

1028	<b>Appendix Contents</b>	
1029	<b>A Related Works</b>	<b>24</b>
1030	<b>B Missing Proof</b>	<b>25</b>
1031	<b>C Dataset Details</b>	<b>27</b>
1032	C.1 AREDS . . . . .	27
1033	C.2 MIMIC-CXR . . . . .	28
1034	C.3 ADNI . . . . .	28
1035	<b>D Algorithms Details</b>	<b>29</b>
1036	D.1 TTE Prediction Models . . . . .	29
1037	D.2 Fairness Algorithms . . . . .	30
1038	D.3 Model Architecture Details . . . . .	31
1039	<b>E Evaluation Metrics</b>	<b>31</b>
1040	E.1 Performance Metrics . . . . .	31
1041	E.2 Fairness Metrics . . . . .	33
1042	E.3 Fairness-Utility Trade-Off Metric . . . . .	33
1043	<b>F Experimental Setup Details</b>	<b>33</b>
1044	F.1 Implementation Details . . . . .	33
1045	F.2 Data Split and Pre-processing . . . . .	34
1046	F.3 Hyperparameter Search . . . . .	34
1047	F.4 Quantifying Source of Bias . . . . .	34
1048	F.5 Constructing Causal Distribution Shift . . . . .	35
1049	<b>G Causal Graphs for Fairness in TTE Prediction</b>	<b>36</b>
1050	G.1 Causal Graphs for Biased and Unbiased Settings . . . . .	36
1051	G.2 Real-world Causal Graph Examples for Fairness in TTE Prediction . . . . .	36
1052	<b>H Additional Results</b>	<b>38</b>
1053	H.1 Predictive Performance and Fairness in TTE Prediction Models . . . . .	38
1054	H.2 Comparison between Pre-Training and Training from Scratch Strategies for TTE	
1055	Prediction Models . . . . .	40
1056	H.3 Comparison with Advanced Image Backbones and Medical Pre-training for TTE	
1057	Prediction Models . . . . .	41
1058	H.4 Fairness in Fair TTE Prediction Models . . . . .	42
1059	H.5 Fairness-Utility Trade-Off Results . . . . .	44
1060	H.6 Additional Results for Predictive Performance and Fairness in Fair TTE Prediction	
1061	Models under Distribution Shift . . . . .	46

## A Related Works

**TTE Prediction.** TTE prediction models can generally be classified into two categories: continuous-time and discrete-time models, each with distinct approaches for handling event timing. Continuous-time models treat time as a continuous variable and often extend traditional models like Cox regression. For example, DeepSurv [34] extends the Cox regression by using a deep neural network with non-linear activation functions in hidden layers. Cox-Time [41] further builds on DeepSurv, introducing time-dependent predictors that allow for the estimation of time-varying effects. In contrast, discrete-time models treat time as a series of distinct intervals and typically use classification techniques. DeepHit [43] learns survival times directly without assuming a specific underlying stochastic process, parameterizing the discrete probability mass function. Another method, Nnet-survival [15], parametrizes the discrete hazard function using a neural network and optimizes the negative log-likelihood loss.

**Fairness in Machine Learning.** Fairness in machine learning has gained significant attention in recent years, with a focus on ensuring models are unbiased and equitable across individuals and groups. *Fairness metrics.* Fairness metrics can be broadly categorized into two types: group fairness [13, 7, 74, 18] and individual fairness [13, 59]. Group fairness ensures that models are fair across different demographic groups, while individual fairness emphasizes that similar individuals should be treated similarly. *Fairness algorithms.* To address fairness and bias issues, bias mitigation methods are generally classified into three approaches: pre-processing, which focuses on modifying the input data before model training [31, 4]; in-processing, which incorporates fairness constraints during model training [32, 75, 45, 60]; and post-processing, which adjusts model outputs to improve fairness [18, 24].

**Fairness in Medical Imaging.** In medical image analysis, machine learning (ML) models have been shown to exhibit systematic biases related to various attributes such as race, gender, and age [57, 42]. These biases are prevalent across different medical imaging modalities, including chest X-rays [56], CT scans [81], and skin dermatology images [38]. While several efforts have been made to benchmark fairness algorithms on medical images, existing datasets [23, 39, 66, 17] and benchmarks [82, 64] primarily focus on diagnostic tasks like image classification and segmentation. Unfortunately, they often overlook the crucial domain of medical prognosis, which involves predicting TTE outcomes.

**Fairness in TTE Prediction.** Despite significant advances in TTE prediction, research on fairness in this area remains limited. *Fairness metrics for TTE prediction.* Fairness metrics for TTE prediction have only recently been defined. These metrics can be roughly classified into three categories based on their objectives: (i) ensuring similar predicted TTE outcomes for similar data points [36, 53, 77, 76, 72], (ii) ensuring similar predicted outcomes for data points from different groups [36, 53, 80], and (iii) ensuring similar predictive performance across different groups [11, 22, 78, 79]. *Fairness algorithms for TTE prediction.* Building on these metrics, several methods have been proposed to achieve fair TTE prediction. One approach incorporates fairness as a regularization term during model training [36, 53, 11], ensuring that the model accounts for fairness constraints throughout its optimization process. Another approach focuses on improving worst-group accuracy by leveraging distributionally robust optimization techniques [22, 20, 55], which aim to enhance performance for underrepresented or disadvantaged groups. In addition to these in-processing methods, recent work has also explored pre- and post-processing strategies to address fairness in TTE prediction [80]. However, these efforts are limited to tabular data and fail to consider medical images, which are essential and pervasive in medical prognosis tasks.

## 1106 B Missing Proof

1107 **Proof for Theorem 1** For any  $a, a' \in \mathcal{A}$ , we have:

$$\begin{aligned}
 \text{Er}(f_{a'}, h, D_{a'}) &\leq \text{Er}(f_a, h, D_a) + |\text{Er}(f_a, h, D_a) - \text{Er}(f_{a'}, h, D_{a'})| \\
 &\stackrel{(1)}{\leq} \text{Er}(f_a, h, D_a) + |\text{Er}(f_a, h, D_a) - \text{Er}(h, h^*, D_a)| \\
 &\quad + |\text{Er}(h, h^*, D_a) - \text{Er}(h, h^*, D_{a'})| + |\text{Er}(h, h^*, D_{a'}) - \text{Er}(f_{a'}, h, D_{a'})| \\
 &\stackrel{(2)}{\leq} \text{Er}(f_a, h, D_a) + \text{Er}(f_a, h^*, D_a) + \mathcal{D}(\mathcal{H}, D_a, D_{a'}) + \text{Er}(f_{a'}, h^*, D_{a'}) \\
 &\stackrel{(3)}{=} \text{Er}(f_a, h, D_a) + \eta(\mathcal{H}, f_a, f_{a'}) + \mathcal{D}(\mathcal{H}, D_a, D_{a'})
 \end{aligned}$$

1108 where  $h^* = \arg \min_{h' \in \mathcal{H}} (\text{Er}(f_a, h', D_a) + \text{Er}(f_{a'}, h', D_{a'}))$ . We have  $\stackrel{(1)}{\leq}$  by using inequality  $|a +$   
 1109  $b| \leq |a| + |b|$ ;  $\stackrel{(2)}{\leq}$  by using triangle inequality for Er metric and  $|\text{Er}(h, h^*, D_a) - \text{Er}(h, h^*, D_{a'})| \leq$   
 1110  $\max_{h', h'' \in \mathcal{H}} |\text{Er}(h', h'', D_a) - \text{Er}(h', h'', D_{a'})| = \mathcal{D}(\mathcal{H}, D_a, D_{a'})$ ;  $\stackrel{(3)}{=}$  because  $\eta(\mathcal{H}, f_a, f_{a'}) =$   
 1111  $(\text{Er}(f_a, h^*, D_a) + \text{Er}(f_{a'}, h^*, D_{a'}))$  by definition. Subtracting  $\text{Er}(f_a, h, D_a)$  from both sides and  
 1112 taking max operator, we have:

$$\mathcal{F}_{\text{Er}}(h) = \max_{a, a' \in \mathcal{A}} |\text{Er}(f_a, h, D_a) - \text{Er}(f_{a'}, h, D_{a'})| \leq \max_{a, a' \in \mathcal{A}} (\eta(\mathcal{H}, f_a, f_{a'}) + \mathcal{D}(\mathcal{H}, D_a, D_{a'}))$$

1113 **Discussion on the assumption of performance metric.** The proof of Theorem 1 relies on the  
 1114 assumption that the performance metric Er satisfies the properties of triangle inequality and symmetry.  
 1115 This assumption is relatively mild and holds for many commonly used performance metrics. For  
 1116 instance, [58] introduced the symmetric discordance index (SDI), a ranking-based metric that adheres  
 1117 to these properties, demonstrating the practical applicability of this assumption in practice.

1118 **Proof of Proposition 2** For  $a \in \mathcal{A}$  with  $I_a(Z, T) = I_a(X, T)$ , we have:

$$\begin{aligned}
 \log P(t|x, a) &= \log \left( \int P(t, z|x, a) dz \right) \\
 &= \log \left( \int P(t|z, a) P(z|x, a) dz \right) \\
 &= \log (\mathbb{E}_{P(z|x)} [P(t|z, a)]) \\
 &\stackrel{(1)}{\geq} \mathbb{E}_{P(z|x)} [\log p(t|z, a)]
 \end{aligned} \tag{5}$$

1119 We have  $\stackrel{(1)}{\geq}$  by using Jensen's inequality.  $\forall a' \in \mathcal{A}$ , taking expectation w.r.t.  $P(x, t|a')$  over both  
 1120 sides, we have:

$$\begin{aligned}
 &\mathbb{E}_{P(x, t|a')} [\log P(t|x, a) - \mathbb{E}_{P(z|x)} [\log P(t|z, a)]] \\
 &= \int \int (\log P(t|x, a) - \mathbb{E}_{P(z|x)} [\log P(t|z, a)]) P(x, t|a') dx dt \\
 &= \int \int (\log P(t|x, a) - \mathbb{E}_{P(z|x)} [\log P(t|z, a)]) P(x, t|a) \frac{P(x, t|a')}{P(x, t|a)} dx dt \\
 &= \mathbb{E}_{P(x, t|a)} \left[ (\log P(t|x, a) - \mathbb{E}_{P(z|x)} [\log P(t|z, a)]) \frac{P(x, t|a')}{P(x, t|a)} \right]
 \end{aligned}$$

$$\begin{aligned}
&\stackrel{(1)}{\leq} \left( \max_{x,t} \frac{P(x,t|a')}{P(x,t|a)} \right) \mathbb{E}_{P(x,t|a)} [\log P(t|x,a) - \mathbb{E}_{P(z|x)} [\log P(t|z,a)]] \\
&= \left( \max_{x,t} \frac{P(x,t|a')}{P(x,t|a)} \right) (\mathbb{E}_{P(x,t|a)} [\log P(t|x,a)] - \mathbb{E}_{P(z,t|a)} [\log P(t|z,a)]) \\
&= \left( \max_{x,t} \frac{P(x,t|a')}{P(x,t|a)} \right) (H_a(T, X) - H_{a'}(T, Z)) \\
&= \left( \max_{x,t} \frac{P(x,t|a')}{P(x,t|a)} \right) ((H_a(T) - H_a(T, Z)) - (H_a(T) - H_a(T, X))) \\
&= \left( \max_{x,t} \frac{P(x,t|a')}{P(x,t|a)} \right) (I_a(T, Z) - I_a(T, X)) \\
&\stackrel{(2)}{=} 0
\end{aligned} \tag{6}$$

1121 We have  $\stackrel{(1)}{\leq}$  because  $\log P(t|x,a) - \mathbb{E}_{P(z|x)} [\log P(t|z,a)] \geq 0$  according to Eq. (5);  $\stackrel{(2)}{=}$  because  
1122  $I_a(T, Z) = I_a(T, X)$ . Based on Eq. (6), we have:

$$\begin{aligned}
\mathbb{E}_{P(x,t|a')} [\log P(t|x,a)] &= \mathbb{E}_{P(x,t|a')} [\mathbb{E}_{P(z|x)} [\log P(t|z,a)]] \\
&= \mathbb{E}_{P(t,z|a')} [\log P(t|z,a)]
\end{aligned} \tag{7}$$

1123 We also have:

$$\begin{aligned}
\mathbb{E}_{P(t,z|a')} [\log P(t|z,a')] &= -H_{a'}(T|Z) \\
&= I_{a'}(T, Z) - H_{a'}(T) \\
&\stackrel{(1)}{\leq} I_{a'}(T, X) - H_{a'}(T) \\
&= -H_{a'}(T|X) \\
&= \mathbb{E}_{P(x,t|a')} [\log P(t|x,a')]
\end{aligned} \tag{8}$$

1124 We have  $\stackrel{(1)}{\leq}$  by using data processing inequality. Finally, we have:

$$\begin{aligned}
&\mathbb{E}_{P(z|a')} [\mathcal{D}_{KL}(P(t|z,a') \parallel P(t|z,a))] \\
&\stackrel{(1)}{=} \mathbb{E}_{P(z|a')} [\mathcal{D}_{KL}(P(t|z,a') \parallel P(t|z,a))] - \mathbb{E}_{P(x|a')} [\mathcal{D}_{KL}(P(t|x,a') \parallel P(t|x,a))] \\
&= \mathbb{E}_{P(t,z|a')} [\log P(t|z,a') - \log P(t|z,a)] - \mathbb{E}_{P(x,t|a')} [\log P(t|x,a') - \log P(t|x,a)] \\
&= (\mathbb{E}_{P(t,z|a')} [\log P(t|z,a')] - \mathbb{E}_{P(x,t|a')} [\log P(t|x,a')]) \\
&\quad + (\mathbb{E}_{P(x,t|a')} [\log P(t|x,a)] - \mathbb{E}_{P(t,z|a')} [\log P(t|z,a)]) \\
&\stackrel{(2)}{=} 0
\end{aligned} \tag{9}$$

1125 We have  $\stackrel{(1)}{=}$  because the shift between two domains w.r.t. input space  $\mathcal{X}$  is covariate shift;  $\stackrel{(2)}{=}$  by  
1126 using Eq. (7) and Eq. (8) and the fact that KL-divergence is non-negative. Note that Eq. (9) implies  
1127 that the shift between these two domains w.r.t. representation space  $\mathcal{Z}$  is also covariate shift (i.e.,  
1128  $P(t|z,a) = P(t|z,a') = P(t|z), \forall a, a' \in \mathcal{A}$ ).



Table A1: Overview of medical image datasets for fair TTE prediction evaluation.

Dataset	Prediction Task	Modality	Subgroup	Attribute	# images	Censoring rate	Mean TTE
AREDS	Late AMD	Retinal Fundus	Age	Total	129708	83.9%	4.4 (years)
				≤70	44224	83.1%	5.1
				>70	85484	84.4%	3.9
			Sex	Female	71837	83.0%	4.4
				Male	57871	85.1%	4.3
			Race	Non-white	4888	99.2%	3.1
White	124820	83.3%	4.4				
MIMIC-CXR	In-hospital Mortality	Chest X-ray	Age	Total	269360	61.7%	488.6 (days)
				≤60	103437	77.3%	503.3
				>60	165923	52.0%	484.2
			Sex	Female	125742	63.9%	514.4
				Male	143618	59.7%	468.4
			Race	Non-white	83234	66.7%	487.2
White	186126	59.4%	489.1				
ADNI	Alzheimer's Disease	Brain MRI	Age	Total	2227	63.2%	35.9 (months)
				≤80	1597	62.2%	37.1
				>80	630	65.7%	32.6
			Sex	Female	986	64.7%	38.5
				Male	1241	62.0%	33.9

## C Dataset Details

### C.1 AREDS

#### C.1.1 Dataset Description

The Age-Related Eye Disease Study (AREDS) [14] was a clinical trial conducted between 1992 and 2001 across 11 retinal specialty clinics in the United States. The primary objective was to study the risk factors for age-related macular degeneration (AMD) and the impact of dietary supplements on AMD progression. The study followed 4,757 participants, aged 55–80 at enrollment, for a median of 6.5 years. Participants were selected with a broad range of AMD severity, from no AMD to late-stage AMD in one eye. At each visit, certified technicians captured color fundus photography images using a standardized imaging protocol, although adherence to the protocol varied, leading to visits at irregular intervals. AMD severity scores were determined by expert graders at the University of Wisconsin Fundus Photograph Reading Center, with late AMD defined as the presence of neovascular AMD or atrophic AMD with geographic atrophy (severity scores from 10 to 12 using severity scale [8]). For this study, we include images from both the left and right eyes of each participant and make predictions separately for each eye. In cases where multiple images were captured per eye during a visit, we select only one image for analysis. Demographic information, including age, sex, and race, is available in the dataset and is used as sensitive attributes in our study.

