Evaluating Domain Generalization for Survival Analysis in Clinical Studies

Florian Pfisterer FLORIAN.PFISTERER@STAT.UNI-MUENCHEN.DE
F.Hoffmann-La Roche AG, Switzerland; Ludwig-Maximilians-Universität München, Germany

Chris Harbron CHRIS.HARBRON@ROCHE.COM
Roche Products Ltd, United Kingdom

Gunther Jansen GUNTHER.JANSEN@ROCHE.COM
F.Hoffmann-La Roche AG, Switzerland

Tao Xu TAO.XU.TX1@ROCHE.COM
F.Hoffmann-La Roche AG, Switzerland

Abstract

Machine learning models are often required to generalize to new populations (domains) unseen during training, which may lead to model underperformance. So far, most research has focused on Domain Generalization methods for image classification tasks, which address the problem by learning domain invariant predictors. In this study, we assess the efficacy of domain generalization methods in survival analysis. The goal is to predict time-to-events such as death or disease progression based on baseline demographic and clinical variables of individuals exposed to medical treatment. We benchmark four domain generalization methods and several conventional/established methods on real world scenarios encountered in clinical practice. This includes tasks such as generalizing between randomized controlled trials to real world data, identification of prognostic models regardless of treatment or disease subtypes. We find that the generalization issue is often not as severe as reported in synthetic scenarios. Furthermore, our results corroborate previous findings that domain generalization often does not consistently outperform classical empirical risk minimization baselines also on low-dimensional data. Finally, to better understand when domain generalization methods can lead to performance gains and thus better outcomes for patients, we quantify the influence of different types of shifts occurring in the data.

Data and Code Availability Our study includes patient outcome and baseline characteristics data collected from several clinical studies on non-small-cell lung carcinoma (NSCLC) and diffuse large B-cell lymphoma (DLBCL). A short description of each dataset along with references is available in the Appendix. Details of the clinical trials can be found on https://clinicaltrials.gov/. Datasets are not publicly available at the time of writing, please contact the study team to obtain data access. Legal review for code sharing is in progress and the code cannot be shared by the time of manuscript submission. Please contact the authors regarding the access to the code.

1. Introduction

Machine learning is increasingly important in medical research. It can be used for a broad array of tasks ranging from improving the understanding of biological or chemical processes, automating and enhancing physician capabilities (e.g., by providing additional annotations or a second opinion in radiology) or providing additional diagnostic scores to predict patient survival. A central assumption for these tasks is that new data points stem from the same underlying distribution as that on which a machine learning model was trained (Widmer and Kubat, 1996; Quiñonero-Candela et al., 2009). This is a reasonable assumption when a large enough and representative data sample about the population we wish to predict for can be collected, or when variation among individuals is limited and central relationships between data and the corresponding property of interest are stable across sub-populations. Unfortunately, this is often not the case in clinical scenarios, e.g., when

data is collected across different sub-populations with access to different hospitals, standards of care and disease heterogeneity (Challen et al., 2019). In such clinical generalization scenarios, clinical models often show lower predictive performance in medical imaging applications (Zhang et al., 2021) such as radiography (Zech et al., 2018; Pooch et al., 2019; Cohen et al., 2020) and MRI imaging (Mårtensson et al., 2020). A popular example of such a scenario is the CAMELEYON17 dataset (Bándi et al., 2019), part of the WILDS (Koh et al., 2021) domain generalization benchmark. Baseline models that attempt to generalize a histology image segmentation task across hospitals saw an average drop in accuracy from 93.2% on training domains to 70.3% on target domains (Koh et al., 2021) due to variations in slide staining and hospital populations.

A solution to this problem is the use of Domain Generalization (DG) methods (Pan and Yang, 2010; Zhou et al., 2021) which identify models that are robust to such shifts in domains by learning domain invariant representations or predictors. Although previous DG research mostly focused on (medical) image classification scenarios, many clinical applications rely on low dimensional tabular data to predict the expected time to a clinical event, using methods from survival analysis. In these scenarios only few highly relevant features are available, and there is considerable error even for a Bayes optimal predictor. This makes clinical time-to-event prediction distinct from the high dimensional scenarios in image classification. In this study we therefore benchmark four DG methods against several Empirical Risk Minimization (ERM) based baseline methods with respect to different types of distribution shifts.

To study the efficacy of domain generalization methods on clinical survival data we ask two questions: 1) how reliable are domain generalization methods in typical clinical survival prediction scenarios and 2) can we provide additional understanding by investigating types of shifts occurring in those scenarios. We use five tabular datasets obtained from randomized controlled trials and from electronic health record derived real-world data. Our main contributions are the following:

- We quantify different types of distribution shifts in the data to characterize scenarios where DG can be successfully applied.
- We find that domain shift and performance degradation of ERM models for survival analysis with tabular data in real-world clinical cases are often smaller than reported in imaging domains (Koh et al., 2021; Zhang et al., 2021).
- We corroborate findings in the imaging literature (Gulrajani and Lopez-Paz, 2020; Koh et al., 2021; Zhang et al., 2021) that DG methods offer an improvement over ERM only in the limited cases in real-world survival analysis scenarios and is correlated with degree of domain shift.

It is important to note that model and hyperparameter selection (Guyon et al., 2010) in DG scenarios is an active field of research and existing strategies often fail to provide a satisfying solution (Gulrajani and Lopez-Paz, 2020). This has drastic empirical consequences, since practitioners, lacking an estimate of a model's generalization error to the target domain, cannot determine whether ERM or DG methods should be preferred and failures w.r.t. generalization can only be detected at the time of prediction.

2. Related work

The goal of domain generalization is to estimate the functional relationship $f(x)$ between a data set X sampled from an input space X and a corresponding outcome of interest $y \in Y$. We further ask that this estimate $f(x)$ generalizes across changes in $P(X)$, $P(Y)$ and $P(Y|X)$ across a set of source domains used for training and a target domain we aim to conduct inference on. In the case of survival analysis, Y is often the cumulative survival distribution $S_t(x)$ for an observation $x \in X$ at time point t. Each data point is assigned a domain d_i. We will denote with $D_{src,i}$ the set of observations sampled from source domain i and with $D_{tgt,i}$ observations sampled from the target domain i ($i \in 1, ..., k$). Models fitted on source domains can now suffer from various distribution shifts that lead to worsened performance. An additional often encountered problem called single-source DG in (Zhou et al., 2021) is that datasets often lack labeled source domains, which either requires identifying domains (Creager et al., 2021) before applying domain generalization, or the use of methods that do not require information about domains (Wang et al., 2019). In contrast to models on high-dimensional data, models studied in our manuscript might suffer less severely from poor generalization as has been shown e.g. in (Simon-Gabriel et al., 2018) under

small transformations (Azulay and Weiss, 2019) or adversarial examples (Szegedy et al., 2013).

2.1. Types of domain shifts

Distribution shifts may occur due to different reasons: Shifts in $P(X)$ might e.g. occur due to population differences between rural and urban hospitals, or shifts in $P(Y|X)$ occurring due to differences between clinical trials and treatment in the real-world. Typically, such shifts in distribution do not occur in isolation but domains exhibit several shifts of differing magnitudes. Depending on the perspective, several types of combined shifts can be identified (Zhang et al., 2015).

