{s_i} as f_{s_i}. Similarly, let D_t be a target survival domain with the corresponding distribution P_t and the true risk function f_t. Assume the following sets: $M_{s_i} = \{x_1, \ldots, x_N | x_j \sim P_{s_i}\}$ and $M_t = \{x_1, \ldots, x_N | x_j \sim P_t\}$ to be sampled, of size N, from the source domains D_{S_i} in S and the target domain D_t, respectively. Also, assume a weighting scheme w_i for the source domain D_{s_i} s.t. $\sum_{i=1}^{K} w_i = 1$. For any hypothesis $h \in \mathcal{H}$, the SDI on the target domain D_t is bound in the following way:*

$$SDI(r_h, f_t; M_t) \leq \eta_D(f_S, f_t) + \sum_{i=1}^{k} w_i \cdot \Big(SDI(r_h, f_{s_i}; M_{s_i}) + D_{SDI-disc}(P_{s_i}, P_t) \Big) \quad (7)$$

where r_h is the risk (or ranking) function induced by h and

$$\eta_D(f_S, f_t) = \min_{h^* \in \mathcal{H}} SDI(r_{h^*}, f_t; M_t) + \sum_{i=1}^{k} w_i \cdot SDI(r_{h^*}, f_{s_i}; M_{s_i})$$

is the minimum joint empirical SDI losses on the sources S and the target D_t, achieved by an optimal hypothesis h^*.

The supplementary material provides the complete proof of Theorem 4. This proof starts by deriving the error bound for a single source domain D_{s_i}. Thanks to the metric properties of SDI, we prove at first that $SDI(r_h, f_t; M_t) \leq SDI(r_h, f_{s_i}; M_{s_i}) + SDI(r_{h^*}, f_t; M_t) + SDI(r_{h^*}, f_{s_i}; M_{s_i}) + D_{SDI-disc}(P_{s_i}, P_t)$. The proof concludes by reweighing and aggregating this inequality for each source domain. The main outcome of Theorem 4 is bounding the symmetric discordance on the target domain by the quantities *i)* the weighted average of the SDI on the survival source domains, *ii)* the weighted mismatch between the target D_t and each of D_{S_i} in terms of the discordance-based distance ($D_{SDI-disc}$), and *iii)* the minimum joint empirical SDI losses on the source and target domains. Based on this result, next, we design an optimization objective for survival domain adaptation.

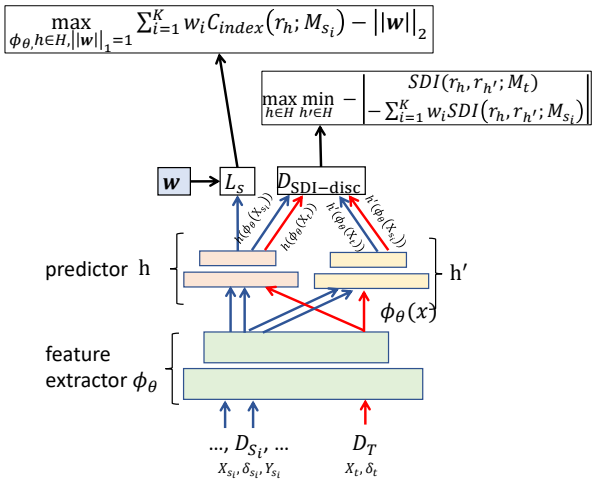


Figure 1: An illustration of how symmetric discordance index (SDI) is employed in our multi-source survival domain adaptation method, MSSDA. The objective includes three terms: 1) the first term enforces the ranking function r_h, to be a good ranker, in terms of C-index, on all source domains; and 2) the second term is an explicit realization of the weighted discordance-based distance ($w_i D_{SDI-disc}(P_t, P_{S_i})$); and 3) the third term is a regularization on the learned weights vector, $\boldsymbol{w}$, that specifies the weight for each source domain concerning the target domain.

3.3 Optimization Problem of Multi-Source Survival Domain Adaptation

We exploit the bound derived in Theorem 4 to enforce distribution matching through an adversarial min-max optimization objective, following domain-adversarial neural networks (DANN) (Ganin, Ustinova et al. 2016). To this end, we search in the hypothesis space $\mathcal{H}$, where each $h \in \mathcal{H}$ defines a risk function r_h, the time-independent function in the hazard Eq. (1). Thus, keeping the proportional hazards assumption. Formally, the hypotheses in $\mathcal{H}$ take the form $h : \mathcal{V} \to \mathbb{R}$, where $\mathcal{V}$ is the feature space. We also search for the feature extractor $\phi_\theta : \mathcal{X} \to \mathcal{V}$, and the weighting $\boldsymbol{w}$ of the source domains, such that:

$$\max_{\substack{\phi_\theta, h \in \mathcal{H} \\ ||\boldsymbol{w}||_1 = 1}} \min_{h' \in \mathcal{H}} \Bigg(\sum_{i=1}^{K} w_i \text{C-index}(r_h; M_{s_i}) - \lambda_1 \left| SDI(r_h, r_{h'}; M_t) - \sum_{i=1}^{K} w_i SDI(r_h, r_{h'}; M_{s_i}) \right| - \lambda_2 ||\boldsymbol{w}||_2 \Bigg) , \quad (8)$$

where r_h and $r_{h'}$ are the ranking functions induced by the hypotheses h and h' respectively. The first term of Eq. (8) enforces the ranking function r_h, to be a good ranker, in terms of C-index, on all source domains; this term is realized by minimizing the negative log-partial likelihood. The second term is an explicit realization of the weighted discordance-based distance ($w_i D_{SDI-disc}(P_t, P_{S_i})$). The third term is a regularization on the learned weights vector,