#### C.1.2 TTE Outcome Construction

In our study, TTE outcomes for fundus images are defined as the duration, in years, from the date an image was captured to the first recorded diagnosis of late AMD in the corresponding eye. For eyes without a recorded diagnosis of late AMD, the censoring dates are set to the time of the last imaging visit. To ensure that the model forecasts the future risk of developing late AMD, all images taken during the final visit for a given eye were excluded from the dataset, as this visit was used solely to determine the TTE outcome. By removing the final visit, we ensured that no images were included with late AMD already present at the time of acquisition. This process resulted in a final dataset of 129,708 fundus images, each paired with corresponding TTE information, enabling a robust analysis of the TTE prediction for AMD progression.

#### C.1.3 Data Access

AREDS is a publicly available dataset hosted in the National Center for Biotechnology Information (NCBI) database of Genotypes and Phenotypes (dbGAP) through controlled access. Researchers can request access to this dataset at dbGAP. Once the application is approved, researchers can access data

1160 at the following address: [https://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/study](https://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/study.cgi?study_id=phs000001.v3.p1)  
1161 [cgi?study\\_id=phs000001.v3.p1](https://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/study.cgi?study_id=phs000001.v3.p1)

## 1162 C.2 MIMIC-CXR

### 1163 C.2.1 Dataset Description

1164 The MIMIC-CXR dataset [27] is a comprehensive, publicly available collection of chest X-ray  
1165 images, along with associated clinical data, from the larger MIMIC-IV [26] database. It includes  
1166 over 370,000 chest X-ray images from more than 65,000 patients, annotated with both structured  
1167 and unstructured clinical information, such as patient demographics, diagnoses, and other relevant  
1168 clinical details. For our study, we utilize the MIMIC-CXR-JPG [27], a processed version of the  
1169 MIMIC-CXR dataset, which provides images in JPG format derived from the original DICOM files.  
1170 Additionally, we link this dataset to the broader MIMIC database to access patient demographic  
1171 information, including age, sex, and race, which we incorporate as sensitive attributes in our fairness  
1172 benchmarking framework.

### 1173 C.2.2 TTE outcome construction

1174 To construct TTE outcomes for chest X-ray images, we extract in-hospital mortality events from the  
1175 MIMIC-IV database. For patients without a recorded date of mortality, the censoring dates are set as  
1176 1 year after their last recorded discharge date. We exclude any images that do not have a matching  
1177 record in the MIMIC-IV patient table, images taken after the latest discharge date, and images taken  
1178 after a recorded date of mortality. TTE is then calculated as the number of days from the image study  
1179 date to either the date of mortality or the censoring date. This process results in a final dataset of  
1180 269,360 chest X-ray images, each paired with corresponding TTE information, enabling a robust  
1181 analysis of the TTE prediction for in-hospital mortality.

### 1182 C.2.3 Data Access

1183 MIMIC-CXR, MIMIC-CXR-JPG, and MIMIC-IV are a publicly available datasets hosted by Phys-  
1184 ioNet, which is a platform providing access to medical data. To access the dataset, researchers first  
1185 need to create an account on PhysioNet and then complete the required training (CITI training). Once  
1186 researchers have completed the CITI training, they will need to request access to the dataset at the  
1187 following address:

- 1188 • MIMIC-IV: <https://physionet.org/content/mimiciv/3.1/>
- 1189 • MIMIC-CXR-JPG: <https://physionet.org/content/mimic-cxr-jpg/2.1.0/>

## 1190 C.3 ADNI

### 1191 C.3.1 Dataset Description

1192 The Alzheimer’s Disease Neuroimaging Initiative (ADNI) [49] is a large, longitudinal study aimed at  
1193 identifying biomarkers for Alzheimer’s disease (AD) and tracking the progression of the disease over  
1194 time. Launched in 2004, ADNI is one of the most comprehensive datasets for studying Alzheimer’s  
1195 disease and other neurodegenerative disorders. It includes a wide range of data types, including clinical  
1196 assessments, neuroimaging (MRI, PET), genetic data, and fluid biomarkers (e.g., cerebrospinal  
1197 fluid and blood samples) from over 1,700 participants. These participants are categorized into different  
1198 diagnostic groups, including cognitively normal individuals, those with mild cognitive impairment  
1199 (MCI), and individuals with Alzheimer’s disease. For our study, we focus on neuroimaging data,  
1200 specifically MRI scans, and consider age and sex as sensitive attributes in our analysis.

### 1201 C.3.2 TTE outcome construction

1202 In our study, TTE outcomes for brain MRI scans are defined as the duration, measured in 6-month  
1203 intervals, from the date an MRI scan was captured to the first recorded diagnosis of Alzheimer’s  
1204 disease. For participants without a recorded diagnosis of Alzheimer’s disease, censoring dates are set  
1205 to the time of their last imaging visit. To ensure that the model predicts the future risk of developing  
1206 Alzheimer’s disease, we excluded all MRI scans taken during the final visit, as this visit was used

solely for determining the TTE outcome. By removing the final visit, we ensured that no MRI scans were included for participants who had already been diagnosed with Alzheimer’s disease at the time of acquisition. This process resulted in a final dataset of 2,227 brain MRI scans, each paired with corresponding TTE information, enabling a comprehensive analysis of TTE prediction for Alzheimer’s disease progression.

### 1212 C.3.3 Data Access

ADNI is a publicly available dataset hosted in the Image and Data Archive (IDA), a secure online resource for archiving, exploring and sharing neuroscience data. Access to the ADNI dataset requires that researchers register for an IDA account. Once the account is created and the ADNI Data Use Agreement is completed, they can access data at the following address: <https://adni.loni.usc.edu>

## 1218 D Algorithms Details

### 1219 D.1 TTE Prediction Models

#### 1220 D.1.1 PMF

PMF [40] is a discrete-time model designed for TTE prediction tasks. It represents the survival time PMF  $P(t|x)$  using a neural network  $f(\cdot; \theta) : \mathcal{X} \rightarrow [0, 1]^L$  where  $\theta$  denotes the model parameters and  $L$  represents the number of time intervals. The model parameters  $\theta$  are estimated by minimizing the negative log-likelihood, averaged over the training data, as follows:

$$\mathcal{L}_{PMF}(\theta) = -\frac{1}{n} \sum_{i=1}^n \left\{ \delta_i \log(f_{\kappa(y_i)}(x_i; \theta)) + (1 - \delta_i) \log \left( \sum_{m=\kappa(y_i)+1}^L f_m(x_i; \theta) \right) \right\}$$

where  $n$  is the number of training samples,  $\kappa(y_i)$  denotes the specific time interval (from  $1, 2, \dots, L$ ) that corresponds to time  $y_i$ , and  $f_m$  represents the predicted probability that the TTE falls within the  $m - th$  interval.

#### 1228 D.1.2 DeepHit

DeepHit [43] extends the PMF model by incorporating a ranking loss function alongside the negative log-likelihood. Given a comparable set defined as  $\mathcal{E} := \{(i, j) \in [n] \times [n] : \delta_i = 1, y_i < y_j\}$ , this ranking loss is calculated over the training data, as follows:

$$\mathcal{L}_{DeepHit}^{rank}(\theta) = \sum_{(i,j) \in \mathcal{E}} \exp \left( \frac{-\left( \sum_{m=1}^{\kappa(y_i)} f_m(x_i; \theta) - \sum_{m'=1}^{\kappa(y_j)} f_{m'}(x_j; \theta) \right)}{\sigma} \right)$$

where  $n$  is the number of training samples,  $\kappa(y_i)$  denotes the specific time interval (from  $1, 2, \dots, L$ ) that corresponds to time  $y_i$ , and  $f_m$  represents the predicted probability that the TTE falls within the  $m - th$  interval.

#### 1235 D.1.3 Nnet-survival

Nnet-survival [15] is another discrete-time model designed for TTE prediction tasks. It represents the hazard function  $h(\cdot|x)$  using a neural network  $f(\cdot; \theta) : \mathcal{X} \rightarrow \mathbb{R}^L$  where  $\theta$  denotes the model parameters and  $L$  represents the number of time intervals. Specifically, Nnet-survival sets the hazard function equal to:

$$h[\ell|x; \theta] := \frac{1}{1 + e^{-f_\ell(x; \theta)}} \quad \text{for } \ell \in [L], x \in \mathcal{X}$$

where  $f_m(x; \theta)$  is the  $m$ -th output of the neural network. The model parameters  $\theta$  are estimated by minimizing the negative log-likelihood, averaged over the training data, as follows:

$$\mathcal{L}_{Net-survival}(\theta) = \frac{1}{n} \sum_{i=1}^n \left\{ \delta_i \log \left( 1 + e^{-f_{\kappa(y_i)}(x_i; \theta)} \right) + (1 - \delta_i) \log \left( 1 + e^{f_{\kappa(y_i)}(x_i; \theta)} \right) + \sum_{m=1}^{\kappa(y_i)-1} \log \left( 1 + e^{f_m(x_i; \theta)} \right) \right\}$$

where  $n$  is the number of training samples and  $\kappa(y_i)$  denotes the specific time interval (from  $1, 2, \dots, L$ ) that corresponds to time  $y_i$ .

## D.2 Fairness Algorithms

### D.2.1 Distributional Group Optimization

Distributional Group Optimization [55, 22] is an optimization technique designed to enhance model robustness by minimizing the worst-case training loss across different groups. Instead of optimizing for the average performance, GroupDRO focuses on the group with the highest loss, ensuring that the model does not disproportionately underperform on any particular group. By incorporating increased regularization, this approach helps mitigate disparities in model performance across diverse subpopulations, making it particularly useful in settings where fairness and reliability across groups are critical.

### D.2.2 Subgroup Rebalancing

This resampling method [31] addresses class imbalance by upsampling minority groups, ensuring that all groups have equal representation during training. By increasing the frequency of underrepresented samples, the model is exposed to a more balanced dataset, reducing bias and improving fairness. This approach helps prevent the model from being overly influenced by the majority group, leading to more equitable predictions across all groups.

### D.2.3 Fair Representation Learning

A common approach to promoting fairness in machine learning is through fair representation learning, which aims to obfuscate sensitive group membership information in the learned representations. By ensuring that the model's latent features do not encode discriminatory patterns, this method helps mitigate bias in downstream predictions. Following [70], we incorporate fairness constraints into representation learning by leveraging kernel-based distribution matching via Maximum Mean Discrepancy. This technique enforces similarity in feature distributions across different groups, reducing disparities while preserving task-relevant information.

### D.2.4 Domain Independence

The Domain Independence [71] is a method that trains separate classifiers for different groups while utilizing a shared encoder. This approach allows the model to capture group-specific patterns through distinct classifiers while maintaining a common feature representation across all groups. By leveraging a shared encoder, DomainInd enhances generalization and reduces the risk of overfitting to individual groups, ultimately improving the model's robustness and fairness in diverse domains.

### D.2.5 Controlled for Sensitive Attribute

This approach [80, 51], similar to Domain Independence, involves training separate models for each group based on the values of a sensitive attribute. By doing so, the method captures group-specific patterns while maintaining model flexibility. During the prediction phase, the outputs of the fitted models are averaged across all groups in the population. This averaging process ensures that no single group disproportionately influences the predictions, promoting fairness and reducing bias in the model's outcomes.

### D.3 Model Architecture Details

Each (fair) TTE prediction model comprises two key networks: an image encoder and a classifier. The image encoder transforms images into representation vectors, while the classifier predicts survival time intervals based on these learned representations. In our study, we employ a 2D EfficientNet [61] backbone as the image encoder for the AREDs and MIMIC-CXR datasets, and a 3D ResNet-18 backbone [65] for the ADNI dataset. To adapt these models for our task, we replace the original fully connected layers with new layers that map images into the representation space. The architectural details of our 2D and 3D image encoders, along with the classifier, are presented in Table A2.

Table A2: Architecture details of (fair) TTE prediction models. In our experiments, we set **n\_channel** = 3 for both 2D and 3D images by duplicating grayscale chest X-ray and brain MRI images to obtain three-channel inputs. **feature\_dim** is set to 64, and **hidden\_dim** is set to 16. **n\_class** is determined using the 'equidistant' discretization method, with values of 14 for AREDs, 28 for ADNI, and 128 for MIMIC-CXR.

Networks	Layers
Image Encoder	2D image <ul style="list-style-type: none"> <li>Conv2d(input channel = <b>n_channel</b>, output channel = 32, kernel = 3)</li> <li>MBConv1(input channel = 32, output channel = 16, kernel = 3)</li> <li>MBConv6(input channel = 16, output channel = 24, kernel = 3) * 2</li> <li>MBConv6(input channel = 24, output channel = 40, kernel = 5) * 2</li> <li>MBConv6(input channel = 40, output channel = 80, kernel = 3) * 3</li> <li>MBConv6(input channel = 80, output channel = 112, kernel = 5) * 3</li> <li>MBConv6(input channel = 112, output channel = 192, kernel = 5) * 4</li> <li>MBConv6(input channel = 192, output channel = 320, kernel = 3)</li> <li>Conv2d(input channel = 320, output channel = 1280, kernel = 1)</li> <li>AdaptiveAvgPool2d(output_size = 1)</li> <li>Linear(input dim = 1280, output dim = feature_dim)</li> <li>Dropout(p=0.5)</li> </ul>
	3D image <ul style="list-style-type: none"> <li>Conv3d(input channel = <b>n_channel</b>, output channel = 64, kernel = 3×7×7)</li> <li>Conv3d(input channel = 64, output channel = 64, kernel = 3×3×3) * 4</li> <li>Conv3d(input channel = 64, output channel = 128, kernel = 3×3×3) * 4</li> <li>Conv3d(input channel = 128, output channel = 256, kernel = 3×3×3) * 4</li> <li>Conv3d(input channel = 256, output channel = 512, kernel = 3×3×3) * 4</li> <li>AdaptiveAvgPool3d(output_size=(1,1,1))</li> <li>Linear(input dim = 512, output dim = <b>feature_dim</b>)</li> <li>Dropout(p=0.5)</li> </ul>
Classifier	Linear(input dim = <b>feature_dim</b> , output dim = <b>hidden_dim</b> ) Linear(input dim = <b>hidden_dim</b> , output dim = <b>n_classes</b> )

## E Evaluation Metrics

### E.1 Performance Metrics

The performance metrics used in our study are defined based on [5].

#### E.1.1 Time-dependent concordance index

Harrell's concordance index (C-index) is one of the most widely used accuracy metrics in TTE prediction. It quantifies the fraction of data point pairs that are correctly ranked by a prediction model among those that can be unambiguously ranked. The C-index values range from 0 to 1, with 1 indicating perfect ranking accuracy. However, a notable limitation of the C-index is its dependence on a risk score function for ranking, which many TTE prediction models do not explicitly learn. To address this limitation, [11] introduced a time-dependent concordance index ( $C^{td}$ ), which leverages predicted survival functions,  $\hat{S}(\cdot|x)$ , to assess model performance more effectively. The  $C^{td}$  is computed as follows.