- Shift in $P(X)$ with constant $P(Y|X)$. This is often referred to as covariate shift in the literature (Zhang et al., 2015). In this case, model performance should theoretically not degrade, but in practice models might be oversimplified and under-fits the conditional models, which causes the predicted Y to depend on the input distribution $P(X)$.
- Shift in $P(Y|X)$. In this case, the optimal model takes into account variations in $P(Y|X)$ between source domains in order to predict the target domain.
- Shift in $P(Y)$: Since we model a functional relationship X→Y, a shift in Y can not occur in isolation (Zhang et al., 2015). define two types of shifts in Y for the reverse causal direction Y→X that result in subsequent changes of $P(X)$ or $P(Y|X)$. We assume effects only in the temporal direction $X \rightarrow Y$ in the remainder of this manuscript.

2.2. Domain generalization methods

Domain Generalization, in contrast to the related concepts of transfer learning and domain adaptation, does not assume access to or knowledge about statistics of the target domain (Pan and Yang, 2010). In recent years, a large variety of domain generalization methods have been proposed. Methods vary from kernel-based methods (Blanchard et al., 2011; Muandet et al., 2013) to approaches that incorporate causal frameworks such as Invariant Causal Predictions (ICP) (Peters et al., 2016; Rothenhäusler et al., 2021) as well as approaches that take a robustness perspective (Krueger et al., 2021; Sagawa et al., 2019). For brevity we only introduce methods relevant to our benchmark, a comprehensive overview over state-of-the-art DG methods is e.g. provided in (Zhou et al., 2021).

3. Method

3.1. ERM and domain generalization methods for survival analysis

3.1.1. Baselines

We investigate two methods based on empirical risk minimization (ERM) as baselines. We choose two widely used models, a Cox Proportional-Hazards model (coxph)(Cox, 1972) and a parametric model using a weibull distribution (weibull) (Kalbfleisch and Prentice, 2011). In coxph, the hazard function $h_i(t) = h_0(t)\, exp(\sum_{k=1}^{p} \theta_k x_{i,k})$, where each feature affects the hazard multiplicatively. In the Weibull model, baseline hazards are defined as $h_i(t) = \lambda\gamma t^{\gamma-1}$ with estimated shape γ and scale λ, as a linear combination of the features X.

3.1.2. Ensemble-based approaches

Ensembles of ML models can be used to obtain better generalizing estimators (Zhou et al., 2021). We investigate survival forests (Ishwaran et al., 2008) as well as more sophisticated survival quilts (Lee et al., 2019) as ensemble baselines for the survival context. Temporal quilting constructs ensembles of survival models assuring that the resulting model is a valid risk function. The core idea is to optimize weights $w_{j,t}$ for risk functions of individual ensemble members j and each time point t optimizing model calibration under a constraint for the predictive error using Bayesian Optimization.

3.1.3. Low-rank decomposition based approaches

Several low-rank decomposition based approaches have been proposed in literature (Khosla et al., 2018; Li et al., 2017; Piratla et al., 2020). We design a strategy heavily inspired by common-specific low-rank decomposition (LRD, (Piratla et al., 2020)). The core assumption is, that for each source domain k, the optimal model's parameters can be written as

$$\theta^\star = \theta_c + \gamma_k \theta_k$$

where θ_k is a domain specific effect and the goal therefore is to find model coefficients θ_c encompassing the

signal that is common across all domains. We perform a low-rank decomposition on the model coefficients of a cox proportional hazards model fitted on each domain in order to find a set of coefficients θ_c containing the domain-independent signal which is used for subsequent prediction on the target domains.

3.1.4. Invariant Risk Minimization (IRM) & Environment Inference for Invariant Learning (EIIL)

Arjovsky et al. (2019) propose Invariant Risk Minimization (IRM), a novel risk minimization strategy with the goal to discover domain-invariant classifiers Φ by solving the following minimization problem:

$$\min_{\Phi} \sum_{i} R^i(\Phi) + \lambda * \|\nabla_{w|w=1} R^i(w \cdot \Phi)\|^2$$

The resulting invariant predictor therefore is balanced by λ between predictive performance and a low gradient at w=1 as a measure of domain invariance (Arjovsky et al., 2019). If domain assignments d_i are latent, domains can be inferred as described in (Creager et al., 2021) (EIIL) by inferring domain assignments d_i, such that the domain invariance in the equation is maximized before training the model using IRM. This method will fit for the single-source DG use cases. We adapt IRM/EIIL a survival by optimizing for the negative log likelihood of the Cox PH risk as R^i (c.f. Kvamme and Borgan (2021)).

3.1.5. Continuously Index Domain Adaptation (CIDA)

Wang et al. (2020) propose an Encoder-Decoder based approach to obtain domain invariant representations E(x) for the scenarios where domain assignments i are continuous (Wang et al., 2020). The goal is to learn an encoder E that allows for training a predictor F on y which simultaneously does not permit predicting domain assignments i.

$$\min_{E,F} \max_{D} L_p(F(E(x, d_i)), y) - \lambda_d L_d(D(E(x, d_i)), d_i)$$

Models can be trained using either the $L_d = L_2$ (CIDA), i.e. the mean squared error loss or a probabilistic loss L_d modeling the mean and variance of a Gaussian distribution and optimizing for the negative log-likelihood (PCIDA). We configured the model both in a linear as well as a deep setting, implementation details can be found in the appendix. We adapt to a survival setting by employing the negative log likelihood of the Cox PH model as a loss for L_p (Kvamme and Borgan, 2021).

3.2. Quantifying shifts

We try to characterize types and magnitudes of shifts occurring in the data to better understand differences between the scenarios and investigate correlations between domain shifts and the improvement by DG methods. In particular, we measure the shift between the target domain and the pooled source domains. We propose 3 metrics allowing for measuring the different shifts, i.e., shifts in $P(X)$, $P(Y)$, and $P(Y|X)$ aforementioned. Other metrics for such shifts have been proposed in histopathology (Stacke et al., 2020), or structured biological data (Borgwardt et al., 2006).

3.2.1. Shift in $P(X)$

Shift in $P(X)$ are summarized using the Wasserstein distance (Dobrushin) between the distribution of the propensity score of data in the source and target domain. The propensity score of a sample in source or target domain was calculated using a logistic model with all features in X.

3.2.2. Shift in $P(Y)$

The distribution shift in $P(Y)$ were measured using the chi square statistics from the log-rank test between the outcomes in the source and target domain.

$$\tilde{\chi_i}^2 = \sum_{t=1}^{n} \frac{(O_{it} - E_{it})^2}{Var(E_{it})}$$

$$Var(E_{it}) = E_{ij}(\frac{N_t - O_t}{N_t})(\frac{N_t - N_{it}}{N_t - 1})$$

where O_{it} represents the observed number of events in the group i (target or source domain) over time, E_{it} represents the expected number of events in the group i over time, N_{it} represents the number of subject at time t in group i.