ID	Acr.	Instances		Pharma.		Rad.		TR
		#	$\delta = 1$	#	$\delta = 1$	#	$\delta = 1$	
1	ACC	80	29	34	13	2	0	
2	BLCA	407	178	111	40	25	15	X
3	BRCA	754	105	238	15	31	2	X
4	CESC	307	72	4	2	35	9	
5	CHOL	36	18	13	7	1	1	
6	ESCA	184	77	12	5	22	5	X
7	HNSC	484	203	6	2	134	46	
8	KIRC	254	76	24	16	7	3	
9	KIRP	290	44	18	13	4	4	
10	LGG	510	124	50	10	70	16	X
11	LIHC	371	128	39	18	12	4	X
12	LUAD	441	157	103	36	34	24	X
13	LUSC	338	137	72	20	19	13	X
14	MESO	86	73	29	26	2	1	
15	PAAD	178	93	75	41	-	-	
16	SARC	259	98	46	22	48	15	X
17	SKCM	97	26	22	9	1	0	
18	STAD	382	147	106	42	1	0	
19	UCEC	410	72	55	18	82	10	X
20	UCS	56	34	15	12	5	5	
21	UVM	80	23	11	6	4	2	

Table 2: Properties of the miRNA data. The treatment columns (Pharmaceutical and Radiation) are collected by matching the data with The Cancer Genome Atlas (TCGA). The TR column indicates whether or not the cancer type is used for evaluating the treatment recommendation.

$\boldsymbol{w}$, that specifies the weight for each source domain concerning the target domain. This adversarial min-max game aims at finding, for the survival source and target domains, a feature extractor ϕ_θ and a ranker r_h such that for any other ranker $r_{h'}$, the weighted distance is minimized, i.e., achieving feature invariance of the target domain and each of the source domains (in a weighted manner). We term our method the multi-source survival domain adaptation as (MSSDA), and acknowledge that comparable min-max objectives were used in (Pei et al. 2018; Saito, Kim et al. 2019; Richard et al. 2020; Shaker, Yu, and Onoro-Rubio 2022) outside of survival analysis. Notice that this algorithm does not optimize for η_D since this term is constant for a single source domain and for the mixture of sources in $\eta_D(f_S, f_t)$ given the weighting $\boldsymbol{w}$.

Figure 1 depicts a graphical illustration of the proposed optimization problem; it shows the details of our method, MSSDA. X_{s_i}, δ_{s_i} and Y_{s_i} are the input samples, the censoring indicators, and the survival times from the source domain D_{s_i}; X_t and δ_t are the input samples and the censoring indicators of the target domain without survival times. The hypothesis h is trained to produce a good ranker r_h in terms of the weighted C-index on the sources. The hypothesis h' tries to increase the discordance-based distance ($D_{SDI-disc}$) between the target distribution and the weighted combination of source domains (i.e., $D_{SDI-disc}$).

4 Empirical Evaluation

To investigate the usefulness of our proposed method to adapt to a target survival domain, we address the following three questions:

- Does the multi-source domain adaption work on survival target domains? How does it perform if the labels for a portion of the target data were used? (Section 4.1.)
- Can we recommend treatment better than what was offered to the patients? (Section 4.2.)
- Do the learned weights on the source domains reveal any useful information about the underlying cancer types? (Section 4.3.)
- How essential is the proposed symmetric discordance index (SDI) for aligning the conditional distributions compared to other domain-invariant regularisation approaches? (Section 4.4.)

Datasets. We utilize two data sets from The Cancer Genome Atlas project (TCGA)[2]. This project analyzes the molecular profiles and the clinical data of 33 cancer types. *(i)* The Messenger RNA data (mRNA) (Li et al. 2016b), which includes eight cancer types. Each patient is represented by 19171 binary features; see Table 1. *(ii)* The micro-RNA data (miRNA) that includes 21 cancer types (Wang et al. 2017); each has a varying number of patients. Table 2 depicts the total number of patients for each cancer and the number of patients that experienced the event (died) during the time of the clinical study ($\delta = 1$). We also extract the treatment performed for each cancer type (if available).

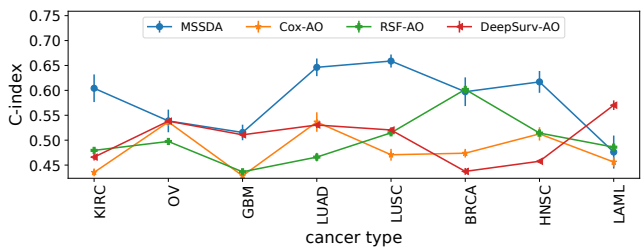
Baselines. For the evaluation, we compare with 1) the Cox proportional hazards model fitted by maximizing the log of the partial likelihood, 2) DeepSurv (Katzman et al. 2018) that introduces the proportional hazards to neural networks, and 3) the survival random forests (RSF) (Ishwaran and Kogalur 2007)[3]. These methods deal with single domains; therefore, we perform separate training on each source domain and use the trained model as a ranking function for the target domain. Each ranker orders the target's instances; we average these orders over all rankers, hence, the abbreviation "average order" (AO). To answer the second part of the first question, we compare with 4) TransferCox (Li et al. 2016b), a transfer learning method that employs multi-task learning on survival domains and requires labels in all domains without prioritizing the target domain.