1300 **Definition 3.** Suppose that we have a survival function estimate  $\hat{S}(\cdot|x)$  for any  $x \in \mathcal{X}$ . Then using  
 1301 the set of comparable pairs  $\mathcal{E} := \{(i, j) \in [n] \times [n] : \delta_i = 1, y_i < y_j\}$ , we define the  $C^{td}$  metric as:

$$C^{td} := \frac{1}{\mathcal{E}} \sum_{(i,j) \in \mathcal{E}} \mathbb{1} \left\{ \hat{S}(y_i|x_i) < \hat{S}(y_i|x_j) \right\}$$

1302 which is between 0 and 1. Higher scores are better.

### 1303 E.1.2 Time-dependent AUC

1304 While the C-index and  $C^{td}$  scores provide valuable single-number summaries of predictive accuracy  
 1305 in TTE prediction, they lack the ability to evaluate accuracy at a specific user-defined time,  $t$ . To  
 1306 address this limitation, [67] introduced time-dependent AUC scores ( $AUC^{td}$ ), which explicitly  
 1307 depend on the chosen time point  $t$ . The core idea behind  $AUC^{td}$  is to frame a binary classification  
 1308 problem for a fixed time  $t$ , where the “positive” class consists of data points that experienced the  
 1309 event no later than  $t$ , and the “negative” class includes those that survived beyond  $t$ . The survival  
 1310 function  $\hat{S}(t|\cdot) : \mathcal{X} \rightarrow [0, 1]$  serves as the probabilistic classifier, predicting survival probabilities. A  
 1311 lower predicted survival probability for a given point  $x$  implies a higher likelihood of belonging to  
 1312 the positive class. The  $AUC^{td}$  score quantifies the classifier’s ability to distinguish between these  
 1313 two classes at time  $t$ , offering a time-specific accuracy assessment of the model’s predictions. The  
 1314  $AUC^{td}$  is computed as follows.

1315 **Definition 4.** Suppose that we have a survival function estimate  $\hat{S}(\cdot|x)$  for any  $x \in \mathcal{X}$ . Then for  
 1316 any  $t > 0$ , using the set of comparable pairs  $\mathcal{E}(t) := \{(i, j) \in [n] \times [n] : \delta_i = 1, y_i \leq t, y_j > t\}$ , we  
 1317 define the  $AUC^{td}(t)$  (the  $AUC^{td}$  at time  $t$ ) as:

$$AUC^{td}(t) := \frac{\sum_{(i,j) \in \mathcal{E}(t)} w_i \mathbb{1} \left\{ \hat{S}(t|x_i) < \hat{S}(t|x_j) \right\}}{\sum_{(i,j) \in \mathcal{E}(t)} w_i}$$

1318 where  $w_1, w_2, \dots, w_n \in [0, \infty)$  are inverse probability of censoring weights to be defined as  
 1319  $w_i := 1/(\hat{S}_{censor}(y_i)\hat{S}_{censor}(t))$ , and  $\hat{S}_{censor}(t)$  is an estimation of  $S_{censor}(t) := P(C > t)$  using  
 1320 Kaplan-Meier estimator [33].  $AUC^{td}(t)$  is between 0 and 1 and higher scores are better. Finally, we  
 1321 can get  $AUC^{td}$  as

$$AUC^{td} := \frac{1}{t_{\max} - t_{\min}} \int_{t_{\min}}^{t_{\max}} AUC^{td}(u) du$$

1322 where  $t_{\min}$  and  $t_{\max}$  are user-specified lower and upper limits of integration.

### 1323 E.1.3 Integrated Brier Score

1324 The Integrated Brier Score ( $IBS$ ) is a performance metric that directly evaluates the error of an  
 1325 estimated survival function  $\hat{S}(\cdot|x)$  without relying on ranking. The  $IBS$  is calculated as follows.

1326 **Definition 5.** Suppose that we have a survival function estimate  $\hat{S}(\cdot|x)$  for any  $x \in \mathcal{X}$ . Then for any  
 1327  $t > 0$ , we define the  $BS(t)$  (the  $IBS$  at time  $t$ ) as:

$$BS(t) := \frac{1}{N} \sum_{i=1}^n \left( \frac{\hat{S}(t|x_i)^2 \delta_i \mathbb{1} \{y_i \leq t\}}{\hat{S}_{censor}(y_i)} + \frac{(1 - \hat{S}(t|x_i)^2) \mathbb{1} \{y_i > t\}}{\hat{S}_{censor}(t)} \right)$$

1328 which is nonnegative. Lower scores are better. Finally, we can get  $IBS$  as

$$IBS := \frac{1}{t_{\max} - t_{\min}} \int_{t_{\min}}^{t_{\max}} BS(u) du$$

1329 where  $t_{\min}$  and  $t_{\max}$  are user-specified lower and upper limits of integration.

## 1330 E.2 Fairness Metrics

1331 In this study, we define fairness metrics as the predictive performance gaps between groups. This  
 1332 kind of metric is used ensure that the model maintains equal predictive performance across different  
 1333 groups. In particular, given a performance metric  $Er$  for TTE prediction task, we define fairness  
 1334 metric  $\mathcal{F}_{Er}$  as follow:

$$\mathcal{F}_{Er}(h) = \max_{a, a' \in \mathcal{A}} |Er_a - Er_{a'}|$$

1335 where  $\mathcal{A}$  is the set of groups considered in TTE prediction task,  $Er_a$  and  $Er_{a'}$  are the predictive  
 1336 performance metrics calculated from subsets containing data from groups  $a$  and  $a'$ , respectively.  
 1337 For each predictive performance metric defined above, we have a corresponding fairness metric as  
 1338 follows.

$$\begin{aligned}\mathcal{F}_{C^{td}} &= \max_{a, a' \in \mathcal{A}} |C_a^{td} - C_{a'}^{td}| \\ \mathcal{F}_{AUC^{td}} &= \max_{a, a' \in \mathcal{A}} |AUC_a^{td} - AUC_{a'}^{td}| \\ \mathcal{F}_{IBS} &= \max_{a, a' \in \mathcal{A}} |IBS_a - IBS_{a'}|\end{aligned}$$

1339 where  $C_a^{td}$ ,  $AUC_a^{td}$ ,  $IBS_a$  are predictive performance metrics calculated from the subset containing  
 1340 data from group  $a$ , and  $C_{a'}^{td}$ ,  $AUC_{a'}^{td}$ ,  $IBS_{a'}$  are predictive performance metrics calculated from the  
 1341 subset containing data from group  $a'$ .

## 1342 E.3 Fairness-Utility Trade-Off Metric

1343 The fairness metrics mentioned above do not capture the fairness-utility trade-off while in medical  
 1344 context, it is essential to balance fairness and utility to ensure that the model is not only fair but also  
 1345 accurate and effective for all groups. To handle this issue, we leverage the equity-scaling metric  
 1346 ( $ES$ ) [44] that takes both utility and fairness into account for evaluation. Similar to fairness metric,  
 1347 for each predictive performance metric, we have a corresponding fairness-utility trade-off metric as  
 1348 follows.

$$\begin{aligned}ES_{C^{td}} &= \frac{C_D^{td}}{1 + \sum_{a \in \mathcal{A}} |C_D^{td} - C_{D_a}^{td}|} \\ ES_{AUC^{td}} &= \frac{AUC_D^{td}}{1 + \sum_{a \in \mathcal{A}} |AUC_D^{td} - AUC_{D_a}^{td}|} \\ ES_{IBS} &= \frac{1 - IBS_D}{1 + \sum_{a \in \mathcal{A}} |IBS_D - IBS_{D_a}|}\end{aligned}$$

1349 The advantage of the equity-scaling metric lies in its intuitive interpretability. Specifically, a higher  
 1350 equity-scaling score indicates that the model is both more accurate and more equitable simultane-  
 1351 ously.

## 1352 F Experimental Setup Details

### 1353 F.1 Implementation Details

#### 1354 F.1.1 Hardware Usage

1355 The experiments were conducted at a supercomputing center utilizing multiple compute nodes. Each  
 1356 node was equipped with an NVIDIA Volta V100 GPU with 16 GB of memory, an Intel Xeon CPU,  
 1357 and 32 GB of RAM, ensuring the computational resources necessary for large-scale experiments.  
 1358 In total, we trained over 20,000 models, requiring approximately 4.56 GPU years of computational  
 1359 effort, highlighting the extensive scale of our study.

#### 1360 F.1.2 Package Usage

1361 The FairTTE benchmark is implemented using Python 3, with PyTorch [47] serving as the framework  
 1362 for deep learning computations. The implementation of TTE models is built on the pycox [41]



package, while the evaluation metrics for TTE prediction leverage pycox, scikit-survival [50], and SurvivalEVAL [52]. Additionally, the training and evaluation pipeline for TTE prediction models is adapted from the demo code provided in [5], ensuring a robust and standardized framework for benchmarking.

## 1367 F.2 Data Split and Pre-processing

1368 **Data Split.** Each dataset in our study was divided into training, validation, and testing sets using a  
 1369 60%:20%:20% split ratio. Models were trained on the training sets, evaluated on the testing sets, and  
 1370 the validation sets were used for model selection. Since a single patient may have multiple medical  
 1371 records, we took precautions to prevent data leakage during model training. Specifically, the data was  
 1372 split by patient, ensuring that no patient appearing in the testing set had any records in the training  
 1373 or validation sets. This approach maintains the integrity of the evaluation process and ensures that  
 1374 model performance is assessed on entirely unseen patient data.

1375 **Data Pre-processing.** Before being fed into the TTE prediction models, chest X-ray and color  
 1376 fundus images are resized to  $224 \times 224$  pixels, while brain MRI scans are resized to  $128 \times 128 \times 96$ .  
 1377 Additionally, all pixel values are normalized to a range of 0 to 1 to ensure stability during training  
 1378 and improve model performance.

1379 We consider binary group setting in our experiment. These groups were constructed according to the  
 1380 following criteria:

- 1381 • Race: 'Non-White' (Group 0), 'White' (Group 1)
- 1382 • Sex: 'Female' (Group 0), 'Male' (Group 1)
- 1383 • Age:
  - 1384 – MIMIC-CXR: ' $\leq 60$ ' (Group 0), '> 60' (Group 1)
  - 1385 – AREDs: ' $\leq 70$ ' (Group 0), '> 70' (Group 1)
  - 1386 – ADNI: ' $\leq 80$ ' (Group 0), '> 80' (Group 1)

## 1387 F.3 Hyperparameter Search

1388 To ensure a fair comparison, we perform a grid-based hyperparameter search using 10 random seeds.  
 1389 The details of the hyperparameter search for the methods used in our experiments are provided below.

- 1390 • TTE prediction models
  - 1391 – Learning rate:  $10^x$  where  $x \sim \text{Uniform}(-4, -3)$
  - 1392 – Decay rate:  $10^x$  where  $x \sim \text{Uniform}(-6, -4)$
- 1393 • Fair TTE prediction models
  - 1394 –  $\eta$  :  $10^x$  where  $x \sim \text{Uniform}(-3, -1)$  (DRO)
  - 1395 –  $\lambda$  :  $10^x$  where  $x \sim \text{Uniform}(-5, 2)$  (FRL)

1396 For standard TTE prediction models, we select the best models based on their predictive performance  
 1397 metrics calculated on the validation sets. In contrast, for fair TTE prediction models, we prioritize  
 1398 fairness metrics when selecting the best models, allowing for up to a 5% reduction in accuracy  
 1399 compared to the baseline TTE models. This approach ensures a balanced trade-off between fairness  
 1400 and predictive performance.

## 1401 F.4 Quantifying Source of Bias

1402 In order to quantify the degree of bias sources in each dataset and sensitive attribute setting, we use  
 1403 several metrics as follows.

1404 **Disparity in mutual information between  $X_Z$  and  $Y$  across groups.** We quantify the disparity  
 1405 in mutual information between  $X_Z$  (i.e., image representation generated from the vision backbones)



and  $Y$  across groups by computing the maximum difference in their normalized mutual information values across all groups, as defined below.

$$Bias_{MI(X_Z, Y)} = \max_{a, a' \in \mathcal{A}} \left| \frac{2I(X_Z, Y|A = a, \Delta = 1)}{H(X_Z|A = a, \Delta = 1) + H(Y|A = a, \Delta = 1)} - \frac{2I(X_Z, Y|A = a', \Delta = 1)}{H(X_Z|A = a', \Delta = 1) + H(Y|A = a', \Delta = 1)} \right|$$

where  $I(\cdot, \cdot|A = a, \Delta = 1)$  represented the mutual information conditioned on  $A = a$  and  $\Delta = 1$  and  $H(\cdot|A = a)$  denotes the entropy conditioned on  $A = a$  and  $\Delta = 1$ .

**Disparity in mutual information between  $X_Z$  and  $\Delta$  across groups.** Similarly, we quantify the disparity in mutual information between  $X_Z$  (i.e., image representation generated from the vision backbones) and  $\Delta$  across groups by computing the maximum difference in their normalized mutual information values across all groups, as defined below.

$$Bias_{MI(X_Z, \Delta)} = \max_{a, a' \in \mathcal{A}} \left| \frac{2I(X_Z, \Delta|A = a)}{H(X_Z|A = a) + H(\Delta|A = a)} - \frac{2I(X_Z, \Delta|A = a')}{H(X_Z|A = a') + H(\Delta|A = a')} \right|$$

**Disparity in TTE distribution across groups.** We measure the disparity in TTE distributions across groups by calculating the maximum Wasserstein distance [63], normalized by the range of TTE, between the TTE distributions of each group. This is defined as follows:

$$Bias_{TTE} = \max_{a, a' \in \mathcal{A}} \left| \frac{\mathcal{W}(P(Y|A = a, \Delta = 1), P(Y|A = a', \Delta = 1))}{\max_{y \in \mathcal{Y}} y} \right|$$

where  $\mathcal{W}(\cdot, \cdot)$  denotes the Wasserstein-1 distance between the two distributions.

**Disparity in image distribution across groups.** We measure the disparity in image distributions across groups by calculating the maximum Wasserstein distance [63], normalized by the range of image feature values, between the image distributions of each group. This is defined as follows:

$$Bias_{Image} = \max_{a, a' \in \mathcal{A}} \left| \frac{\mathcal{W}(P(X_Z|A = a), P(X_Z|A = a'))}{\max_{y \in \mathcal{Y}} y} \right|$$

where  $\mathcal{W}(\cdot, \cdot)$  denotes the Wasserstein-1 distance between the two distributions. Due to the high dimensionality of image representations, we implement sliced Wasserstein distance [3], a variant of the Wasserstein distance that approximates the full Wasserstein distance between high-dimensional distributions by projecting them onto one-dimensional subspaces and averaging the resulting 1D Wasserstein distances.

**Disparity in censoring rate across groups.** We quantify the disparity in censoring rates across groups by calculating the maximum normalized difference between the means of the censoring distributions for each group, as defined below.

$$Bias_{Censoring} = \max_{a, a' \in \mathcal{A}} \left| \frac{\mathbb{E}[\Delta|A = a] - \mathbb{E}[\Delta|A = a']}{\mathbb{E}[\Delta]} \right|$$

## F.5 Constructing Causal Distribution Shift

To construct distribution shift between training and testing data, we modify the training data by introducing correlations between the sensitive attribute and other RVs in the causal graph (Figure 2). This adjustment simulates real-world scenarios where biases in data collection or underlying relationships may lead to disparities across groups. The details of this process, including the specific modifications applied to establish these correlations, are outlined below.