3.2.3. Shift in $P(Y|X)$

To measure the shift in $P(Y|X)$, we used the difference between a model trained on the source domain (Φ^{src}) and a model trained on the target domain (Φ^*, fitted with the training split of the target domain) in the Akaike information criterion (AIC) computed on the data from target domain for both models. Note

that the Φ^* is fitted in the target domain, which usually have a much smaller sample size than the source domain in our experiments.

$$D_{x,y} = AIC(Y_{tgt}, \Phi^*(X_{tgt})) - AIC(Y_{tgt}, \Phi^{src}(X_{tgt}))$$

3.3. Model selection

Since DG assumes no access to target domain data, no reliable estimates for the generalization error GE_{tgt} are available for model selection or hyperparameter tuning (Sagawa et al., 2019; Gulrajani and Lopez-Paz, 2020). Since we are interested in the performance of models, we investigate models from a *post-hoc* perspective by reporting *best-in-class* performance on target domain data from each approach (Oracle, ERM, DG). Since this is not possible in practice, we additionally investigate a setting where a model is selected based on performance on a 30% validation sample collected from each source domain. Then, the best model is refitted on the full data to compute GE_{tgt} after model selection. We investigate differences to post-hoc selection in order to assess the effect of model selection.

3.4. Synthetic domain shift

Besides scenario A, all other scenarios has natural domains that can be identified. For scenario A, we create a label based on the quantiles (0-20%, 21-40%, 41-60%, 61-80%, 81-100%) of propensity score of one sample, which primarily categorize data based on the distribution of P(X). We use the propensity labels together with the original domain labels to rearrange the source and target domains to 10 different subdomains. By different combinations of these subdomains, we are able to create new source and target datasets with different degrees of distribution shifts and to test the performance of domain generalization under these scenarios (see Appendix for more details).

3.5. Data

In this study, we focus on patient-level tabular data from the oncology domain, in particular diffuse large B-cell lymphoma (DLBCL) as well as squamous and non-squamous small cell lung cancer (NSCLC). Datasets are obtained either from randomized controlled trials (RCTs) or electronic health record based real-world data (RWD) collection efforts (for more details see Appendix). Measured covariates contain demographic information such as sex and age as well as clinical variables, e.g., Eastern Cooperative Oncology Group (ECOG) score (Oken et al., 1982) and lab test results. For NSCLC datasets (dataset D,E), we select five covariates following Alexander et al. (2017), while for DLBCL (dataset A,C,E) we select five variables chosen for the International Prognostic Index (IPI) (International Non-Hodgkin's Lymphoma Prognostic Factors Project, 1993). For both diseases, patients' death of all causes is used as the target event for survival analysis. We further created synthetic datasets based on the DLBCL data in case A, using the method described above.

The five different domain shift cases from real clinical data are summarized in Table 1. In order to provide an overview, we indicate the presence of detected shifts. The datasets contain between 733 and 3218 samples distributed into 5-8 source domains while target domains are pooled into a single domain for evaluation. Target domain sizes vary from 107 to 733 (Table 1). Distributions across datasets reflect shifts that might typically be encountered in clinical practice: A) build a prognostic model for DLBCL on RCT data and apply the model in the real-world dataset; B) train a model on squamous NSCLC trials and apply it to a non-squamous NSCLC trial, which may reflect the case of generalization of models between disease subtypes; C) train a model on a set of randomly selected treatment groups in the NSCLC trials and apply to other treatment groups in NSCLC trials, which aims to investigate the potential influence by the change of care; D) train a model based on younger DLBCL patient groups (0-60 yrs) and apply the model to the older population (60+), which tests the generalizability between demographic subpopulations; E) train a model on low/intermediate risk populations and apply to the high risk population of DLBCL patients, which tests generalizability across risk groups (Wang et al., 2021).

4. Domain Generalization on clinical data

We design an experiment comparing a set of DG methods reflective of existing approaches to ERM baselines including cox proportional-hazards models (Cox, 1972) and weibull models (Kalbfleisch and Prentice, 2011). Together with ensemble models presented above, this allows us to go one step further beyond questions posed in previous publications (Gulrajani and Lopez-Paz, 2020; Zhang et al., 2021): analyz-

Table 1: **Overview of datasets and generalization cases.** Criteria for check marks: if the Wasserstein distance of features between the datasets is larger than 0.1, we consider it as a case with shift in P(X); if the differences on survival outcomes between the source and target domains is significant under a log-rank test ($p-value < 0.05$), we consider it as a case with shift in P(Y); if the difference on δAIC is larger than 6, we consider it as a case with shift in $P(Y|X)$.

ID	Generalization	$P(X)$	$P(Y)$	$P(Y\|X)$	Samples size (target domain)	Sample size (source domain)	Number of source domains
A	RCT to RWD	✓	✓	✓	733	1060	1
B	Cancer subtypes	-	✓	✓	107	3111	6
C	Treatment groups	-	✓	✓	254	2857	8
D	Age group	-	-	-	191	542	4
E	Disease risk	-	✓	✓	151	582	4

ing whether DG not only improves over simple baselines but instead also over more advanced methods that are readily available in practice. We adapt the following experimental protocol for all studied settings: After splitting data into fixed source and target domains, we impute missing values using their median/mode and subsequently fit the model. Each experiment is replicated 10 times on a randomly drawn 90% sample of the data to obtain an estimate of the variance of results.

4.1. Performance metrics

Given that our outcome of interest is a continuous survival distribution, we measure the target domain generalization error (GE_{tgt}) in terms of the C-index (Harrell Jr et al., 1996) on data obtained from the target domain.

$$GE_{tgt} = C(Y_{tgt}, \Phi(X_{tgt}))$$

We further estimate the source domain generalization error GE_{src} using a random 70-30 split of source data from each source domain as training and validation data. The GE_{src} are used for the model selection process as the validation data strategy.

$$GE_{src} = C(Y^{'}_{src}, \Phi(X^{'}_{src}))$$

4.2. Oracle models

We further study an oracle condition to answer how much performance gain could be had, if access to the target domain data was available. Measured differences in C-index between source model and the oracle model stem from training on either a subset of the target domain (oracle model) or source domain data (source model). The resulting gap δ_{oracle} thus reflects differences in training data between source and target domain.

5. Results

5.1. Domain shifts and model performance degradation

In case A, we observe a significant difference in $P(X)$ (Wasserstein distance $W_p = 0.25$), $P(Y)$ ($\chi^2 = 23.49$, $p < 0.001$) and $P(Y|X)$ ($\Delta AIC = 12.62$) between the cohorts from RWD and the RCT. Case B and C both show distribution shifts in $P(Y)$ and $P(Y|X)$, but the metrics are larger in case B. In case D, we observe a moderate shift in $P(Y)$ ($\chi^2 = 5.55$, p = 0.02), but not in $P(Y|X)$ or $P(X)$. In case E, we found a significant difference in P(Y) ($\chi^2 = 8.75$, p = 0.003), and $P(Y|X)$ ($\Delta AIC = 16.33$). Metrics of domain shifts are summarised in Table 2 (for additional details see Fig. 3 in the Appendix).