For both MSSDA[4] and DeepSurv, we use the same architecture, a two-layered feature extractor with 200 and 20 units in the first and second hidden layers, respectively. The detailed architecture and the hyper-parameters search are explained in the supplementary material. We model the log-risk function as the non-linear function $h \circ \phi_\theta(x)$ learned by the fitted network architecture, i.e., $r(x) = e^{h \circ \phi_\theta(x)}$. MSSDA and DeepSurv are trained for 20 epochs.

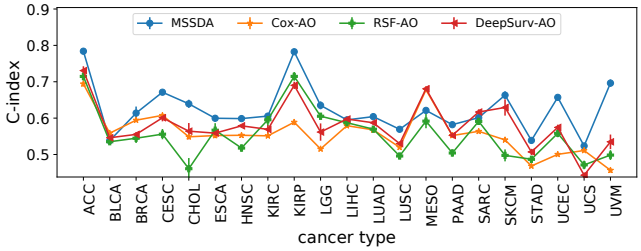
[2]https://www.cancer.gov/about-nci/organization/ccg/research/structural-genomics/tcga

[3]https://square.github.io/pysurvival/

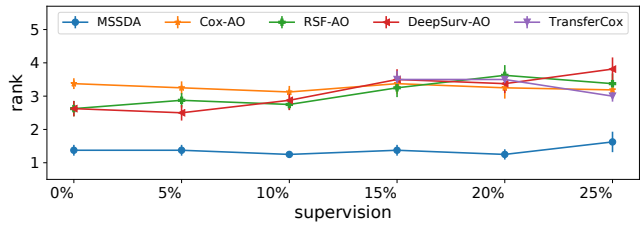
[4]https://github.com/shaker82/MSSDA



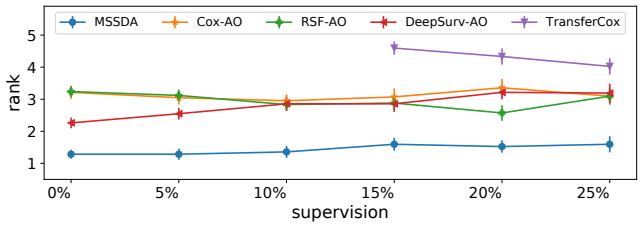
(a) C-index on mRNA with no supervision.



(b) C-index on miRNA with no supervision.



(c) Rank of the different methods based on the C-index on mRNA.



(d) Rank of the different methods based on the C-index on miRNA.

Figure 2: Performance comparison and ranking, on the miRNA and mRNA data, in terms of the C-index.

In the supervised target case, we allow a small portion of the target domain to be labeled and used for training. We use these percentages, 5%, 10%, 15%, 20%, and 25%. Except for the TransferCox, the target samples are appended to each source domain's samples. For TransferCox, the target samples are added as a new domain, which is why TransferCox can not be tested when the target domain contains very few samples (less than 20%).

Evaluation. To measure the performance of each method, we employ the C-index, Eq.(2), to measure the concordance between the inversely ordered predicted risks and the actual lifetime. For the unsupervised and supervised cases, we measure the C-index on the target domain's samples after removing the ones used for the supervision. We also propose C-index$'$ that measures the concordance on the whole target domain, including the samples used for supervision. Notice that C-index$'$ includes only a tiny portion of the pairs used for the training. In the case of 25% supervision, we show in the supplementary material the advantage of measuring the C-index$'$ and that the ratio of reused pairs is only 6.25%. All results are averaged over five folds.

Optimizing the SDI The counting-based comparison in the SDI is implemented using the MarginRankingLoss MRS[5]: $MRS(x1, x2, y) = max(0, -y(x1 - x2) + m)$, where m is the margin. For example, we implement $I\Big((r(\boldsymbol{x}_i) < r(\boldsymbol{x}_j)\Big)$ in Eq. (3) using the surrogate $MRS(exp(r(\boldsymbol{x}_i) - r(\boldsymbol{x}_j)), 0, 1)$ with $m = 1$.

4.1 Evaluation of Survival Prediction

Table 3 depicts the performance in terms of the C-index on the mRNA data; it shows that MSSDA outperforms all other methods in both the unsupervised and the partially supervised (5%) cases while always achieving the first rank (the last row). In the supplementary material, results show that MSSDA still dominates the remaining supervision settings at 10%, 15%, 20%, and 25%; these results are graphically depicted in Figures 2a and 2c and confirm the superiority of MSSDA performance and its first rank. In general, MSSDA performs best on five of seven cancer types, achieving the best rank, followed by RSF in low supervision and TransferCox in high supervision settings. A similar performance is evident when considering the C-index$'$, as confirmed in tables shown in the supplementary material.

Similarly, MSSDA achieves a superb performance on miRNA on supervision and no supervision settings, as confirmed by the dominating C-index in Figure 2b, and the best rank in Figure 2d. These figures summarize tables and figures that are kept in the supplementary material. Again MSSDA performs best on 17, 18, 18, 15, 15, 16 out of 21 cancer types for the 0%, 5%, 10% 15%, 20% and 25% supervision settings, respectively. Hence, MSSDA always ranks first.