- **Distribution shift on  $X$ :** Images from disadvantaged groups are degraded using a Gaussian blur filter to simulate lower-quality data.
- **Distribution shift on  $Y$ :** TTE labels for disadvantaged groups are corrupted by adding noise sampled from a uniform distribution.
- **Distribution shift on  $\Delta$ :** To simulate biased censoring, we flip the censoring indicators for 90% of uncensored samples within disadvantaged groups.

## 1441 G Causal Graphs for Fairness in TTE Prediction

### 1442 G.1 Causal Graphs for Biased and Unbiased Settings

1443 Figure A1 presents the causal graphs for the unbiased and biased scenarios. In the unbiased scenario  
 1444 (Figure A1a), the sensitive attribute  $A$  is unrelated to the TTE outcome and influences only  $X_A$ ,  
 1445 with no effect on other variables in the graph. In contrast, in the biased scenarios (Figure A1b and  
 1446 Figure A1c),  $A$  also affects additional variables, resulting in dependencies between  $A$  and the TTE  
 1447 outcome. These causal pathways may be direct (Figure A1b), mediated through unobserved variables  
 1448  $U$  (Figure A1c), or both.

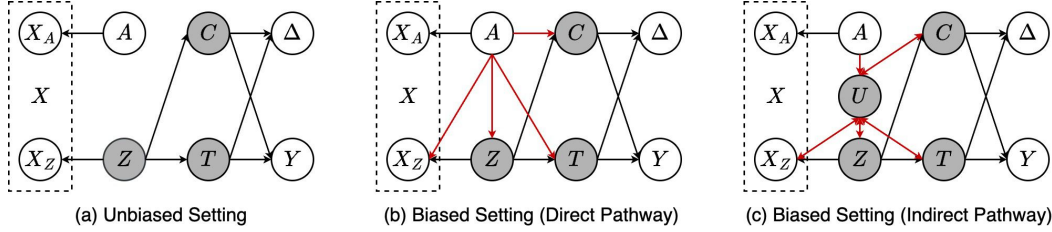


Figure A1: Causal structure in TTE prediction. Gray circles denote unobserved random variables. (a) Unbiased setting, where the sensitive attribute  $A$  influences only  $X_A$ . (b) Biased setting with direct causal pathways, where  $A$  is directly associated (red arrows) with other variables in the graph. (c) Biased setting with indirect causal pathways, where  $A$  influences (red arrows) other variables through unobserved variables  $U$ .

### 1449 G.2 Real-world Causal Graph Examples for Fairness in TTE Prediction

1450 In this section, we present causal graphs illustrating real-world scenarios in time-to-event (TTE)  
 1451 prediction using medical imaging. Many of these examples are adapted from diagnostic settings  
 1452 in prior work [28]. We describe four scenarios in which the sensitive attribute  $A$  influences other  
 1453 variables in the causal graph—namely, the medical image  $X$ , the underlying condition  $Z$ , the time-to-  
 1454 event  $T$ , and the censoring time  $C$ —leading to disparities in group-specific data distributions. For  
 1455 each scenario, we include two examples: one where the causal pathway from  $A$  is valid (red arrows),  
 1456 appearing in both training and testing data, and one where the pathway is spurious (blue arrows),  
 1457 representing bias present only in the training data and absent in the testing data.

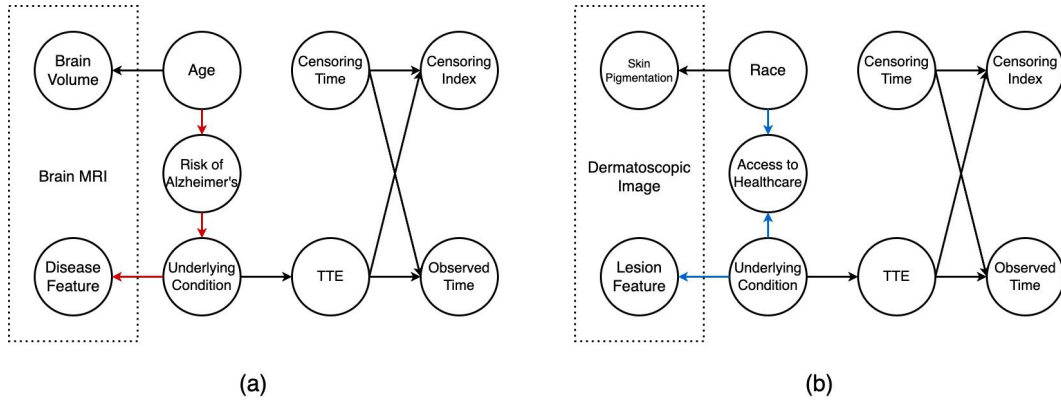


Figure A2: Causal graphs illustrating scenarios where the sensitive attribute  $A$  affects the underlying condition  $Z$ . (a) Valid pathway: age is a known clinical risk factor for Alzheimer's disease. (b) Invalid pathway: race appears spuriously correlated with  $Z$  due to disparities in healthcare access.

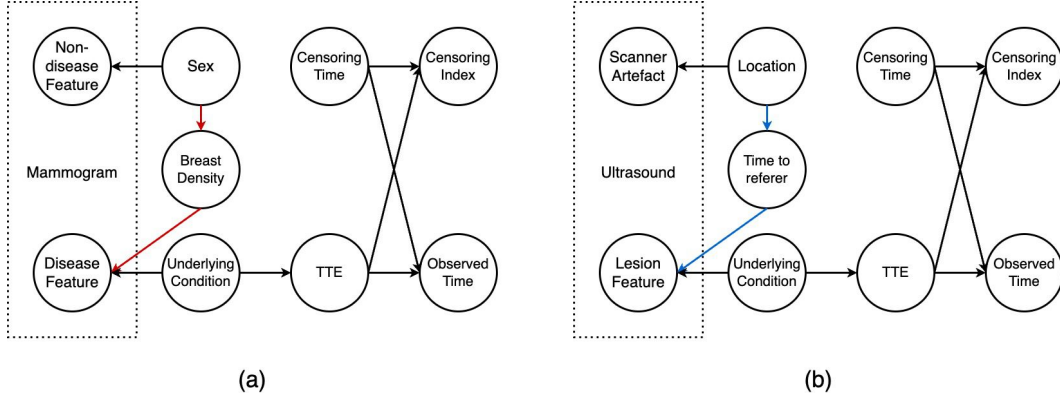


Figure A3: Causal graphs illustrating scenarios where the sensitive attribute  $A$  influences the medical image  $X$ . (a) Valid pathway: breast cancer presents differently in men and women due to inherent differences in breast tissue. (b) Invalid pathway: spurious correlation arises when patients in different locations are imaged at varying disease stages due to inconsistent ultrasound referral policies.

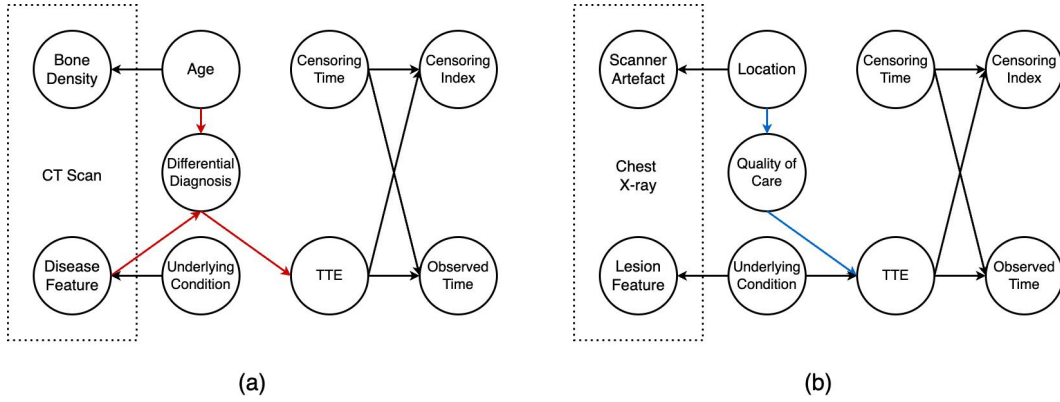


Figure A4: Causal graphs illustrating scenarios where the sensitive attribute  $A$  influences the time-to-event outcome  $T$ . (a) Valid pathway: age contributes to differential diagnosis in epidemiology and legitimately affects disease progression. (b) Invalid pathway: a spurious correlation arises when patients from different locations receive healthcare services of varying quality, impacting  $T$  in a non-causal manner.

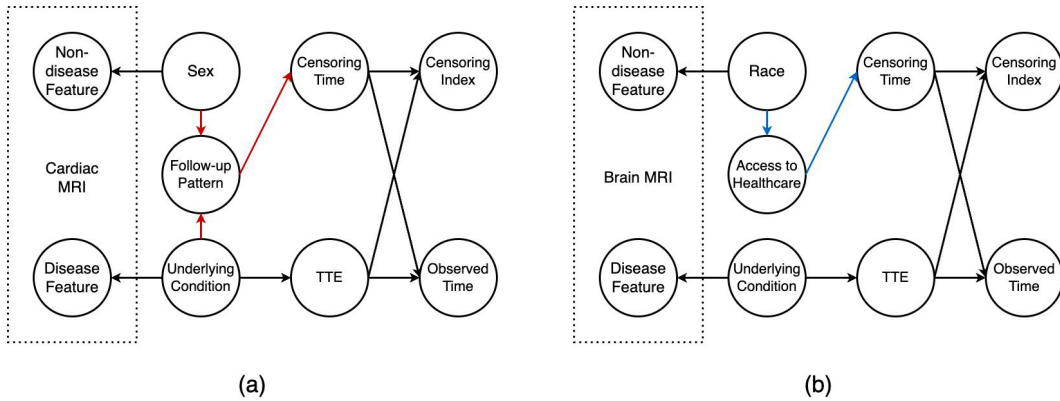


Figure A5: Causal graphs illustrating scenarios where the sensitive attribute  $A$  affects the censoring time  $C$ . (a) Valid pathway: Women may be less likely to receive aggressive follow-up or diagnostic imaging for cardiac conditions, resulting in higher censoring for female patients. (b) Invalid pathway: race appears spuriously correlated with censoring time due to disparities in healthcare access.

## 1458 H Additional Results

### 1459 H.1 Predictive Performance and Fairness in TTE Prediction Models

1460 Figure A6 presents the complete per-group performance results of TTE prediction models—DeepHit,  
 1461 Nnet-Survival, and PMF—across all dataset, sensitive attribute, and metric combinations, while  
 1462 Figure A7 reports the corresponding significance tests using the two-sided Wilcoxon signed-rank test.  
 1463 As shown, performance gaps between groups are observed across all settings.

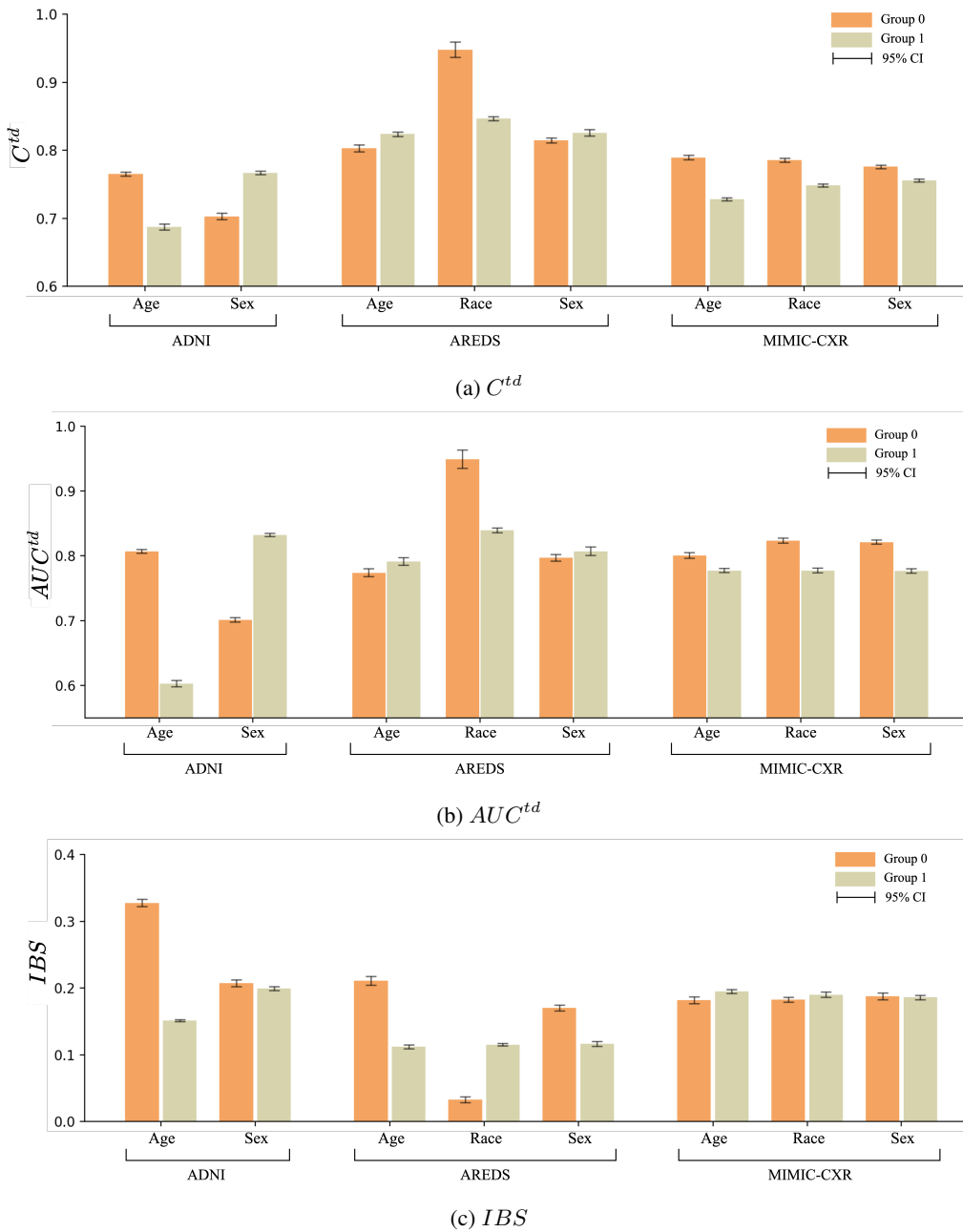


Figure A6: Per-group predictive performances of TTE prediction models across various datasets and sensitive attribute combinations. The visualized performances correspond to the best models determined by model selection conducted on the validation sets. The 95% confidence intervals (CIs) are calculated using bootstrapping over the test sets. a) Results measured by  $C^{td}$ ; b) Results measured by  $AUC^{td}$ ; c) Results measured by  $IBS$ .

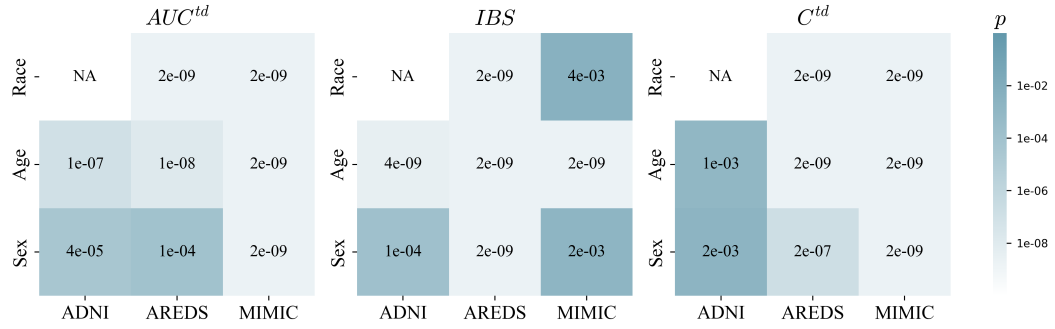


Figure A7: P-values from the two-sided Wilcoxon signed-rank test computed across all TTE prediction models and random seeds. A p-value  $< 0.05$  indicates that there is significant differences in predictive performance between groups.