When measuring performance degradation by comparing GE_{src} and GE_{tgt}, differences range between 0.01 and 0.05 across all experiments (Fig. 4 in the Appendix). Interestingly, we find that performance of the ERM models on the source domain are worse than on the target domain in the conducted experiments. When using the oracle models as a comparator, oracle models are also not always superior to the ERM models. The performances of oracle models are worse than ERM models in the cases D by 0.03 and E by 0.01. Note, that model performances can be worse than random guessing if models over-fit on source domain data.

Table 2: **Summary of experimental results.** We compare the best performing baseline with the best performing DG method selected as best-of-class (DG) and selected according to the validation data strategy (DG'). We report the domain shift by each metric for each tested scenario. For the model performance, we report mean C-index along with standard deviations (in brackets) over 10 monte-carlo cross-validation iterations. The best class is denoted in bold.

	A	**B**	**C**	**D**	**E**
Domain shift metrics					
Shift in $P(X)$ (W_p)	0.25	0.02	0.01	0.01	0.03
Shift in $P(Y)$ (χ^2)	23.49	12.43	3.89	5.55	8.75
Shift in $P(Y\|X)$ (ΔAIC)	12.62	19.24	6.32	2.76	16.33
C-index by method class					
Oracle	0.70(0.02)	0.65(0.07)	0.63(0.04)	0.66(0.08)	0.67(0.07)
ERM	0.70(0.01)	0.59(0.01)	0.62(0.01)	**0.71(0.01)**	**0.69(0.01)**
Ensemble	0.65(0.01)	0.61(0.01)	0.640(0.00)	0.64(0.01)	0.66 (0.01)
DG	0.70(0.01)	**0.63(0.01)**	**0.66(0.01)**	0.70(0.01)	0.67(0.01)
DG'	0.65(0.01)	0.60(0.01)	0.66(0.01)	0.66(0.01)	0.66(0.01)

5.2. Performance of DG methods

We report best-of-class results across all investigated DG and ERM methods to provide an overview. We find DG only outperforms ERM in two out of five real-world cases, with improvement on C-index between 0.03 and 0.05. In case A, the difference between the best DG and ERM was almost negligible. In case D and E, ERM gives yields better C-indices than DG (Table 2). Among the selected DG methods, LRD performed competitively well in all cases. CIDALinear was the best performing method for case C, and outperformed ERMs in two out of five cases. Full results can be found in Table 3 in the Appendix.

When applying the *validation data strategy* for model selection, the selected methods (*DG'*, Table 2) have a loss from 0.02 to 0.05 in C-index compared to the best-of-class results.

5.3. Correlation between domain shift and the improvement by DG methods

Since the introduced domain shift metrics such as chi-square statistics and ΔAIC are comparable with the same target domain, we derive an experiment allowing us to evaluate correlations with a fixed testing dataset while changing source domains to mimic scenarios with different degrees of shifts between source and target domain.

We first tested the correlation in case B. While keeping the same testing dataset, we remove clinical trials from the training datasets, two at each time, to create new training datasets. This generates shifts with ΔAIC ranging from 16.6 to 20.5. With increased domain shift, we observe a larger advantage in performance of DG over ERM based models (from 0.63 to 0.59 for DG, 0.63 to 0.56 for ERM, Figure 1).

Using synthetic domain labels, we applied the same approach and created four scenarios with increasing domain shift between the target and source domains based on case A (Fig. 5 in the Appendix). The ERM and DG models have similar performance in the initial scenario (median C-index of 0.70 in both cases with ΔAIC of 14). However, when the ΔAIC increases to 26 and 39, the performance was worse in ERM models (median C-index 0.68 and 0.66) compared to DG models (0.69 and 0.67, Fig. 2).

6. Discussion

6.1. Domain shift and generalizability issues in clinical data

A comparison between GE_{src} and GE_{tgt} is a common method to evaluate model performance under domain shift. In contrast to the literature on image data, where significant performance degradation has been reported in target domains (Arjovsky et al., 2019; Koh et al., 2021), we observed relatively small performance loss in most of the tested cases, which are low dimensional survival analysis settings. In our experiment we find domain shift are not reflected in differences between GE_{src} and GE_{tgt}, which is often

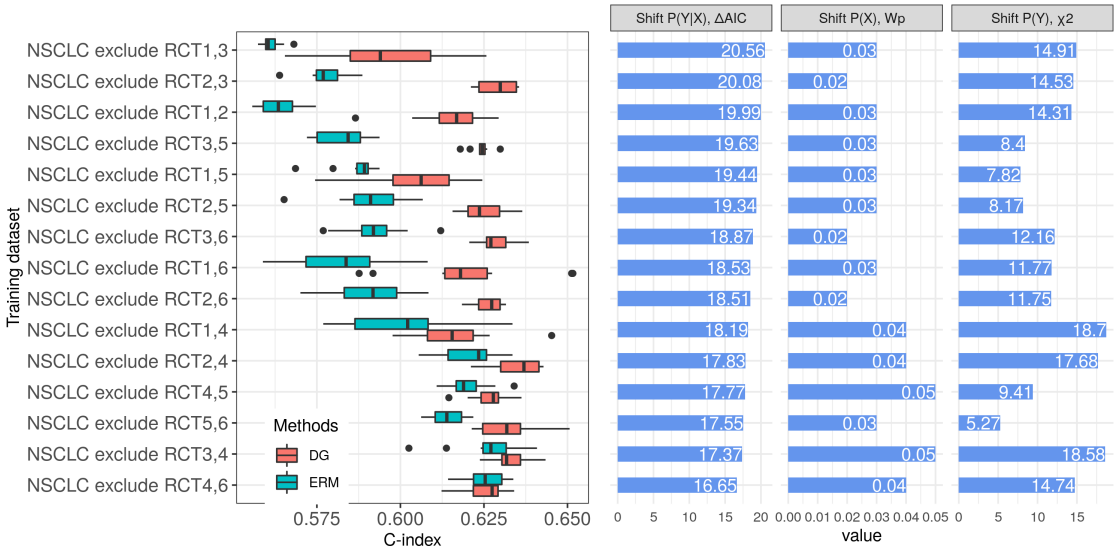

Figure 1: **Performance of ERM and DG models with increased domain shift in case B**. While keeping the same target domain, we create different training datasets by excluding trials from the source domain. See appendix for the details of the NSCLC RCTs.

used as an indicator for detecting potential generalization issues. We observe that models often exhibit $GE_{tgt} < GE_{src}$, which might stem from different Bayes error rates (Fukunaga, 1990) between the domains.

Oracle models are often used in previous studies to benchmark the performance of DG methods. However, oracle models may suffer from small sample sizes and large variations in the target domain. In the current study, oracle models also do not seem to be the upper limit of the performance as suggested in other literature (Zhang et al., 2021). In practice, comparing the ERM model with an oracle model trained on a small dataset from the target domain may not necessarily help to identify potential generalizability issues either. An additional open challenge is model selection from a set of candidate models, which can lead to severe performance degradation in comparison to the *post-hoc* best algorithm.

Because no single metric can directly measure the generalizability of a model, a set of carefully designed experiments are required to understand the underlying issues.