The supplementary material includes figures that depict the rank when using C-index$'$, and a deeper discussion on the discrepancy between the result of C-index and C-index$'$.

We compute the p-value for the upper-tailed Wilcoxon signed-ranks test between each method and MDSSA in each setting on both data sets. The null hypothesis can be rejected on all mRNA data at the significance level of $\alpha = 0.05$ for both performance measures. On the miRNA, the null hypothesis can be rejected in all cases at $\alpha = 0.1$ except for the two cases (RSF, 20% supervision, C-index) and (TransferCox, 25% supervision, C-index$'$).

[5]https://pytorch.org/docs/stable/generated/torch.nn.MarginRankingLoss.html

superv.	.00%				5%			
method	MSSDA	Cox-AO	RSF-AO	DeepSurv -AO	MSSDA	Cox-AO	RSF-AO	DeepSurv -AO
KIRC	**.604**	.435	.480	.466	**.618**	.444	.521	.479
±	.028	.008	.002	.007	.019	.007	.004	.004
OV	**.539**	.537	.497	.538	**.563**	.523	.490	.552
±	.023	.008	.002	.002	.015	.013	.006	.002
GBM	**.516**	.428	.436	.511	.493	.443	.494	**.515**
±	.016	.004	.002	.006	.017	.009	.008	.007
LUAD	**.646**	.536	.466	.530	**.661**	.556	.466	.507
±	.018	.020	.008	.006	.014	.020	.015	.005
LUSC	**.659**	.471	.515	.520	**.658**	.533	.494	.508
±	.013	.012	.005	.005	.024	.007	.029	.006
BRCA	.597	.474	**.602**	.437	**.599**	.492	.536	.418
±	.029	.009	.004	.007	.033	.019	.017	.009
HNSC	**.617**	.513	.514	.458	**.646**	.430	.504	.441
±	.022	.013	.003	.003	.017	.009	.016	.005
LAML	.476	.456	.486	**.570**	.533	.451	.514	**.551**
±	.033	.010	.002	.010	.024	.003	.014	.008
P-value		.005	.02	.02		.003	.009	.009
Rank	**1.38**	3.38	2.63	2.63	**1.38**	3.25	2.88	2.50

Table 3: The performance comparison on the eight cancer types in the mRNA data in terms of C-index. The following settings were used: no supervision and 5% supervision. The numbers in brackets depict the standard error. The last row shows the rank in each supervision group. The p-value row depicts the p-value for the upper-tailed Wilcoxon signed-ranks test between each method and MDSSA. The null hypothesis can be rejected at the significance level of 0.05.

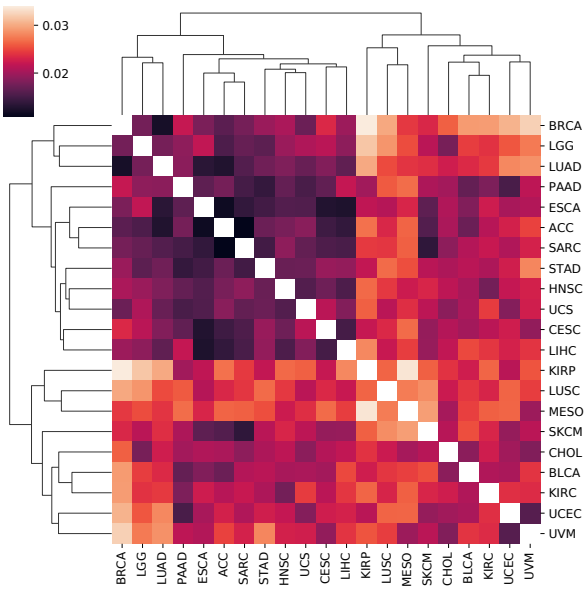


Figure 3: Heatmap of the matrix computed from the learned weights' distances on the miRNA data.

4.2 Treatment Recommendation

For the treatment recommendation experiments, we collect the type of treatment (either pharmaceutical (P) or radiation (R)) given for each of the patients (if available). Table 2 depicts the number of cases for each treatment type. We finally select only those cancer types with at least two non-censored and one censored patient for each treatment type, as indicated in column "TR"; see Table 2. Following the procedure proposed in DeepSurv (Katzman et al. 2018), we annotate each instance from the source domains by a dummy binary attribute identifying the type of treatment (P or R). After learning on the source domains, for each sample $\boldsymbol{x}$ from the target domain, we measure the recommendation $rec(x)_{PR} = log\frac{r(x^P)}{r(x^R)} = h \circ \phi_\theta(x^P) - h \circ \phi_\theta(x^R)$, where $\boldsymbol{x}^P$ and $\boldsymbol{x}^R$ are the same target domain sample once considered to be treated by pharmaceutical and once by radiation, respectively. A positive $rec(x)_{PR}$ means that the patient has a higher risk when treated by "P" than when treated by "R". Hence, it is recommended to prescribe "R". By comparing with the ground truth treatments, we group the patients into the Υ_{recom} group when the recommended treatment aligns with the true treatment and the Υ_{anti} group containing patients that received a recommendation contradicting the actual treatment. Thereafter, the median survival times of the two groups are compared.