## 1464 H.2 Comparison between Pre-Training and Training from Scratch Strategies for TTE 1465 Prediction Models

### 1466 H.2.1 Comparison in Predictive Performance

1467 Figure A8 presents the complete per-group predictive performance gap between pre-training and  
1468 training from scratch approaches across all dataset, sensitive attribute, and metric combinations. As  
1469 shown, pre-training consistently improves the predictive performance of TTE models across most  
1470 settings.

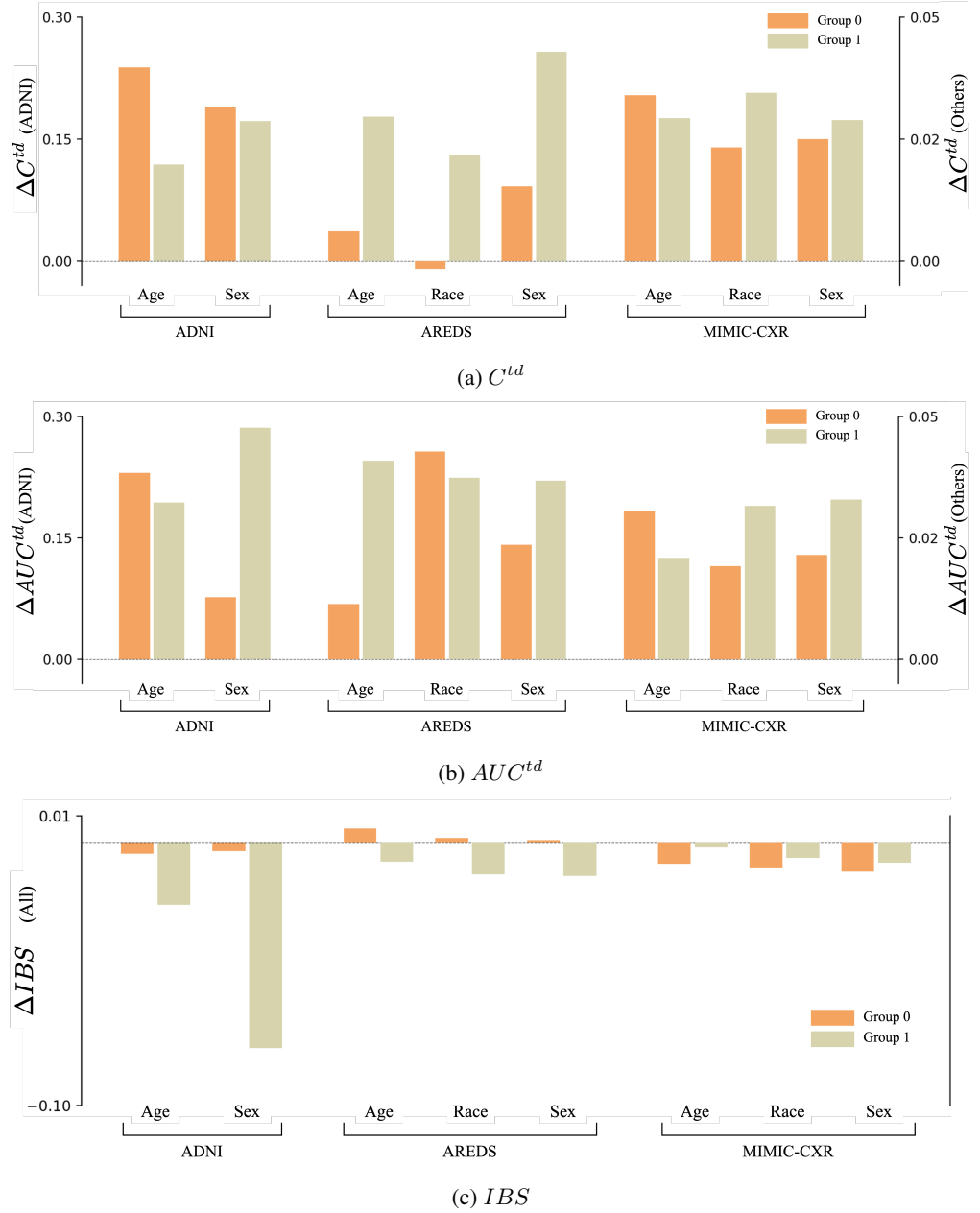


Figure A8: Per-group average performance gap for TTE prediction models using a pre-training strategy compared to training from scratch across various datasets and sensitive attribute combinations. Positive  $\Delta C^{td}$  and  $\Delta AUC^{td}$  and negative  $\Delta IBS$  values indicate that the pre-training strategy enhances predictive performance relative to training from scratch. a) Results measured by  $C^{td}$ ; b) Results measured by  $AUC^{td}$ ; c) Results measured by  $IBS$ .

## H.2.2 Comparison in Fairness

Figure A9 presents the significant differences in terms of fairness between pre-training and training from scratch strategies. As shown, we do not observe a significant improvement with pre-training compared to training from scratch. Specifically, the p-values from one-sided Wilcoxon signed-rank tests are larger than 0.05 in 18 out of 24 settings, suggesting that pre-training does not lead to more equitable predictions in most cases.

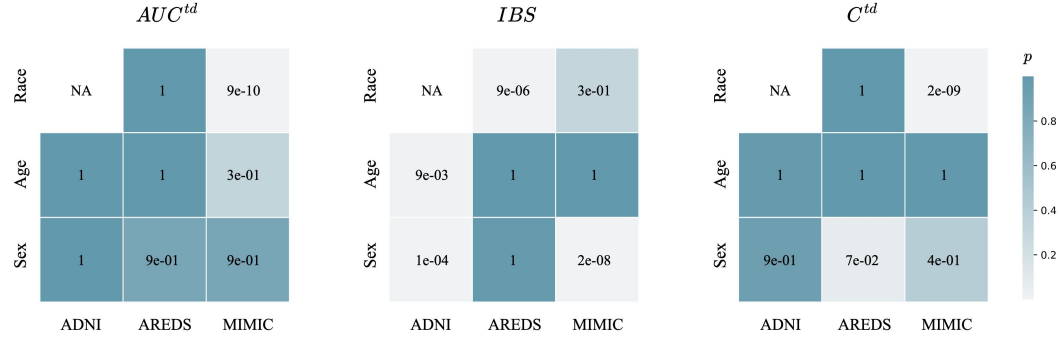


Figure A9: P-values from the one-sided Wilcoxon signed-rank test computed across all TTE prediction models and random seeds. A p-value  $> 0.05$  suggests that pre-training does not result in significantly more equitable predictions compared to training from scratch.

## H.3 Comparison with Advanced Image Backbones and Medical Pre-training for TTE Prediction Models

To investigate whether more advanced vision backbones pretrained on medical imaging data can enhance the predictive performance and fairness of TTE prediction models, we conduct an additional experiment using the late AMD progression prediction task on the AREDS dataset. Specifically, we adopt RETFound—a widely recognized eye-specific foundation model based on the Vision Transformer (ViT) architecture [12] and pretrained on millions of retinal images—as the image backbone for our TTE models. As shown in Table A3, RETFound with ViT does not demonstrate any improvement over EfficientNet pretrained on ImageNet for this task, suggesting that general-purpose backbones may remain competitive despite the availability of domain-specific pretraining.

Table A3: Predictive and fairness performances of DeepHit using EfficientNet and Vision Transformer as vision backbones on AREDS dataset. EfficientNet is pretrained on ImageNet dataset while Vision Transformer is initialized with pretrained weights provided by RETFound.

		EfficientNet			Vision Transformer		
Sensitive Attribute		$AUC^{td} \uparrow$	$IBS \downarrow$	$C^{td} \uparrow$	$AUC^{td} \uparrow$	$IBS \downarrow$	$C^{td} \uparrow$
Accuracy	Age	78.41	15.37	81.30	78.04	15.22	80.65
	Sex	79.08	15.36	81.77	78.01	14.51	80.72
	Race	81.78	11.74	84.53	80.99	11.99	83.91
Sensitive Attribute		$\mathcal{F}_{AUC^{td}} \downarrow$	$\mathcal{F}_{IBS} \downarrow$	$\mathcal{F}_{C^{td}} \downarrow$	$\mathcal{F}_{AUC^{td}} \downarrow$	$\mathcal{F}_{IBS} \downarrow$	$\mathcal{F}_{C^{td}} \downarrow$
Fairness	Age	1.58	12.56	2.20	0.35	9.50	1.44
	Sex	0.76	3.84	1.32	0.85	4.26	0.44
	Race	14.00	10.14	11.09	9.32	10.47	10.27

#### 1487 H.4 Fairness in Fair TTE Prediction Models

1488 Table A4 and Figure A10 present the results of statistical significance testing for fair TTE prediction  
 1489 models, conducted using the Friedman test followed by the Nemenyi post-hoc test.

Table A4: P-values from the Friedman test followed by a Nemenyi post-hoc test computed across all dataset and sensitive attribute combinations. A p-value < 0.05 indicates that the significant difference in terms of fairness between the two corresponding methods.

Metrics	Models	DI	CSA	DRO	DeepHit	FRL	SR
$C^{td}$	DI	1.000	0.995	0.684	0.420	1.000	0.967
	CSA	0.995	1.000	0.340	0.765	0.985	1.000
	DRO	0.684	0.340	1.000	<b>0.011</b>	0.765	0.206
	DeepHit	0.420	0.765	<b>0.011</b>	1.000	0.340	0.894
	FRL	1.000	0.985	0.765	0.340	1.000	0.937
	SR	0.967	1.000	0.206	0.894	0.937	1.000
$AUC^{td}$	DI	1.000	0.206	0.894	0.894	0.340	0.596
	CSA	0.206	1.000	0.836	0.836	1.000	0.985
	DRO	0.894	0.836	1.000	1.000	0.937	0.995
	DeepHit	0.894	0.836	1.000	1.000	0.937	0.995
	FRL	0.340	1.000	0.937	0.937	1.000	0.999
	SR	0.596	0.985	0.995	0.995	0.999	1.000
$IBS$	DI	1.000	0.995	0.985	0.985	0.596	0.894
	CSA	0.995	1.000	1.000	0.836	0.894	0.995
	DRO	0.985	1.000	1.000	0.765	0.937	0.999
	DeepHit	0.985	0.836	0.765	1.000	0.206	0.507
	FRL	0.596	0.894	0.937	0.206	1.000	0.995
	SR	0.894	0.995	0.999	0.507	0.995	1.000



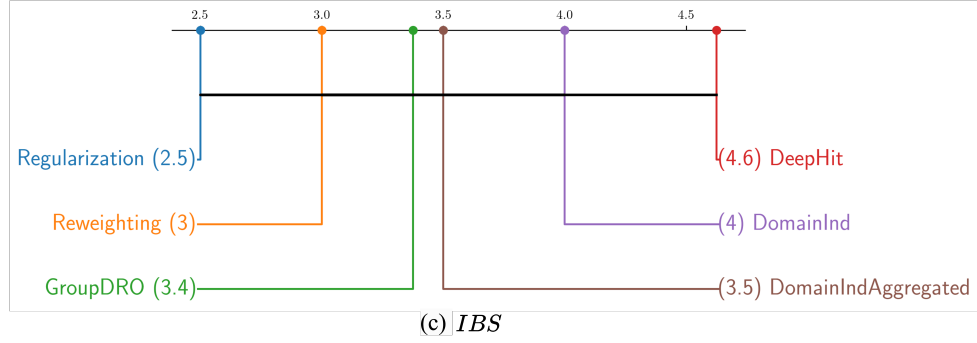
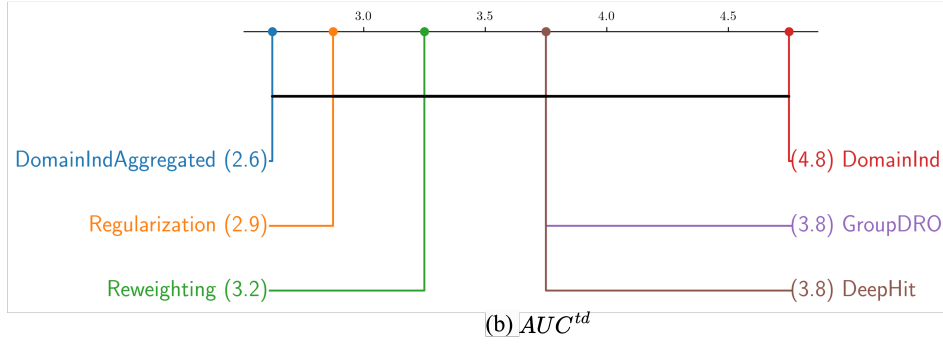
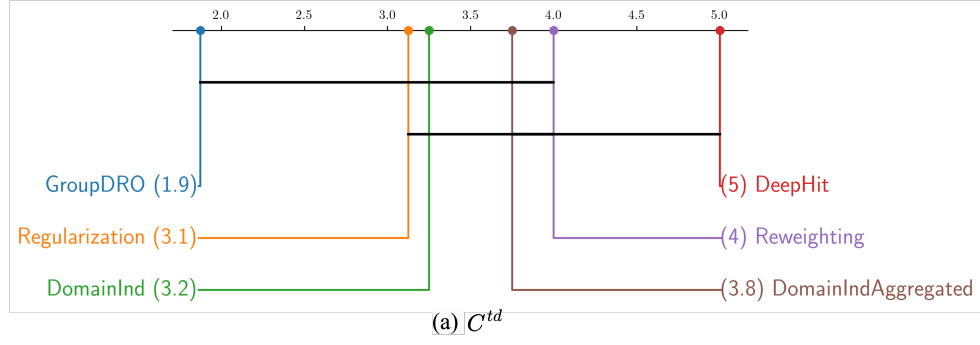


Figure A10: Critical Difference diagrams for all methods calculated from all dataset and sensitive attribute combinations. Although fairness algorithms are generally ranked higher than DeepHit in all settings, there is no significant difference in terms of fairness as indicated by the connections between fairness algorithms and DeepHit in the diagrams. a) Diagram for  $C^{td}$ ; b) Diagram for  $AUC^{td}$ ; c) Diagram for  $IBS$ .

## 1490 H.5 Fairness-Utility Trade-Off Results

1491 Incorporating fairness shifts the objective from pure utility optimization to balancing utility and fair-  
 1492 ness. To assess this trade-off in fair TTE prediction methods, we compute equity scaling scores [44]  
 1493 across datasets and sensitive attributes under both in-distribution and distribution shift scenarios. As  
 1494 shown in Figures A11–A14, different methods exhibit varying fairness-utility trade-offs, with CSA  
 1495 achieving the most favorable balance in most settings.

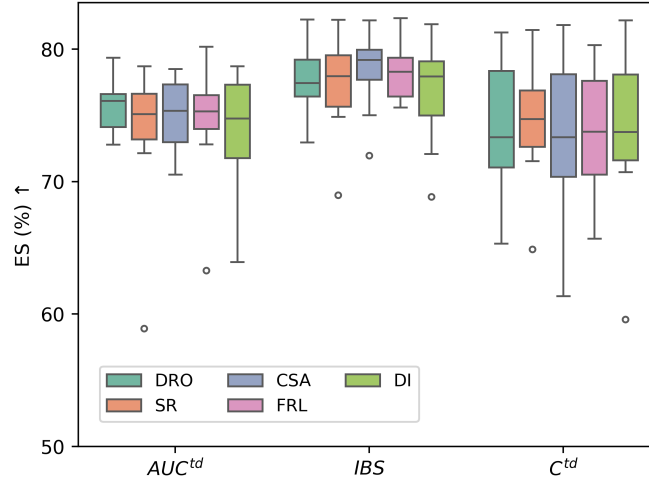


Figure A11: Fairness-utility trade-offs of fairness algorithms for TTE prediction across various utility metrics. For each metric, we compute the corresponding equity scaling score as a measure of the trade-off. The results for each fairness algorithm are aggregated across all dataset and sensitive attribute combinations.