6.2. Synthetic data for DG assessment

Synthetic data allows to test DG methods in controlled experiments. However, previous publications have shown that DG methods only outperform ERMs in some special settings, particularly influenced by spurious correlations generated from the synthetic process. Synthetic DG scenarios such as coloured MNIST (Arjovsky et al., 2019) or artificial features often introduce spurious correlations in the source domains that are reversed on the target domain, which is perhaps rarely observed in real datasets (Zhang et al., 2021; Arjovsky et al., 2019).

Instead of creating synthetic datasets, we introduce a method to create synthetic domain labels based on propensity scores. One advantage of this approach is that the method only attempts to identify subclusters within the sample to be used as domains, and thus does not change the correlations between the features and the outcomes. This may mimic more realistic domain shift scenarios than directly modifying the distribution of the original data.

6.3. Factors influencing domain generalization methods

Previous studies reported that domain generalization provides no advantage in the case of more subtle data

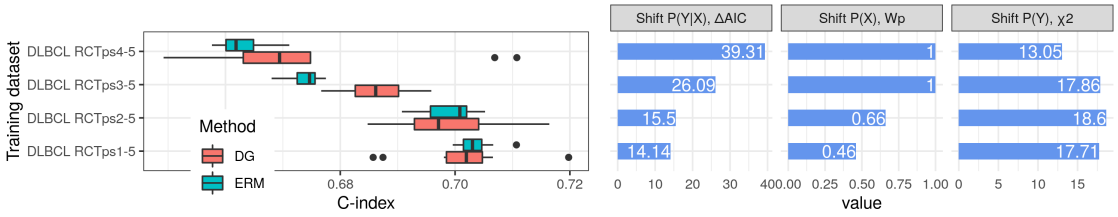

Figure 2: **Performance of ERM and DG models with increased domain shift in the synthetic data**. While keeping the same target domain as in case A, we create multiple training datasets by removing subgroups of patients from the original DLBCL trial data according to their propensity scores categories, which creates training datasets with different degrees of deviation from the target domain. As an example, RCTps2-5 includes patients in RCT with propensity score category 2-5. Models are tested on patients from RWD dataset with propensity score category 1 and 2.

shifts (Zhang et al., 2021; Gulrajani and Lopez-Paz, 2020). We observe that the DG methods show more benefits when the source and target domains have larger deviations. From a practical point of view, when the training datasets are expected to be close to the target domains, training with ERM methods might be sufficient; when the training data are more likely to be different from the target domain, e.g. from a different disease subtype (case B) or patients with different characteristics (the synthetic case), DG method could yield improvements.

It is worth noting that $\Delta AICs$ from different datasets are not directly comparable, comparisons are only meaningful when the models share the same test dataset. Although Case A and E have similar $\Delta AICs$ as in case B, it does not necessarily mean that they have the same degree of domain shift. On the other hand, there might be other factors influencing the efficacy of DG methods, such as the diversity of the source domains (Zhou et al., 2021) and sample size. In case A, since only one dataset is used in the training, the source domains are created by a random split of the dataset, which may result in very homogeneous source domains. Similarly in case E, although the target domain includes patients with higher risk, the source domains are similarly included patients with low/intermediate risk patients. Additionally, case E has a relatively small sample size in each source domain (100), which may influence the model fitting.

The process of model selection is another factor influencing the final performance of the DG methods (Gulrajani and Lopez-Paz, 2020). In our experiments, the selected models have a loss in c-index between 0.02 to 0.05 compared with post-hoc model selection. The model selection should not only include hyper-parameter tuning of the algorithms, but also feature selection. None of the original publications of the tested DG methods address the proper optimization under the DG scenario. As suggested previously (Zhang et al., 2021; Gulrajani and Lopez-Paz, 2020), the model selection strategy needs to be an integral part of a domain generalization method and its evaluation. Without it, the validity of the reported performance of these methods is limited.

6.4. Clinical applications and regulatory hurdles

It is encouraging for practitioners in the field that in most cases ERM methods are performing competitively. But we should also be aware that in many other scenarios DG methods do outperform ERM models. Both successful and failed attempts to use DG methods in different clinical applications abound in the literature (Lafarge et al.; Guo et al., 2021; Jin et al.). As observed here and elsewhere (Wang et al.), these conflicting observations may be explained by different degrees of domain shifts as well as the quantity and diversity of training data. The challenge thus lies in the correct choice of proper methods contingent on the recognition of the specific type of scenario. To aid this choice, we propose a set of metrics that can be used to qualify domain shifts, and to understand the diversity within the source domain. Additionally, estimating potential shifts in the target domain of the intended use cases will be a potential required step for identifying the proper use scenarios (Gossmann et al.).

For a clinical algorithm, the regulatory requirement plays a critical role. With the rising question on trustworthiness of the machine learning models, the request is not only on accuracy, but increasingly on transparency of the model training process, which includes a demonstration of model design tailored to the available data and intended use (FDA, 2021). These may imply a requirement of evidence to justify the use of selected methods, their applicable scenarios and potential risks, such as over-fitting, performance degradation, and security risks. In essence, this calls for comprehensive evaluation of methods before real-world application.

7. Conclusion

In this study, we evaluated four recently published domain generalization methods for their ability to generalize to an unseen data domain with real clinical data. Similar to previous findings with imaging data, these methods provided improvement over ERM for survival analysis in limited settings. However, our study is limited to comparatively low-dimensional settings (less than 10 features), which are often encountered in clinical practice. Richer settings employing a larger number of variables might result in different results due to possibly stronger over-fitting. We propose several metrics of domain shifts, and analyze the factors influencing the efficacy of the DG methods, which is a first step to find the right method that fits a particular domain shift scenario. Furthermore, data used throughout our study comes only from the US, a broader study across different populations could lead to interesting results.

Most of the current DG methods were developed for tasks outside low-dimensional clinical settings and may therefore not have been optimized for clinical use cases. We hope our work will encourage researchers in the field to further develop suitable DG methods for clinical research, as well as to develop more fitting evaluation frameworks and datasets to benchmark these methods.

Institutional Review Board (IRB)

The data used in this study were all published previously, the study did not require an IRB approval. For the original clinical studies, approval from the Independent Review Board (IRB)/Independent Ethics Committee (IEC) were obtained before the start of the studies, and all patients provided written informed consent.

Acknowledgments

The project is funded by F.Hoffmann-La Roche AG. The work was done while FP was interning at Fa. Hoffman-La Roche AG. We would like to thank Prof. Bernd Bischl for supporting FP taking the intern project and continuing with the write up after the internship.

References

Marliese Alexander, Rory Wolfe, David Ball, Matthew Conron, Robert G Stirling, Benjamin Solomon, Michael MacManus, Ann Officer, Sameer Karnam, Kate Burbury, et al. Lung cancer prognostic index: a risk score to predict overall survival after the diagnosis of non-small-cell lung cancer. *British journal of cancer*, 117(5):744–751, 2017.

Martin Arjovsky, Léon Bottou, Ishaan Gulrajani, and David Lopez-Paz. Invariant risk minimization. *arXiv preprint arXiv:1907.02893*, 2019.

Aharon Azulay and Yair Weiss. Why do deep convolutional networks generalize so poorly to small image transformations? *Journal of Machine Learning Research*, 20:1–25, 2019.