A smaller median survival time of the Υ_{anti} indicates that the patients could have had a potentially longer survival time had they been given the model's recommended treatment. MSSDA is employed as in the previous experiments using the labeled multi-source domains and an unlabelled target domain. For DeepSurv, we train a different model for each source domain and then compute the average order (rank) of each sample of the target domain for each treatment. On DeepSurv, the groups Υ_{recom} and Υ_{anti} are computed using the difference in the predicted patients' ranks for the two treatments. Cox model is omitted since it recommends the same treatment for all instances, as proven in (Katzman et al. 2018). TransferCox is also omitted since it requires labeled target data.

Part[A] of Table 4 shows that in the case of contradiction with the administered treatment, our method in 64.4% and 66% of the cases gives a better recommendation for the radiation and pharmaceutical treatments, compared to 50% and 46.6% achieved by DeepSurv. This result is computed over five folds for each treatment and each cancer type. Part[B] of Table 4 shows a detailed comparison of the median survival time for each treatment and cancer pair when merging the samples of all folds. The medians are struck through upon equality and compared otherwise. Again, the results show a median survival time in the Υ_{anti} group smaller than that of the Υ_{recom} group in 5/5 and 5/8 of the cases when MSSDA is employed. Whereas, Deepsurv achieves this only on 2/5 and 3/8 of the cases. RSF fails experimentally to identify and employ the treatment indicator, which has led to failing to induce two different risk models for the different treatments. Therefore, RSF is omitted in the experiment.

	[A] Success rate on 5 folds				[B] Median time of all folds merged											
	MSSDA		DeepSurv-AO		MSSDA						DeepSurv-AO					
	Rad.	Pharma.	Rad.	Pharma.	Rad.				Pharma.				Rad.				Pharma.			
Name	success ratio	success ratio	success ratio	success ratio	median recomm.		median anti-recom.		median recomm.		median anti-recom.		median recomm.		median anti-recom.		median recomm.		median anti-recom.	
BLCA	2/5	2 / 5	4/5	2/5	~~370~~		~~370~~		536.5	$\not\geq$	547		651	$\geq$	324		539	$\not\geq$	544	
BRCA	1/2	4 / 5	3/3	3/5	~~1330.5~~		~~1330.5~~		~~1032~~		~~1032~~		2296	$\geq$	365		~~1032~~		~~1032~~	
ESCA	2/4	4 / 4	3/5	3/5	~~283~~		~~283~~		496	$\geq$	480		~~283~~		~~283~~		496	$\geq$	480	
LGG	4/5	3 / 5	1/5	4/5	1368	$\geq$	794		1106	$\geq$	933		1011	$\not\geq$	1335		1106	$\geq$	758	
LIHC	5/5	4 / 5	1/4	1/5	643	$\geq$	432		639	$\geq$	633		432	$\not\geq$	643		612	$\not\geq$	639	
LUAD	4/5	3 / 5	2/5	2/5	677	$\geq$	561		574	$\not\geq$	594		~~633~~		~~633~~		503	$\not\geq$	594	
LUSC	2/5	2 / 5	2/5	1/5	387	$\geq$	345		559	$\not\geq$	562		~~387~~		~~387~~		559	$\not\geq$	573	
SARC	3/5	5 / 5	3/5	2/5	~~695~~		~~695~~		1013.5	$\geq$	550		~~695~~		~~695~~		591	$\not\geq$	599	
UCEC	4/5	2 / 5	1/5	3/5	1279	$\geq$	1127		666	$\geq$	610		1016	$\not\geq$	1317		670	$\geq$	610	
$\sum$	**5.8/9**	**6/9**	4.5/9	4.2/9	**5/5**				**5/8**				2/5				3/8			

Table 4: Results of the treatment recommendation experiments on the miRNA data. Table A, on five folds, depicts the ratios of better-recommended treatments over the valid folds. Table B presents the median survival times of the recommendation and anti-recommendation groups across all folds after removing the censored patients.

cancer	MSSDA	MDAN	KuiperUB -KM	MMD	MMD-KM	D-index
ACC	**.784**	.729	.690	.688	.682	.702
BLCA	**.538**	.507	.505	.513	.507	.520
BRCA	**.614**	.560	.538	.542	.534	.562
CESC	**.671**	.620	.615	.606	.617	.632
CHOL	**.639**	.553	.579	.598	.582	.582
ESCA	**.600**	.557	.555	.561	.553	.589
HNSC	**.599**	.579	.561	.564	.560	.576
KIRC	**.606**	.571	.575	.583	.576	.588
KIRP	**.782**	.707	.677	.669	.676	.691
LGG	**.635**	.566	.553	.542	.554	.565
LIHC	**.595**	.554	.546	.554	.542	.561
LUAD	**.604**	.566	.547	.544	.545	.559
LUSC	**.569**	.554	.539	.538	.537	.554
MESO	.621	**.632**	.600	.588	.595	.610
PAAD	**.582**	.568	.555	.553	.553	.570
SARC	**.601**	.573	.571	.573	.571	.583
SKCM	**.663**	.595	.551	.533	.545	.566
STAD	.539	.515	.527	.531	.526	**.543**
UCEC	**.657**	.547	.529	.531	.526	.548
UCS	**.524**	.496	.496	.493	.494	.504
UVM	**.696**	.537	.551	.534	.553	.586
Rank	**1.1**	3.19	4.57	4.62	5.1	2.43

Table 5: The performance comparison in the ablation analysis on the miRNA data in terms of C-index. The numbers in brackets depict the standard error. The last row shows the rank of each method.