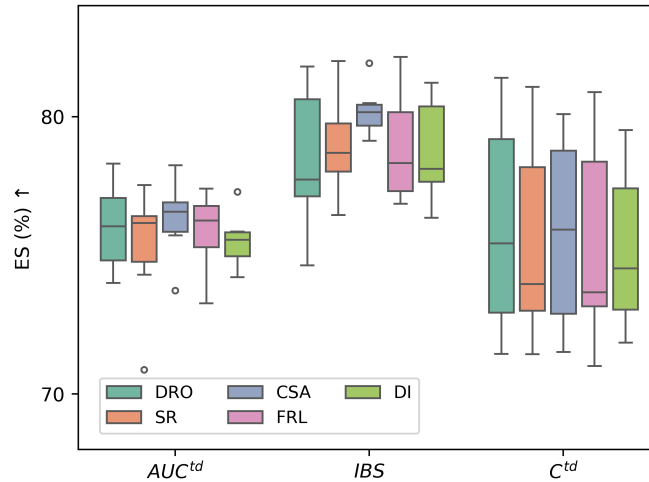


Figure A12: Fairness-utility trade-offs of fairness algorithms for TTE prediction across various utility metrics under distribution shift created by flipping censoring indices (shift on  $X$ ). For each metric, we compute the corresponding equity scaling score as a measure of the trade-off. The results for each fairness algorithm are aggregated across all dataset and sensitive attribute combinations.

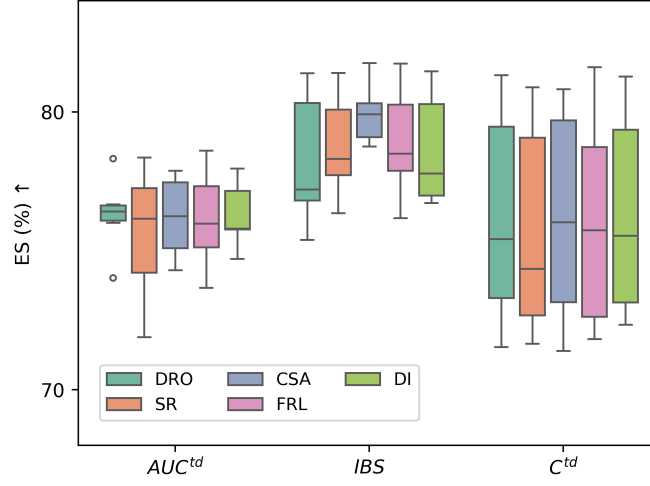


Figure A13: Fairness-utility trade-offs of fairness algorithms for TTE prediction across various utility metrics under distribution shift created by flipping censoring indices (shift on  $Y$ ). For each metric, we compute the corresponding equity scaling score as a measure of the trade-off. The results for each fairness algorithm are aggregated across all dataset and sensitive attribute combinations.

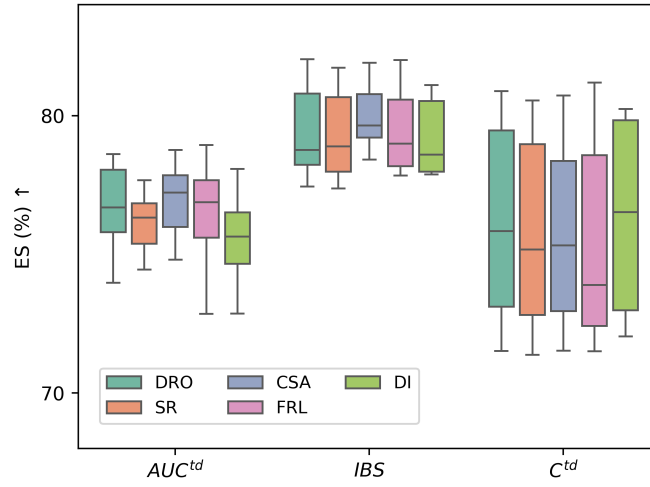


Figure A14: Fairness-utility trade-offs of fairness algorithms for TTE prediction across various utility metrics under distribution shift created by flipping censoring indices (shift on  $\Delta$ ). For each metric, we compute the corresponding equity scaling score as a measure of the trade-off. The results for each fairness algorithm are aggregated across all dataset and sensitive attribute combinations.

## 1496 H.6 Additional Results for Predictive Performance and Fairness in Fair TTE Prediction 1497 Models under Distribution Shift

1498 This section presents the complete results for fair TTE prediction under distribution shift scenarios.  
1499 As illustrated in Figure 3, we define distribution shift as a setting where correlations between the  
1500 sensitive attribute and other variables in the causal graph are present in the training data but absent  
1501 in the testing data. To simulate such shifts, we intervene on one group (the intervened group) by  
1502 corrupting specific aspects of the data—namely, the images ( $X$ ), TTE labels ( $Y$ ), or censoring  
1503 indicators ( $\Delta$ )—while leaving the other group unchanged. Detailed procedures for generating these  
1504 shifts are provided in Appendix F.5.

### 1505 H.6.1 Results for Distribution Shift in $X$

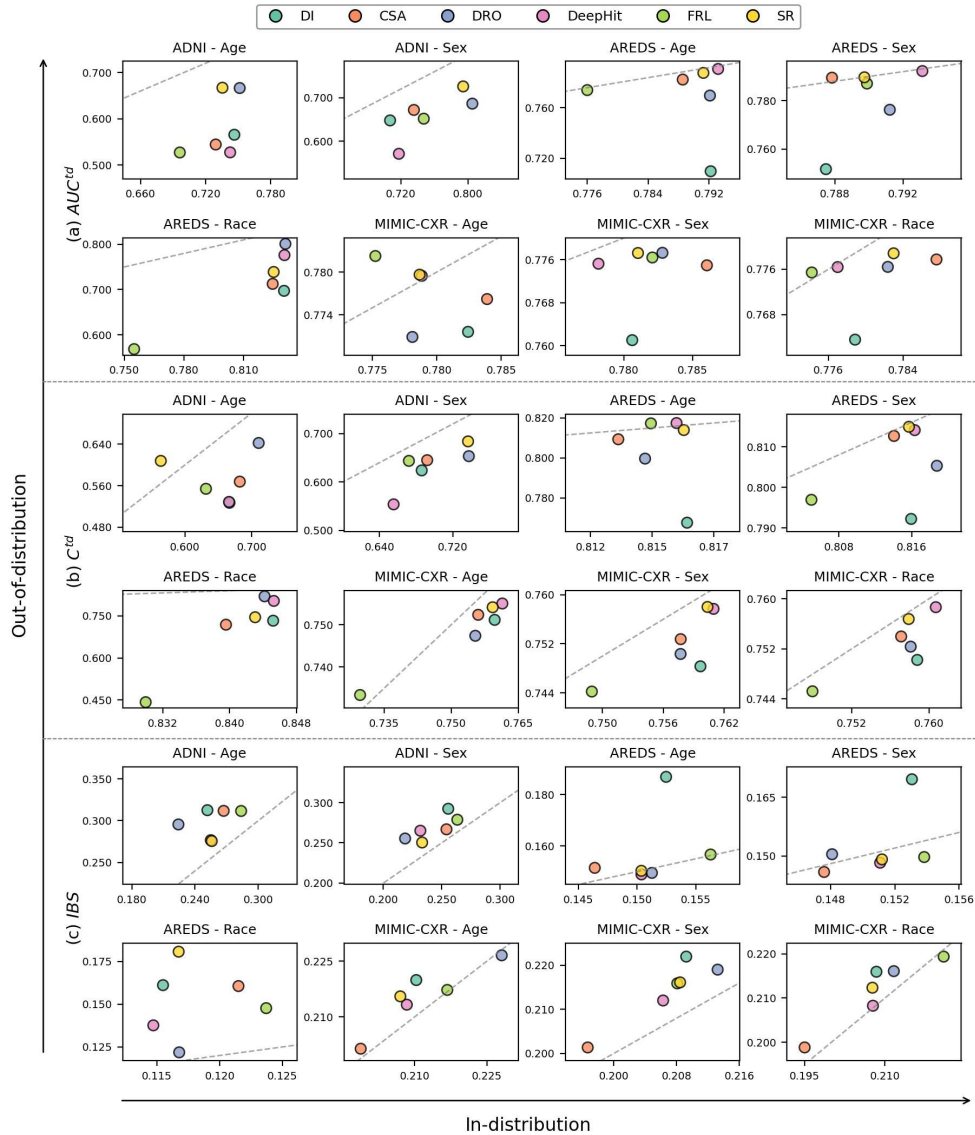


Figure A15: Comparison of predictive performance for (fair) TTE prediction models in in-distribution vs. out-of-distribution (i.e., shift in  $X$ ) learning scenarios, evaluated across all dataset and sensitive attribute combinations. The displayed results represent the average performance across all random seeds. Points on the dashed line indicate equal performance in both scenarios. a) Results for  $AUC^{td}$ ; b) Results for  $C^{td}$ ; c) Results for  $IBS$ .

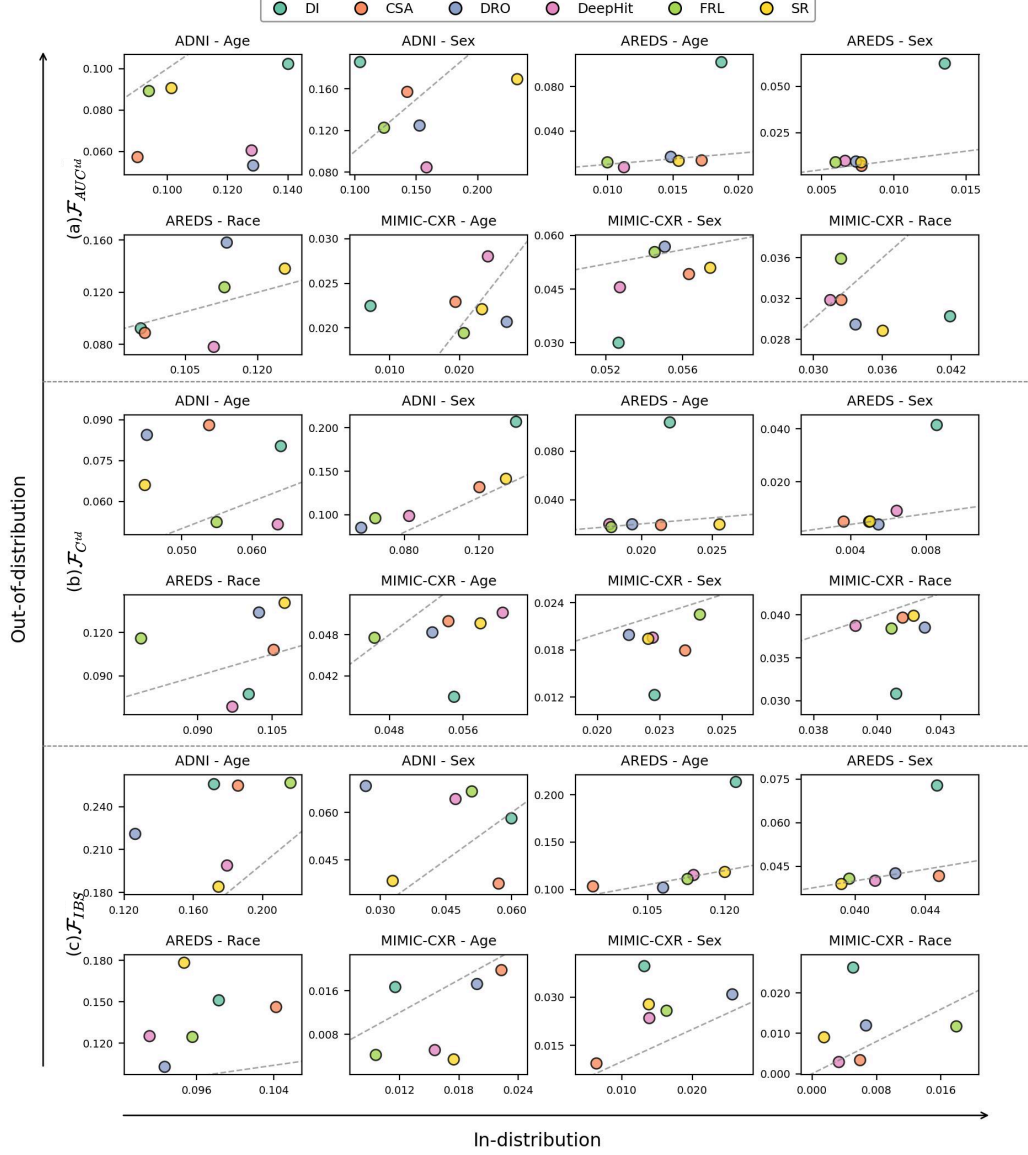


Figure A16: Comparison of fairness for (fair) TTE prediction models in in-distribution vs. out-of-distribution (i.e., shift in  $X$ ) learning scenarios, evaluated across all dataset and sensitive attribute combinations. The displayed results represent the average performance across all random seeds. Points on the dashed line indicate equal performance in both scenarios. a) Results for  $\mathcal{F}_{AUC^{td}}$ ; b) Results for  $\mathcal{F}_{C^{td}}$ ; c) Results for  $\mathcal{F}_{IBS}$ .

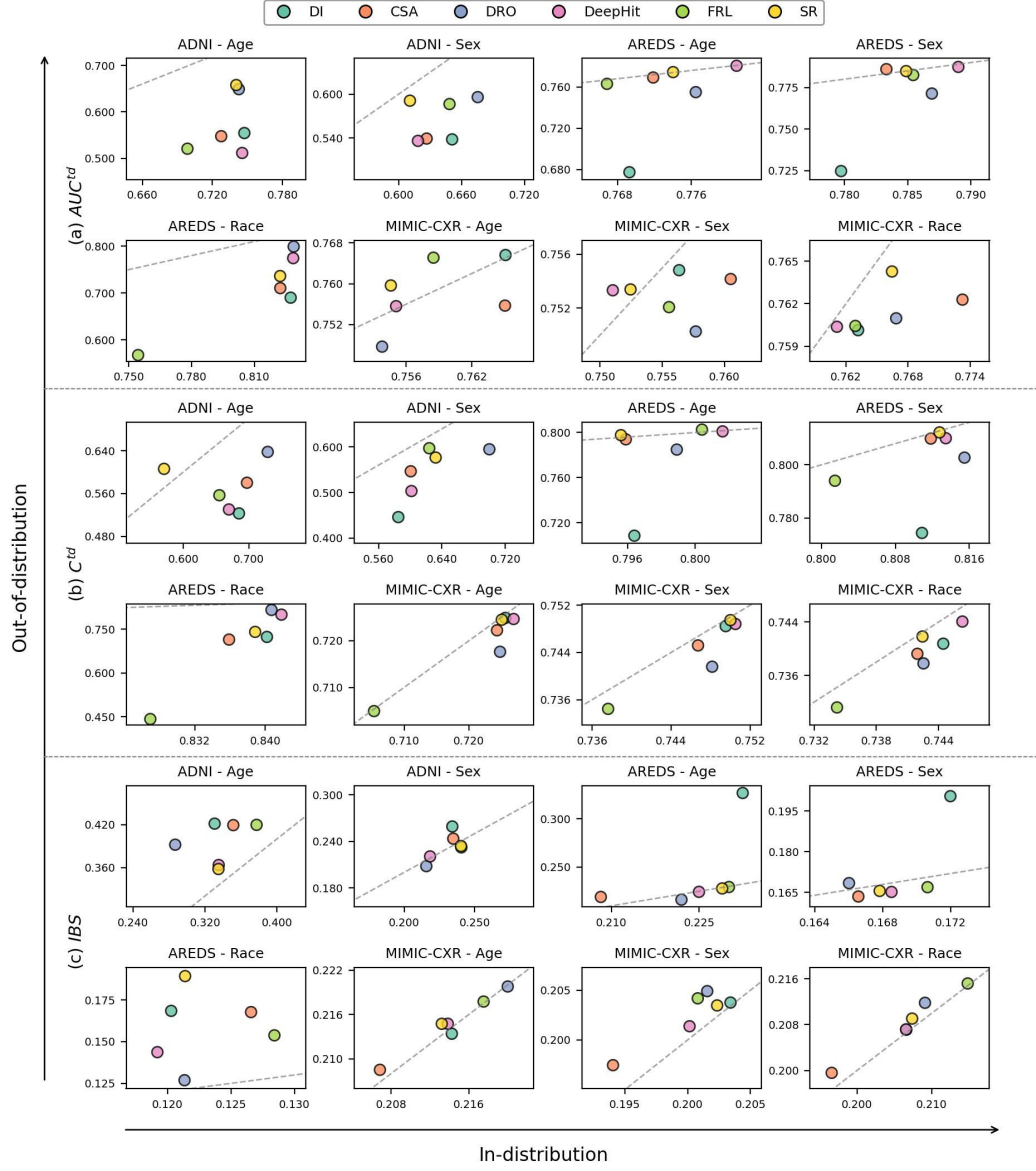


Figure A17: Comparison of predictive performance on the intervened group for (fair) TTE prediction models in in-distribution vs. out-of-distribution (i.e., shift in  $X$ ) learning scenarios across all dataset and sensitive attribute combinations. The displayed results represent the average performance across all random seeds. Points on the dashed line indicate equal performance in both scenarios. a) Results for  $AUC^{td}$ ; b) Results for  $C^{td}$ ; c) Results for  $IBS$ .