Gilles Blanchard, Gyemin Lee, and Clayton Scott. Generalizing from several related classification tasks to a new unlabeled sample. *Advances in neural information processing systems*, 24:2178–2186, 2011.

Karsten M Borgwardt, Arthur Gretton, Malte J Rasch, Hans-Peter Kriegel, Bernhard Schölkopf, and Alex J Smola. Integrating structured biological data by kernel maximum mean discrepancy. *Bioinformatics*, 22(14):e49–e57, 2006.

Péter Bándi, Oscar Geessink, Quirine Manson, Marcory Van Dijk, Maschenka Balkenhol, Meyke Hermsen, Babak Ehteshami Bejnordi, Byungjae Lee, Kyunghyun Paeng, Aoxiao Zhong, Quanzheng Li, Farhad Ghazvinian Zanjani, Svitlana Zinger, Keisuke Fukuta, Daisuke Komura, Vlado Ovtcharov, Shenghua Cheng, Shaoqun Zeng, Jeppe Thagaard, Anders B. Dahl, Huangjing Lin, Hao Chen, Ludwig Jacobsson, Martin Hedlund, Melih Çetin, Eren Halıcı, Hunter Jackson,

Richard Chen, Fabian Both, Jörg Franke, Heidi Küsters-Vandevelde, Willem Vreuls, Peter Bult, Bram van Ginneken, Jeroen van der Laak, and Geert Litjens. From detection of individual metastases to classification of lymph node status at the patient level: The camelyon17 challenge. *IEEE Transactions on Medical Imaging*, 38(2):550–560, 2019. doi: 10.1109/TMI.2018.2867350.

Robert Challen, Joshua Denny, Martin Pitt, Luke Gompels, Tom Edwards, and Krasimira Tsaneva-Atanasova. Artificial intelligence, bias and clinical safety. *BMJ Quality & Safety*, 28(3):231–237, 2019.

Joseph Paul Cohen, Mohammad Hashir, Rupert Brooks, and Hadrien Bertrand. On the limits of cross-domain generalization in automated x-ray prediction. In *Medical Imaging with Deep Learning*, pages 136–155. PMLR, 2020.

David R Cox. Regression models and life-tables. *Journal of the Royal Statistical Society: Series B (Methodological)*, 34(2):187–202, 1972.

Elliot Creager, Joern-Henrik Jacobsen, and Richard Zemel. Environment inference for invariant learning. In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 2189–2200. PMLR, 18–24 Jul 2021.

R. L. Dobrushin. Prescribing a system of random variables by conditional distributions. 15(3):458–486. ISSN 0040-585X. doi: 10.1137/1115049. Publisher: Society for Industrial and Applied Mathematics.

Health Center for Devices and Radiological FDA. Artificial Intelligence and Machine Learning in Software as a Medical Device. *FDA*, September 2021. Publisher: FDA.

Keinosuke Fukunaga. *Introduction to Statistical Pattern Recognition (Computer Science & Scientific Computing)*. Academic Press, hardcover edition, 10 1990. ISBN 978-0122698514.

Alexej Gossmann, Kenny H. Cha, and Xudong Sun. Performance deterioration of deep neural networks for lesion classification in mammography due to distribution shift: an analysis based on artificially created distribution shift. In *Medical Imaging 2020: Computer-Aided Diagnosis*, volume 11314, page 1131404. International Society for Optics and Photonics.

Ishaan Gulrajani and David Lopez-Paz. In search of lost domain generalization. *CoRR*, abs/2007.01434, 2020.

Lin Lawrence Guo, Stephen R Pfohl, Jason Fries, Alistair Johnson, Jose Posada, Catherine Aftandilian, Nigam Shah, and Lillian Sung. Evaluation of domain generalization and adaptation on improving model robustness to temporal dataset shift in clinical medicine. *medRxiv*, 2021. doi: 10.1101/2021.06.17.21259092.

Isabelle Guyon, Amir Saffari, Gideon Dror, and Gavin Cawley. Model selection: Beyond the bayesian/frequentist divide. *Journal of Machine Learning Research*, 11(3):61–87, 2010.

Frank E Harrell Jr, Kerry L Lee, and Daniel B Mark. Multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. *Statistics in medicine*, 15(4):361–387, 1996.

International Non-Hodgkin's Lymphoma Prognostic Factors Project. A predictive model for aggressive non-hodgkin's lymphoma. *N Engl J Med*, 329(14):987–994, 1993.

Hemant Ishwaran, Udaya B. Kogalur, Eugene H. Blackstone, and Michael S. Lauer. Random survival forests. *The Annals of Applied Statistics*, 2(3):841–860, September 2008. ISSN 1932-6157, 1941-7330. doi: 10.1214/08-AOAS169. Publisher: Institute of Mathematical Statistics.

Xin Jin, Cuiling Lan, Wenjun Zeng, Zhibo Chen, and Li Zhang. Style normalization and restitution for generalizable person re-identification. URL http://arxiv.org/abs/2005.11037.

John D Kalbfleisch and Ross L Prentice. *The statistical analysis of failure time data*, volume 360. John Wiley & Sons, 2011.

Aditya Khosla, Tinghui Zhou, Tomasz Malisiewicz, Alexei Efros, and Antonio Torralba. Undoing the damage of dataset bias, Jun 2018.

Pang Wei Koh, Shiori Sagawa, Sang Michael Xie, Marvin Zhang, Akshay Balsubramani, Weihua Hu, Michihiro Yasunaga, Richard Lanas Phillips, Irena

Gao, Tony Lee, et al. Wilds: A benchmark of in-the-wild distribution shifts. In *International Conference on Machine Learning*, pages 5637–5664. PMLR, 2021.

David Krueger, Ethan Caballero, Joern-Henrik Jacobsen, Amy Zhang, Jonathan Binas, Dinghuai Zhang, Remi Le Priol, and Aaron Courville. Out-of-distribution generalization via risk extrapolation (rex). In *International Conference on Machine Learning*, pages 5815–5826. PMLR, 2021.

Håvard Kvamme and Ørnulf Borgan. Continuous and discrete-time survival prediction with neural networks. *Lifetime Data Analysis*, pages 1–27, 2021.

Maxime W. Lafarge, Josien P. W. Pluim, Koen A. J. Eppenhof, and Mitko Veta. Learning domain-invariant representations of histological images. 6: 162. ISSN 2296-858X. doi: 10.3389/fmed.2019.00162.

Changhee Lee, William Zame, Ahmed Alaa, and Mihaela Schaar. Temporal quilting for survival analysis. In *The 22nd international conference on artificial intelligence and statistics*, pages 596–605. PMLR, 2019.

Da Li, Yongxin Yang, Yi-Zhe Song, and Timothy M Hospedales. Deeper, broader and artier domain generalization. In *Proceedings of the IEEE international conference on computer vision*, pages 5542–5550, 2017.

Gustav Mårtensson, Daniel Ferreira, Tobias Granberg, Lena Cavallin, Ketil Oppedal, Alessandro Padovani, Irena Rektorova, Laura Bonanni, Matteo Pardini, Milica G Kramberger, et al. The reliability of a deep learning model in clinical out-of-distribution mri data: a multicohort study. *Medical Image Analysis*, 66:101714, 2020.