4.3 Explanation of Learned Weights

Finally, we would like to investigate if our method has learned any meaningful relations between the different cancer types. Therefore, we compute the matrix of pair-wise Euclidean distance between each pair of cancer types i and j by removing the i-th and j-th entries from their learned weight vectors. After that, we perform hierarchical clustering on the computed matrix, as shown in Figure 3. We notice two major groups of cancer types in the resulting clustering. Following the classification of the solid tumor types in Hoadley et al. (2018), the figure shows closeness in the hierarchical clustering between cancers from the same solid tumor types. For example, for the urologic type, we find that BLCA and KIRC are clustered together, and KIRP also belongs to the same major group that contains BLCA and KIRC. We observe the same for the thoracic type (LUSC and MESO). ESCA and STAD, from the core gastrointestinal type, are within a small distance. The same applies to the types: cancers affecting melanocytes in skin and eye (UVM and SKCM) and soft tissues (SARC and UCS). We find a weaker confirmation for the gynecologic types where BRCA and CESC are in the same major cluster. The same can be observed for cancers in the developmental gastrointestinal type (LICH and PAAD).

Moreover, we found overlaps and similarities when comparing with the unsupervised clustering performed in Hoadley et al. on the DNA methylation. For example, HNSC, CESC, and ESCA were clustered within small proximity by MSSDA and belong to the same clusters (METH2 and MET3) by Hoadley et al. (2018). The same observation can be made for ESCA and STAD that we find to be within a small distance and belong to the same branch of clusters.

Our observations are of high importance since our system learned the relations between the cancer types by only fitting the risk functions of unlabeled targets and not directly from

the data as in (Hoadley et al. 2018).

4.4 Ablation Analysis

In this section, we perform an ablation study by replacing the proposed (SDI) with the following domain-invariant distances and regularizers: *(i)* MDAN, a domain classifier as proposed in (Fernandez and Gretton 2019), *(ii)* KuiperUB-KM which tightens the upper bound of the p-value of the two-sample Kuiper test (Kuiper 1960) that is applied on the Kaplan–Meier (KM) curve (Kaplan and Meier 1958a), *(iii)* MMD, the maximum mean discrepancy (Gretton et al. 2006) (which does not take censoring into consideration), *(iv)* MMD-KM, the maximum mean discrepancy on the KM curve, and *(v)* the D-index. Fernandez and Gretton (2019) propose an adaptation to the maximum mean discrepancy (MMD) for data with censored cases. We couldn't compare with this distance since it is a one-sample test against the uniform distribution.

Results in Table 5 show the superiority and benefit of SDI over the other methods in forcing the representation's conditional invariance. This is mainly because SDI takes censoring into consideration (which is ignored by MDAN and MMD), aligns the conditional distributions (which is ignored also by MDAN and MMD), and guarantees symmetry (symmetry is not guaranteed in KuiperUB-KM and D-index).

5 Related Work

Multi-Source Domain Adaptation (MSDA) Ben-David et al. (2007) define the distance d_A between two distributions and prove a VC dimension-based generalization bound for domain adaptation in binary tasks. Mansour, Mohri, and Rostamizadeh (2009) generalized this bound further to a broader set of problems and used it in a tighter bound with the Rademacher complexity. Ben-David et al. (2010) introduce the $H\Delta H$ as a discrepancy measure between distributions and show how to approximate it merely from a finite sample of unlabeled target data. Cortes and Mohri (2014) define the discrepancy measure D_{disc} between distributions regardless of the true labeling function and present an algorithm for adaptation using Discrepancy minimization. Most MSDA methods employ bounds based on these seminal works; for example, domain adversarial neural network (DANN) (Ganin, Ustinova et al. 2016) performs distribution matching by a min-max game; this work was extended to the multiple domains in MDAN (Zhao et al. 2017). Li et al. (2016b) show a tighter bound using a Wasserstein-like distance extending the $H\Delta H$ divergence. Richard et al. (2020) employ the D_{disc} for regression target domains. Shaker, Yu, and Onoro-Rubio (2022) propose to align the conditional distributions in the multi-source domain adaptation setting using a symmetric form of the conditional von Neumann divergence (Yu et al. 2020). Our proposed method can be interpreted as aligning the conditional distributions in the feature space while conditioning on the rankings in the output space.