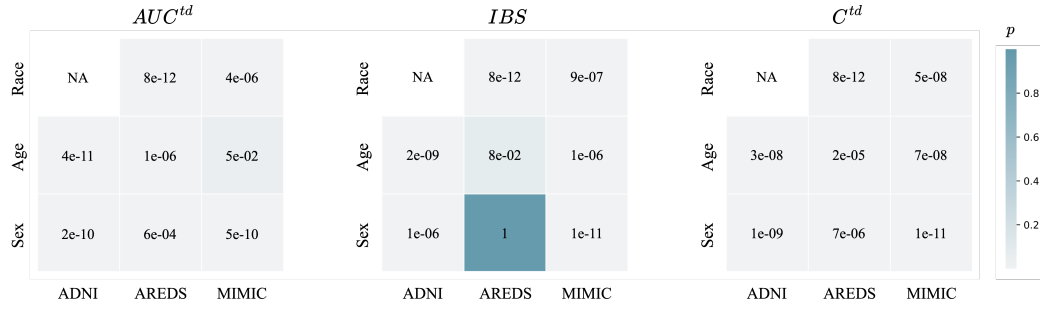


Figure A18: P-values from the one-sided Wilcoxon signed-rank test computed across all fair TTE prediction models and random seeds. A p-value  $< 0.05$  suggests distribution shift on  $X$  significantly degrades TTE predictive performance compared no distribution shift.

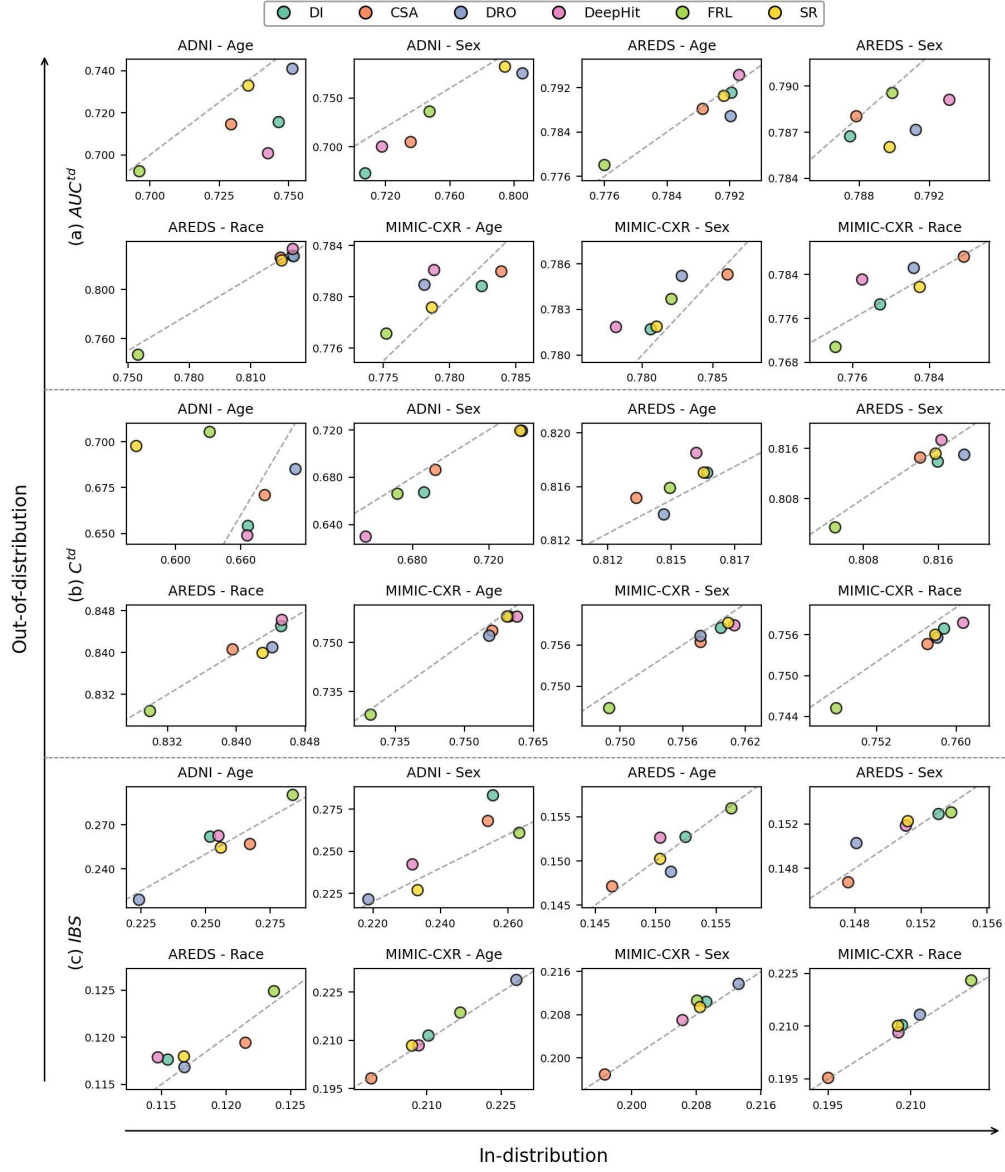


Figure A19: Comparison of predictive performance for (fair) TTE prediction models in in-distribution vs. out-of-distribution (i.e., shift in  $Y$ ) learning scenarios, evaluated across all dataset and sensitive attribute combinations. The displayed results represent the average performance across all random seeds. Points on the dashed line indicate equal performance in both scenarios. a) Results for  $AUC^{td}$ ; b) Results for  $C^{td}$ ; c) Results for  $IBS$ .



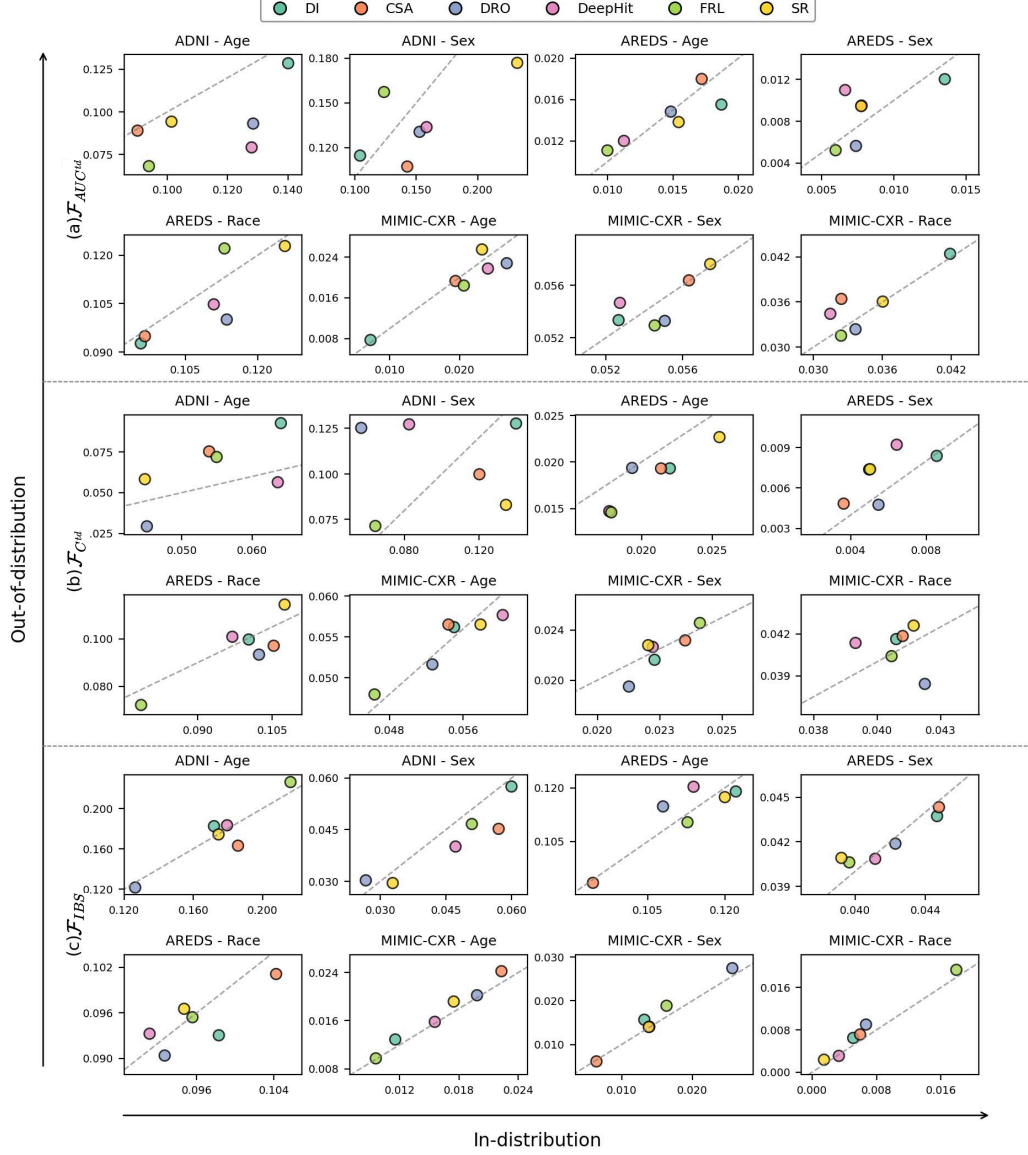


Figure A20: Comparison of fairness for (fair) TTE prediction models in in-distribution vs. out-of-distribution (i.e., shift in  $Y$ ) learning scenarios, evaluated across all dataset and sensitive attribute combinations. The displayed results represent the average performance across all random seeds. Points on the dashed line indicate equal performance in both scenarios. a) Results for  $\mathcal{F}_{AUC^{td}}$ ; b) Results for  $\mathcal{F}_{C^{td}}$ ; c) Results for  $\mathcal{F}_{IBS}$ .

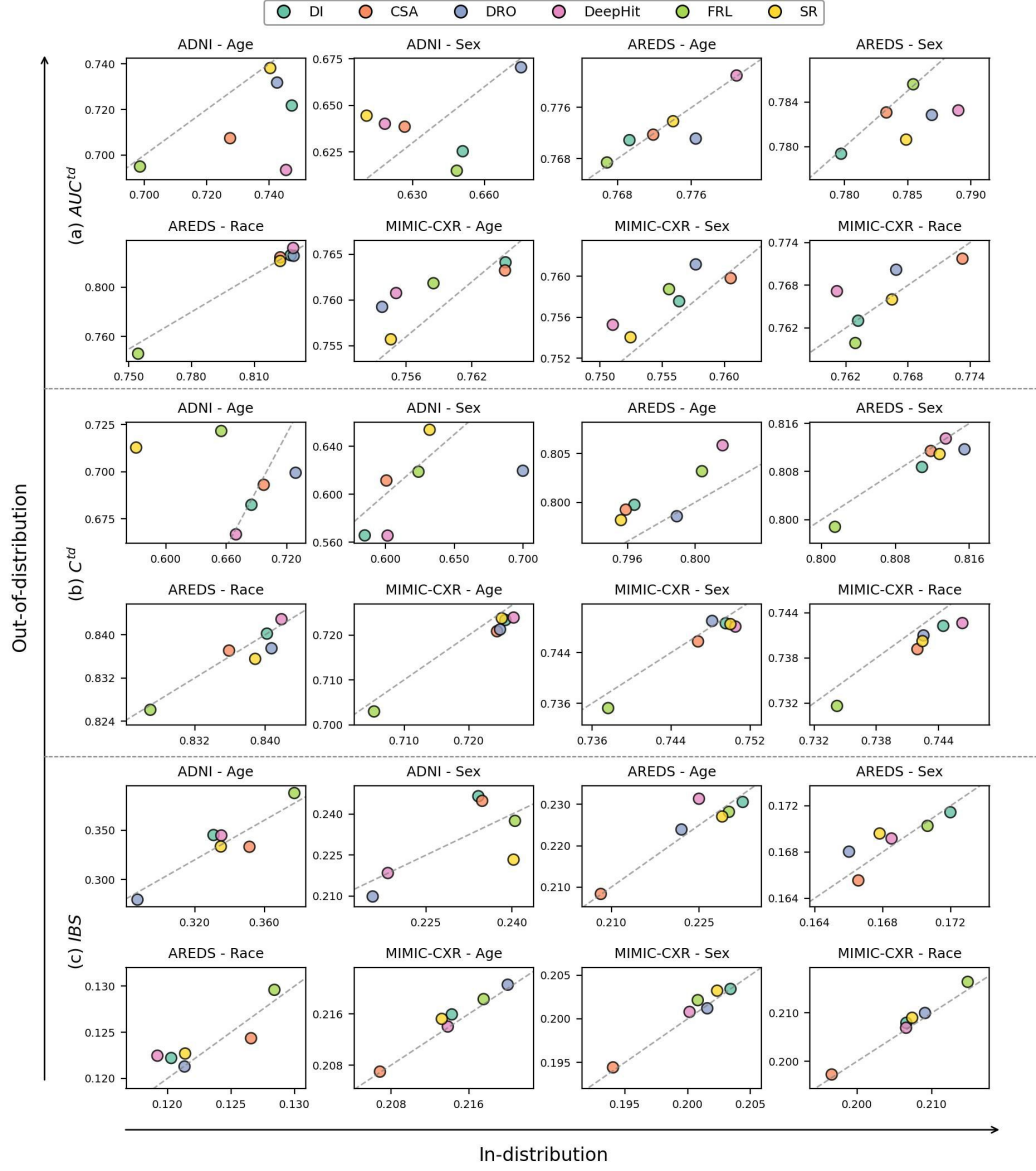


Figure A21: Comparison of predictive performance on the intervened group for (fair) TTE prediction models in in-distribution vs. out-of-distribution (i.e., shift in  $Y$ ) learning scenarios across all dataset and sensitive attribute combinations. The displayed results represent the average performance across all random seeds. Points on the dashed line indicate equal performance in both scenarios. a) Results for  $AUC^{td}$ ; b) Results for  $C^{td}$ ; c) Results for  $IBS$ .