Krikamol Muandet, David Balduzzi, and Bernhard Schölkopf. Domain generalization via invariant feature representation. In *International Conference on Machine Learning*, pages 10–18. PMLR, 2013.

Martin M Oken, Richard H Creech, Douglass C Tormey, John Horton, Thomas E Davis, Eleanor T McFadden, and Paul P Carbone. Toxicity and response criteria of the eastern cooperative oncology group. *American journal of clinical oncology*, 5(6): 649–656, 1982.

Sinno Jialin Pan and Qiang Yang. A survey on transfer learning. *IEEE Transactions on Knowledge and Data Engineering*, 22(10):1345–1359, 2010. doi: 10.1109/TKDE.2009.191.

Jonas Peters, Peter Bühlmann, and Nicolai Meinshausen. Causal inference by using invariant prediction: identification and confidence intervals. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, pages 947–1012, 2016.

Vihari Piratla, Praneeth Netrapalli, and Sunita Sarawagi. Efficient domain generalization via common-specific low-rank decomposition. In Hal Daumé III and Aarti Singh, editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 7728–7738. PMLR, 13–18 Jul 2020.

Eduardo HP Pooch, Pedro L Ballester, and Rodrigo C Barros. Can we trust deep learning models diagnosis? the impact of domain shift in chest radiograph classification. *arXiv preprint arXiv:1909.01940*, 2019.

Joaquin Quiñonero-Candela, Masashi Sugiyama, Neil D Lawrence, and Anton Schwaighofer. *Dataset shift in machine learning*. Mit Press, 2009.

Dominik Rothenhäusler, Nicolai Meinshausen, Peter Bühlmann, and Jonas Peters. Anchor regression: Heterogeneous data meet causality. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 83(2):215–246, 2021.

Shiori Sagawa, Pang Wei Koh, Tatsunori B Hashimoto, and Percy Liang. Distributionally robust neural networks for group shifts: On the importance of regularization for worst-case generalization. *arXiv preprint arXiv:1911.08731*, 2019.

Carl-Johann Simon-Gabriel, Yann Ollivier, Léon Bottou, Bernhard Schölkopf, and David Lopez-Paz. Adversarial vulnerability of neural networks increases with input dimension. 2018.

Raphael Sonabend, Franz J Király, Andreas Bender, Bernd Bischl, and Michel Lang. mlr3proba: An r package for machine learning in survival analysis. *Bioinformatics*, 02 2021. ISSN 1367-4803. doi: 10.1093/bioinformatics/btab039.

Karin Stacke, Gabriel Eilertsen, Jonas Unger, and Claes Lundström. Measuring domain shift for deep learning in histopathology. *IEEE journal of biomedical and health informatics*, 25(2):325–336, 2020.

Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199*, 2013.

Hao Wang, Hao He, and Dina Katabi. Continuously indexed domain adaptation. *arXiv preprint arXiv:2007.01807*, 2020.

Haohan Wang, Zexue He, Zachary C Lipton, and Eric P Xing. Learning robust representations by projecting superficial statistics out. *arXiv preprint arXiv:1903.06256*, 2019.

Jindong Wang, Cuiling Lan, Chang Liu, Yidong Ouyang, and Tao Qin. Generalizing to unseen domains: A survey on domain generalization. In *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence*, pages 4627–4635. International Joint Conferences on Artificial Intelligence Organization. ISBN 978-0-9992411-9-6. doi: 10.24963/ijcai.2021/628.

Xuejian Wang, Wenbin Zhang, Aishwarya Jadhav, and Jeremy Weiss. Harmonic-mean cox models: A ruler for equal attention to risk. In Russell Greiner, Neeraj Kumar, Thomas Alexander Gerds, and Mihaela van der Schaar, editors, *Proceedings of AAAI Spring Symposium on Survival Prediction - Algorithms, Challenges, and Applications 2021*, volume 146 of *Proceedings of Machine Learning Research*, pages 171–183. PMLR, 22–24 Mar 2021. URL https://proceedings.mlr.press/v146/wang21a.html.

Gerhard Widmer and Miroslav Kubat. Learning in the presence of concept drift and hidden contexts. *Machine learning*, 23(1):69–101, 1996.

John R Zech, Marcus A Badgeley, Manway Liu, Anthony B Costa, Joseph J Titano, and Eric Karl Oermann. Variable generalization performance of a deep learning model to detect pneumonia in chest radiographs: a cross-sectional study. *PLoS medicine*, 15(11):e1002683, 2018.

Haoran Zhang, Natalie Dullerud, Laleh Seyyed-Kalantari, Quaid Morris, Shalmali Joshi, and Marzyeh Ghassemi. *An Empirical Framework for Domain Generalization in Clinical Settings*, page 279–290. Association for Computing Machinery, New York, NY, USA, 2021. ISBN 9781450383592.

Kun Zhang, Mingming Gong, and Bernhard Schölkopf. Multi-source domain adaptation: A causal view. In *Twenty-ninth AAAI conference on artificial intelligence*, 2015.

Kaiyang Zhou, Ziwei Liu, Yu Qiao, Tao Xiang, and Chen Change Loy. Domain generalization: A survey, 2021.

Appendix A. Data and Methods

A.1. Datasets

The data are collected from the following studies: NSCLC:

- NCT01351415 (RCT1): A Study of Bevacizumab in Combination With Standard of Care Treatment in Participants With Advanced Non-squamous Non-small Cell Lung Cancer (NSCLC).
- NCT01496742 (RCT2): A Study of Onartuzumab (MetMAb) in Combination With Bevacizumab (Avastin) Plus Platinum And Paclitaxel or With Pemetrexed Plus Platinum in Patients With Non-Squamous Non-Small Cell Lung Cancer.
- NCT01903993 (RCT3): A Randomized Phase 2 Study of Atezolizumab (an Engineered Anti-PDL1 Antibody) Compared With Docetaxel in Participants With Locally Advanced or Metastatic Non-Small Cell Lung Cancer Who Have Failed Platinum Therapy - "POPLAR".
- NCT02008227 (RCT4): A Study of Atezolizumab Compared With Docetaxel in Participants With Locally Advanced or Metastatic Non-Small Cell Lung Cancer Who Have Failed Platinum-Containing Therapy (OAK).
- NCT02366143 (RCT5): A Study of Atezolizumab in Combination With Carboplatin Plus (+) Paclitaxel With or Without Bevacizumab Compared With Carboplatin+Paclitaxel+Bevacizumab in Participants With Stage IV Non-Squamous Non-Small Cell Lung Cancer (NSCLC) (IMpower150).
- NCT02657434 (RCT6): A Study of Atezolizumab in Combination With Carboplatin or Cisplatin + Pemetrexed Compared With Carboplatin or Cisplatin + Pemetrexed in Participants Who Are Chemotherapy-Naive and Have Stage IV Non-Squamous Non-Small Cell Lung Cancer (NSCLC) (IMpower 132).
- NCT01519804 (target): A Study of Onartuzumab (MetMAb) Versus Placebo in Combination With Paclitaxel Plus Platinum in Patients With Squamous Non-Small Cell Lung Cancer.