Machine Learning for Survival Analysis While the non-parametric methods, such as the Kaplan-Meier (KM) estimator (Kaplan and Meier 1958b), can be efficient for moderate data volumes, they have a major limitation in relating the survival function to the covariates. Cox proportional hazards models (Cox 1972; Cox and Oakes 1984) assume the proportionality of hazards between instances and model the risk by a log-linear function of the instance's covariates. A broad spectrum of machine learning methods has been adapted to deal with the challenge of censoring. Ridge-Cox (Tibshirani 1997) and lasso-Cox (Verweij and Van Houwelingen 1994) add l_1 and l_2 regularization terms to the original Cox model, respectively. Wang, Li, and Reddy (2019) reveal a recent survey on the intersection between survival analysis and machine learning research. Survival random forest (RSF) adopts ensemble learning to cope with censored cases (Ishwaran and Kogalur 2007; Ishwaran et al. 2008), and Khan and Zubek (2008) introduce support vector regression for censored data (SVRC). In (Shaker and Hüllermeier 2014; Krempl et al. 2014), the authors propose the continuous and adaptive learning of parallel hazard functions in non-stationary environments under the instantaneous PH assumption, whereas, Lee et al. (2018) deal with learning time-variant survival functions while allowing multiple events and risks per patient, thus, relaxing the PH assumption. Knowledge-transfer between survival models has been the focus of transfer (Li et al. 2016b) and multi-task learning (Li et al. 2016a; Wang et al. 2017) for survival analysis. DeepSurv (Katzman et al. 2018) implement the PH assumption using a deep neural network. The work in (Mouli et al. 2019) defines a clustering objective over survival distributions of samples by tightening the upper bound of the p-value of the two-sample Kuiper test (Kuiper 1960). In (Nagpal et al. 2021), individual survival distributions are fit as a mixture of Cox regression functions. Despite these advancements in research, there is still the need for methods that perform adaptation between survival domains. This work is the first attempt to fill this gap.

6 Conclusion

We presented multi-source survival domain adaptation (MSSDA), which is, to the best of our knowledge, the first multi-source domain adaptation work for survival domains. Adapting to a particular target survival domain is essential for rare or new illness types. In survival analysis, we are faced with the additional difficulty of censored data. To not lose this partial information about survival, we define a new symmetric index for survival data that can handle censored data, show that it is a metric, and use it to bound the generalization error on target domains. This bound is explicitly employed in our method MSSDA. We confirm in experimental results that: (1) our method outperforms existing methods on target survival domains in terms of survival ranking; (2) it can offer better treatment recommendations; (3) it allows us to inspect how different domains relate, offering medical professionals additional insights. We hope our method can aid in identifying better treatments for rare or new illnesses. In the future, we hope to extend our method so that medical professionals can better understand its predictions to improve precision medicine for individuals.

Ethics Statement

Our work aims to introduce domain adaptation to the field of survival data. This line of work can have a positive impact on saving human life by providing precision medicine in the form of personalized treatment recommendations and a better understanding of how diseases could be related and correlated with each other. However, our method is still in the research stage. Therefore, we do not recommend its use in a medical setting without first extensively verifying that a learned model performs as expected. Ultimately all medical decisions should remain in the hands of a medical professional, who is better qualified to judge whether an AI model's prediction should be followed or not.

In addition, domain adaptation, with knowledge transfer, helps learn from fewer data. This positively affects the environment by reducing computational power and run-time to train models, hence, less electricity consumption and less CO_2 emissions.

Acknowledgments

We thank Brandon Malone and Anja Moesch for their feedback on the paper and the insightful discussions.

References

Ben-David, S.; Blitzer, J.; Crammer, K.; Kulesza, A.; Pereira, F.; and Vaughan, J. W. 2010. A theory of learning from different domains. *Machine learning*, 79(1): 151–175.

Ben-David, S.; Blitzer, J.; Crammer, K.; and Pereira, F. 2007. Analysis of representations for domain adaptation. In *NeurIPS*, 137–144.

Cicirello, V. A. 2019. Kendall tau sequence distance: Extending Kendall tau from ranks to sequences. *arXiv preprint arXiv:1905.02752*.

Cortes, C.; and Mohri, M. 2014. Domain adaptation and sample bias correction theory and algorithm for regression. *Theoretical Computer Science*, 519: 103–126.

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

Cox, D. R.; and Oakes, D. 1984. *Analysis of Survival Data*. London, UK: Chapman & Hall.

Fernandez, T.; and Gretton, A. 2019. A maximum-mean-discrepancy goodness-of-fit test for censored data. In *The 22nd International Conference on Artificial Intelligence and Statistics*, 2966–2975. PMLR.

Ganin, Y.; Ustinova, E.; et al. 2016. Domain-adversarial training of neural networks. *JMLR*, 17(1): 2096–2030.

Gretton, A.; Borgwardt, K.; Rasch, M.; Schölkopf, B.; and Smola, A. 2006. A kernel method for the two-sample-problem. *Advances in neural information processing systems*, 19.

Haider, H.; Hoehn, B.; Davis, S.; and Greiner, R. 2020. Effective ways to build and evaluate individual survival distributions. *Journal of Machine Learning Research*, 21(85): 1–63.

Harrell, F. E.; Califf, R. M.; Pryor, D. B.; Lee, K. L.; and Rosati, R. A. 1982. Evaluating the yield of medical tests. *Jama*, 247(18): 2543–2546.

Harrell Jr, F. E.; Lee, K. L.; Califf, R. M.; Pryor, D. B.; and Rosati, R. A. 1984. Regression modelling strategies for improved prognostic prediction. *Statistics in medicine*, 3(2): 143–152.

Hoadley, K. A.; Yau, C.; Hinoue, T.; Wolf, D. M.; Lazar, A. J.; Drill, E.; Shen, R.; Taylor, A. M.; Cherniack, A. D.; Thorsson, V.; et al. 2018. Cell-of-origin patterns dominate the molecular classification of 10,000 tumors from 33 types of cancer. *Cell*, 173(2): 291–304.

Ishwaran, H.; and Kogalur, U. B. 2007. Random survival forests for R. *R news*, 7(2): 25–31.