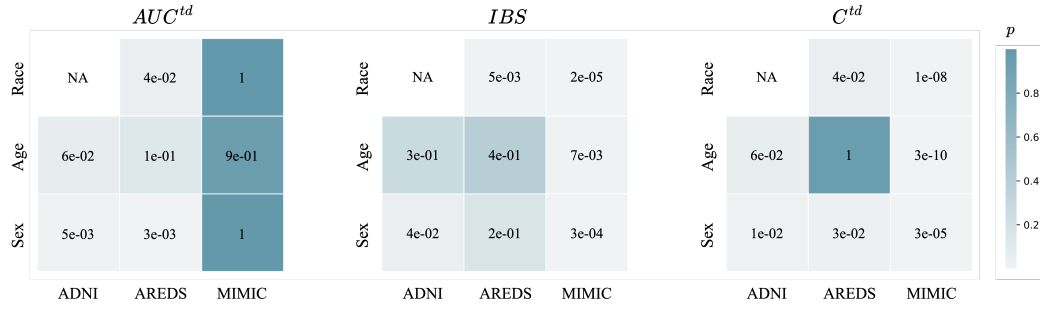


Figure A22: P-values from the one-sided Wilcoxon signed-rank test computed across all fair TTE prediction models and random seeds. A p-value < 0.05 suggests distribution shift on  $Y$  significantly degrades TTE predictive performance compared no distribution shift.

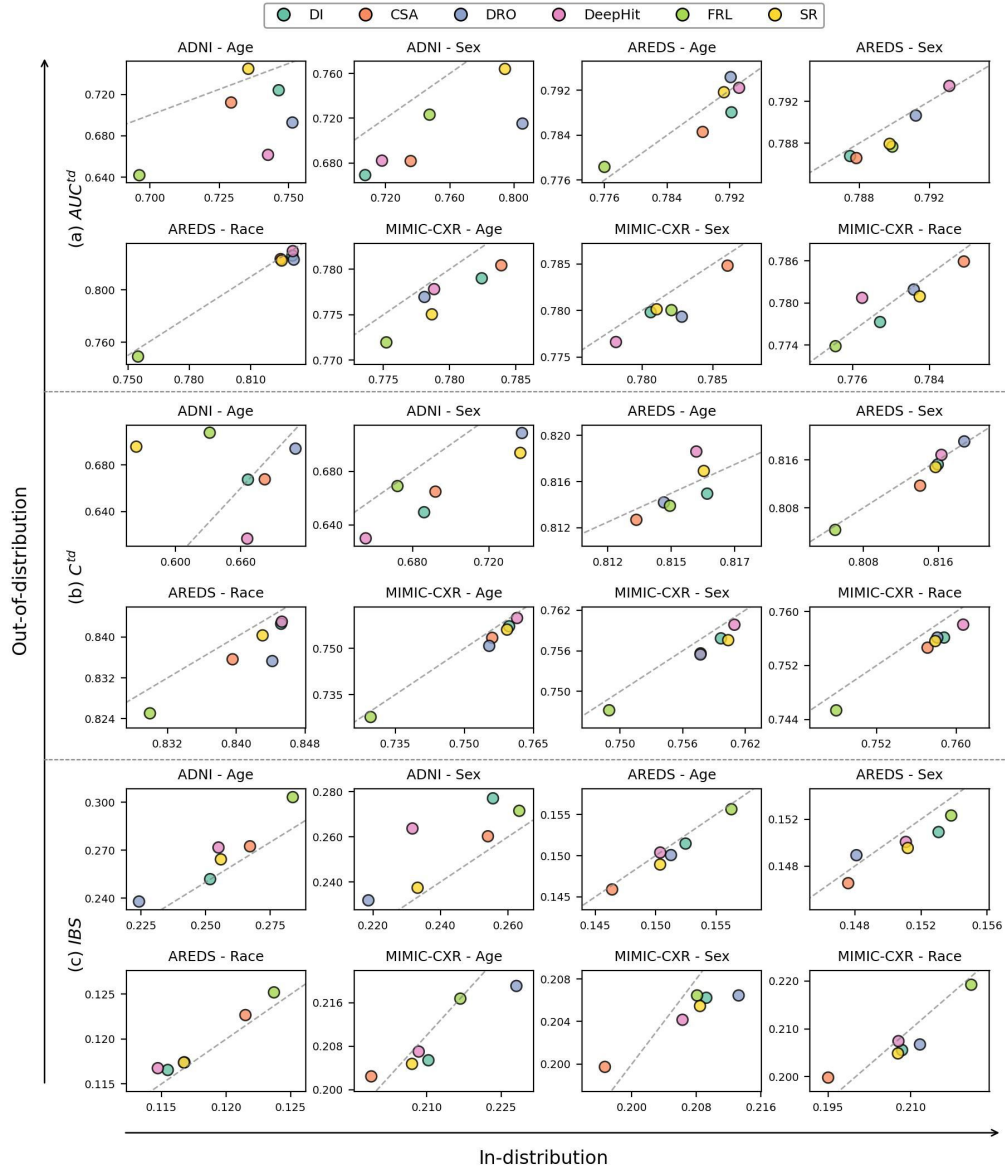


Figure A23: Comparison of predictive performance for (fair) TTE prediction models in in-distribution vs. out-of-distribution (i.e., shift in  $\Delta$ ) learning scenarios, evaluated across all dataset and sensitive attribute combinations. The displayed results represent the average performance across all random seeds. Points on the dashed line indicate equal performance in both scenarios. a) Results for  $AUC^{td}$ ; b) Results for  $C^{td}$ ; c) Results for  $IBS$ .

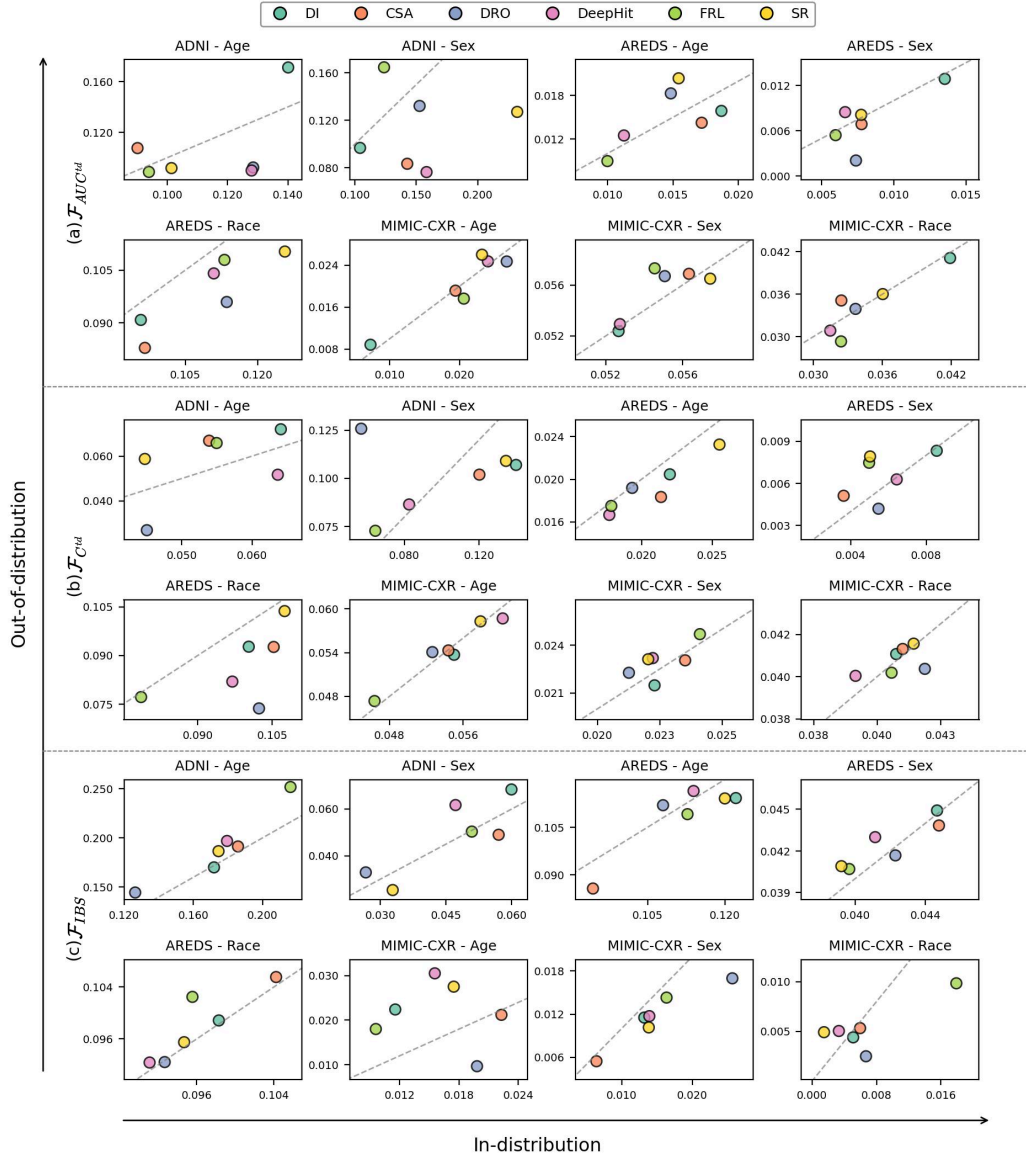


Figure A24: Comparison of fairness for (fair) TTE prediction models in in-distribution vs. out-of-distribution (i.e., shift in  $\Delta$ ) learning scenarios, evaluated across all dataset and sensitive attribute combinations. The displayed results represent the average performance across all random seeds. Points on the dashed line indicate equal performance in both scenarios. a) Results for  $\mathcal{F}_{AUC^{td}}$ ; b) Results for  $\mathcal{F}_{C^{td}}$ ; c) Results for  $\mathcal{F}_{IBS}$ .

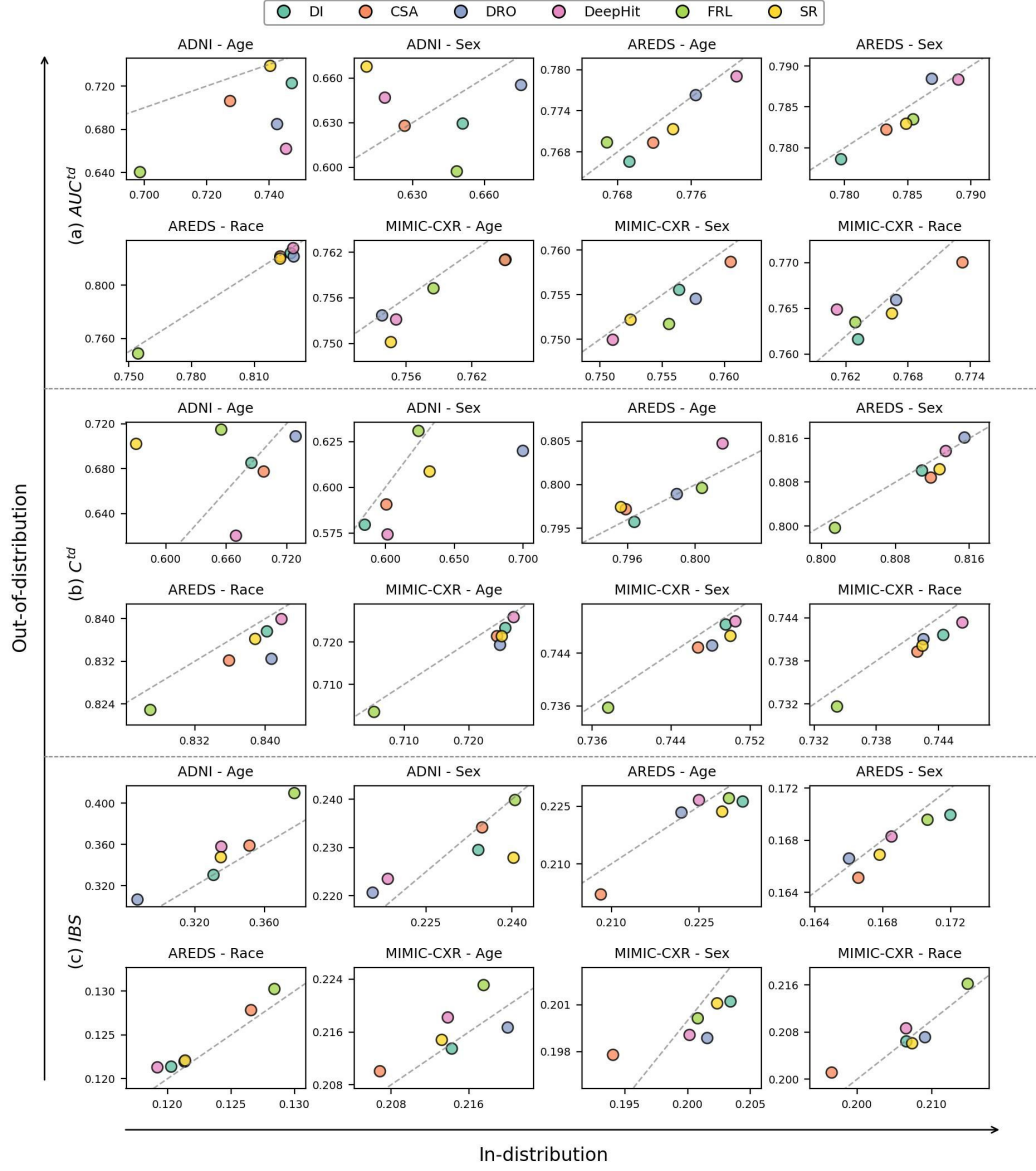


Figure A25: Comparison of predictive performance on the intervened group for (fair) TTE prediction models in in-distribution vs. out-of-distribution (i.e., shift in  $\Delta$ ) learning scenarios across all dataset and sensitive attribute combinations. The displayed results represent the average performance across all random seeds. Points on the dashed line indicate equal performance in both scenarios. a) Results for  $AUC^{td}$ ; b) Results for  $C^{td}$ ; c) Results for  $IBS$ .

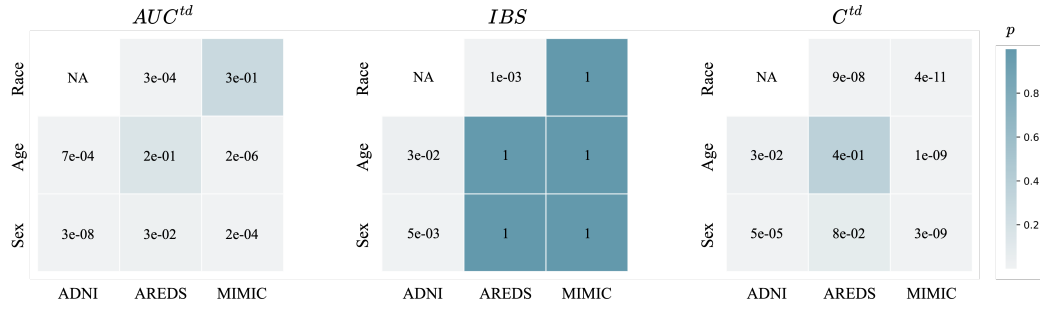


Figure A26: P-values from the one-sided Wilcoxon signed-rank test computed across all fair TTE prediction models and random seeds. A p-value  $< 0.05$  suggests distribution shift on  $\Delta$  significantly degrades TTE predictive performance compared no distribution shift.