DLBCL:

- NCT01287741 (RCT): A Study of Obinutuzumab in Combination With CHOP Chemotherapy Versus Rituximab With CHOP in Participants With CD20-Positive Diffuse Large B-Cell Lymphoma (GOYA)
- FlatironHealth (RWD): This study used the nationwide Flatiron Health electronic health record (EHR)-derived de-identified database. We selected a subset of patients from the DLBCL cohort for the analyses.

A.2. Model Selection

Throughout the manuscript, we report the *best-in-class* model. That is, for each group of models (ERM, DG, Oracle condition), we report the best model based on the average performance on the held-out 30% validation data from each split.
Since DG assumes no access to source domain data, no reliable estimates for the generalization error GE are available. Mainly three strategies have been proposed in literature (Sagawa et al., 2019; Gulrajani and Lopez-Paz, 2020):

- Validation Data Measure generalization error using average performance on a hold-out sample from each source domain.
- Worst-case analysis Measure generalization error as a method's performance on the worst domain (Sagawa et al. 2019).
- Validation Domain Measure generalization error as a method's performance on a (randomly) held out source domain.

While each method can help to obtain better estimates of eventual performance on a target domain, their efficacy heavily depends on the (dis-)similarity between source and target domains.

A.3. Implementation Details

IRM & EIL

Since we consider low-dimensional datasets with only few observations, we consider simplistic (linear) neural networks in our benchmarks. The regularization parameter λ for both is tuned on a grid of values: $, 1e-7, 1e-5, 1e-3, 1e-2, 1e-1, .5$ and trained using the Adam optimizer with a learning rate of 0.01.

CIDA

We consider versions of CIDA and PCIDA that include only a linear predictor (CIDALinear and PCIDALinear) as well as a deep version including 4 layers of widths $(8, 12, 12, 8)$ respectively (CIDA, PCIDA). Since we study settings with ~ 5 covariates, we consider a width of 12 to be appropriately big for our neural networks.

A.4. Building synthetic domain shift

We created the synthetic dataset based on case A to mimic scenario different degree of domain shift. The case A contains only one dataset in the source domain. For the main task, we simply separated the data according to the age groups of the population (0-50, 51-60, 61-70, 71-80, 80+), however, this is an over simplification of categorizing the heterogeneous subgroups within the population.

We applied the method described to create synthetic domain labels. Firstly, a logistic regression model was fitted on the combined source and target domain to calculate the propensity of each sample being in one of the domains. Based on the propensity scores, we stratified the whole population according to the quantiles of the propensity scores (0-20%, 21-40%, 41-60%, 61-80%, 81-100%), which are used as the propensity labels (ps_1-5, Fig. 5). We then combined the propensity labels and the original domain labels (RCT or RWD) to create a new domain label for each stratum in the population (e.g. RCT1, indicates the patient comes from the RCT data and belongs to the propensity category ps_1). Based on different combinations of stratum we are able to created scenarios with different degrees of domains shifts.

Appendix B. Figures and Tables

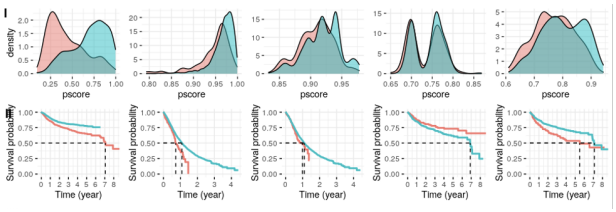

Figure 3: **Summary of the distribution of feature space X and outcome Y.** I) propensity score of data from the source and target domain; II) Kaplan-Meier curve of the source and target domain.

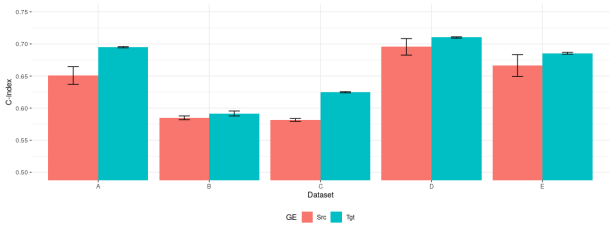

Figure 4: GE_{src} and GE_{tgt} of the ERM Model with error bars for the reported mean.

Table 3: **Full experimental results.** We report c-index across replications for all experiments by method and scenarios. Bold: best method according to GE_{tgt} except Oracle model, underlined: chosen via validation data strategy. We report averages along with standard deviations (in brackets).

Algorithm	A	B	C	D	E
Oracle	0.703 (0.023)	0.648 (0.069)	0.634 (0.036)	0.664 (0.078)	0.666 (0.069)
ERM					
coxph	**0.695 (0.003)**	0.592 (0.012)	0.625 (0.003)	**0.710 (0.003)**	0.686 (0.005)
weibull	0.694 (0.003)	0.588 (0.011)	0.624 (0.003)	0.710 (0.003)	0.687 (0.004)
LRD	0.683 (0.027)	**0.630 (0.002)**	0.629 (0.004)	0.698 (0.014)	0.654 (0.009)
CIDA					
CIDA	0.631 (0.062)	**0.588 (0.040)**	**0.601 (0.024)**	0.665 (0.044)	0.642 (0.043)
PCIDA	0.593 (0.067)	0.551 (0.054)	0.556 (0.051)	0.654 (0.036)	0.639 (0.049)
CIDALinear	0.648 (0.002)	0.601 (0.014)	**0.658 (0.014)**	0.659 (0.012)	0.659 (0.013)
PCIDALinear	0.646 (0.001)	0.595 (0.015)	0.607 (0.015)	0.683 (0.020)	0.674 (0.014)
IR					
IRM	0.345 (0.034)	0.465 (0.016)	0.478 (0.035)	0.605 (0.055)	0.518 (0.074)
EIIL	0.364 (0.019)	0.452 (0.002)	0.471 (0.002)	0.519 (0.110)	0.292 (0.007)
Ensemble					
surv.quilts	0.635 (0.049)	0.613 (0.006)	0.623 (0.002)	0.635 (0.007)	0.662 (0.009)
surv.forest	0.650 (0.004)	0.602 (0.008)	0.639 (0.003)	0.610 (0.004)	0.655 (0.008)

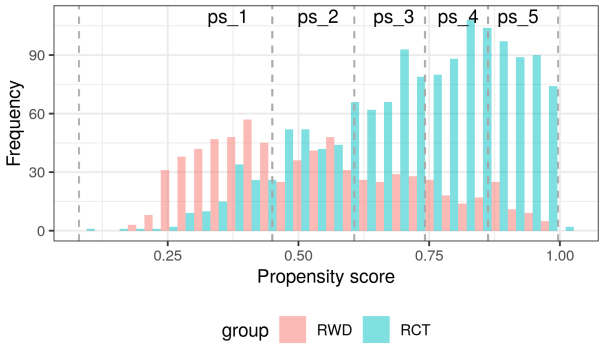

Figure 5: **Create synthetic domain labels based on propensity score in case A**. The RCT and RWD data were each categorized based on the quantiles of the propensity score (ps_1-5). Combining with the original domain labels, the target and source domains were divided into 10 different subdomains, labeled as RCT1-5 and RWD1-5.