Ishwaran, H.; Kogalur, U. B.; Blackstone, E. H.; and Lauer, M. S. 2008. Random survival forests. *The annals of applied statistics*, 2(3): 841–860.

Kaplan, E. L.; and Meier, P. 1958a. Nonparametric estimation from incomplete observations. *Journal of the American statistical association*, 53(282): 457–481.

Kaplan, E. L.; and Meier, P. 1958b. Nonparametric Estimation from Incomplete Observations. *Journal of the American Statistical Association*, 53(282): 457–481.

Katzman, J. L.; Shaham, U.; Cloninger, A.; Bates, J.; Jiang, T.; and Kluger, Y. 2018. DeepSurv: personalized treatment recommender system using a Cox proportional hazards deep neural network. *BMC medical research methodology*, 18(1): 1–12.

Kendall, M. G. 1948. *Rank correlation methods.* London, UK: Charles Griffin and Co. Ltd.

Khan, F. M.; and Zubek, V. B. 2008. Support vector regression for censored data (SVRc): a novel tool for survival analysis. In *2008 Eighth IEEE International Conference on Data Mining*, 863–868. IEEE.

Krempl, G.; Žliobaite, I.; Brzeziński, D.; Hüllermeier, E.; Last, M.; Lemaire, V.; Noack, T.; Shaker, A.; Sievi, S.; Spiliopoulou, M.; et al. 2014. Open challenges for data stream mining research. *ACM SIGKDD explorations newsletter*, 16(1): 1–10.

Kuiper, N. H. 1960. Tests concerning random points on a circle. In *Nederl. Akad. Wetensch. Proc. Ser. A*, volume 63, 38–47.

Lee, C.; Zame, W.; Yoon, J.; and Van Der Schaar, M. 2018. Deephit: A deep learning approach to survival analysis with competing risks. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32.

Lee, E. T. 1992. *Statistical Methods for Survival Data Analysis*. Hoboken, NJ, USA: John Wiley & Sons, Inc., 2nd edition.

Li, Y.; Wang, J.; Ye, J.; and Reddy, C. K. 2016a. A multi-task learning formulation for survival analysis. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, 1715–1724.

Li, Y.; Wang, L.; Wang, J.; Ye, J.; and Reddy, C. K. 2016b. Transfer learning for survival analysis via efficient l2, 1-norm regularized cox regression. In *2016 IEEE 16th In-*

ternational Conference on Data Mining (ICDM), 231–240. IEEE.

Mansour, Y.; Mohri, M.; and Rostamizadeh, A. 2008. Domain adaptation with multiple sources. *Advances in neural information processing systems*, 21.

Mansour, Y.; Mohri, M.; and Rostamizadeh, A. 2009. Domain adaptation: Learning bounds and algorithms. In *22nd Conference on Learning Theory, COLT 2009*.

Mouli, S. C.; Teixeira, L.; Neville, J.; and Ribeiro, B. 2019. Deep lifetime clustering. *arXiv preprint arXiv:1910.00547*.

Nagpal, C.; Yadlowsky, S.; Rostamzadeh, N.; and Heller, K. 2021. Deep Cox mixtures for survival regression. In *Machine Learning for Healthcare Conference*, 674–708. PMLR.

Pei, Z.; Cao, Z.; Long, M.; and Wang, J. 2018. Multi-adversarial domain adaptation. In *AAAI*, volume 32.

Richard, G.; de Mathelin, A.; Hébrail, G.; Mougeot, M.; and Vayatis, N. 2020. Unsupervised Multi-Source Domain Adaptation for Regression. In *ECML*.

Saito, K.; Kim, K.; et al. 2019. Semi-supervised domain adaptation via minimax entropy. In *IEEE ICCV*, 8050–8058.

Shaker, A.; and Hüllermeier, E. 2014. Survival analysis on data streams: Analyzing temporal events in dynamically changing environments. *International Journal of Applied Mathematics and Computer Science*, 24(1).

Shaker, A.; Yu, S.; and Onoro-Rubio, D. 2022. Learning to Transfer with von Neumann Conditional Divergence. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, 8231–8239.

Tibshirani, R. 1997. The lasso method for variable selection in the Cox model. *Statistics in medicine*, 16(4): 385–395.

Verweij, P. J.; and Van Houwelingen, H. C. 1994. Penalized likelihood in Cox regression. *Statistics in medicine*, 13(23-24): 2427–2436.

Wang, L.; Li, Y.; Zhou, J.; Zhu, D.; and Ye, J. 2017. Multi-task survival analysis. In *2017 IEEE International Conference on Data Mining (ICDM)*, 485–494. IEEE.

Wang, P.; Li, Y.; and Reddy, C. K. 2019. Machine learning for survival analysis: A survey. *ACM Computing Surveys (CSUR)*, 51(6): 1–36.

Yu, S.; Shaker, A.; Alesiani, F.; and Principe, J. C. 2020. Measuring the Discrepancy between Conditional Distributions: Methods, Properties and Applications. In *IJCAI*, 2777–2784.

Zhao, H.; Zhang, S.; Wu, G.; Costeira, J. P.; Moura, J. M.; and Gordon, G. J. 2017. Multiple source domain adaptation with adversarial training of neural networks. *arXiv preprint arXiv:1705.09684*.