

SUPPLEMENTARY MATERIAL FOR ENABLING COUNTERFACTUAL SURVIVAL ANALYSIS WITH BALANCED REPRESENTATIONS

Anonymous authors

Paper under double-blind review

1 GENERAL LOG-LIKELIHOOD

The general likelihood-based loss hypothesis that accounts for informative censoring is formulated as:

$$-\ell_{h,\Phi,\nu}(t_a, c_a, y, \delta) = \log p_{h,\Phi,\nu}(T_A, C_A|X = x) \quad (1)$$

$$= \log p_{h,\Phi}(T_A|X = x) + \log p_{\nu,\Phi}(C_A|X = x), \quad (2)$$

where (2) follows from the conditional independence (informative censoring) assumption $T \perp\!\!\!\perp C|X, A$. For some parametric formulations of event $p_{h,\Phi}(T_A|X = x)$ and censoring $p_{\nu,\Phi}(C_A|X = x)$ time distributions, *e.g.*, exponential, Weibull, log-Normal, *etc.*, then $-\ell_{h,\Phi,\nu}(t_a, c_a, y, \delta)$ is the closed-form log-likelihood, where:

$$\log p_{h,\Phi}(T_A|X = x) \triangleq \delta \cdot \log f_{h,\Phi}(t_a|x) + (1 - \delta) \cdot \log S_{h,\Phi}(t_a|x), \quad (3)$$

$$\log p_{\nu,\Phi}(C_A|X = x) \triangleq (1 - \delta) \cdot \log e_{\nu,\Phi}(c_a|x) + \delta \cdot \log G_{\nu,\Phi}(c_a|x), \quad (4)$$

where $\{S_{h,\Phi}(\cdot), G_{\nu,\Phi}(\cdot)\}$ and $\{f_{h,\Phi}(\cdot), e_{\nu,\Phi}(\cdot)\}$ are survival and density functions respectively.

2 METRICS

2.1 ESTIMANDS OF INTEREST

Several common estimands of interest include (Zhao et al., 2012; Trinquart et al., 2016):

- Difference in expected lifetime: $\text{ITE}(t, x) = \int_0^{t_{\max}} \{S_1(t|x) - S_0(t|x)\} dt = \mathbb{E}\{T_1 - T_0|X = x\}$.
- Difference in survival function: $\text{ITE}(t, x) = S_1(t|x) - S_0(t|x)$.
- Hazard ratio: $\text{ITE}(t, x) = \lambda_1(t|x)/\lambda_0(t|x)$.

In our experiments, we consider both the hazard ratio and difference in expected lifetime. The difference of expected lifetime is expressed in terms of both survival functions and expectations:

$$\begin{aligned} \mathbb{E}[T|X = x] &= \int_{-\infty}^{\infty} t f(t|x) dt \\ &= \int_0^{\infty} (1 - F(t|x)) dt - \int_{-\infty}^0 F(t|x) dt \end{aligned} \quad (5)$$

$$= \int_0^{t_{\max}} S(t|x) dt, \quad (6)$$

where (5) follows from standard properties of expectations and (6) from $1 - F(t|x) = S(t|x)$ and $\int_{-\infty}^0 F(t|x) dt = 0$. Below we formulate an approach for estimating the individualized and population hazard ratio.

2.2 NONPARAMETRIC HAZARD RATIO

From standard survival function definitions (Kleinbaum & Klein, 2010), the relationship between survival and hazard function is formulated as

$$\lambda(t|x) = \lim_{dt \rightarrow 0} \frac{P(t < T < t + dt|X = x)}{P(T > t|X = x)dt} = -\frac{d \log S(t|x)}{dt} = \frac{f(t|x)}{S(t|x)}. \quad (7)$$

We propose a new *nonparametric hazard ratio*, model-free estimator for computing $\text{HR}(t)$.

Definition 1 We define the *nonparametric marginal Hazard Ratio* and its approximation, $\hat{\text{HR}}(t)$, as

$$\text{HR}(t) = \frac{\lambda_1(t)}{\lambda_0(t)} = \frac{S_0(t)}{S_1(t)} \cdot \frac{S'_1(t)}{S'_0(t)}, \quad \hat{\text{HR}}(t) = \frac{\hat{S}_0^{\text{PKM}}(t)}{\hat{S}_1^{\text{PKM}}(t)} \cdot \frac{m_1(t)}{m_0(t)}, \quad (8)$$

where for $\text{HR}(t)$ we leveraged (7) and $S'(t) \triangleq dS(t)/dt$. For the estimator $\hat{\text{HR}}(t)$, provided that $S(t)$ is a monotonically decreasing function, for simplicity, we fit a linear function $S(t) = m \cdot t + c$ and set $S'(t) \approx m$. Further, we leverage $\hat{S}^{\text{PKM}}(t)$ in (Chapfuwa et al., 2020), defined as the model-free population point-estimate-based nonparametric Kaplan-Meier (Kaplan & Meier, 1958) estimator. We denote J distinct and ordered observed event times (censored and non-censored) by the set $\mathcal{T} = \{t_j | t_j > t_{j-1} > \dots > t_0\}$ from N realizations of Y . Formally, the *population survival* $\hat{S}_A^{\text{PKM}}(t)$ is recursively formulated as

$$\hat{S}_A^{\text{PKM}}(t_j) = \left(1 - \frac{\sum_{n:\delta_n=1} \mathbb{I}(t_{j-1} \leq \gamma(T_A^{(n)}) < t_j)}{N - \sum_{n=1}^N \mathbb{I}(\gamma(T_A^{(n)}) < t_{j-1})} \right) \hat{S}_A^{\text{PKM}}(t_{j-1}), \quad (9)$$

where $\hat{S}_A^{\text{PKM}}(t_0) = 1$, and $\mathbb{I}(b)$ represent an indicator function such that $\mathbb{I}(b) = 1$ if b holds or $\mathbb{I}(b) = 0$ otherwise. Further, $\gamma(\cdot)$ is a deterministic transformation for summarizing T_A , in our experiments, $\gamma(\cdot) = \text{median}(\cdot)$, computed over samples from $t_a \sim p_{h,\Phi}(T_A|X=x)$. Note from (9), we marginalize both *factual* and *counterfactual* predictions given covariates x .

A similar formulation for the conditional, *individualized* $\text{HR}(t|x)$, can also be derived, where the cumulative density $F_A(t|x) = 1 - S_A(t|x)$, is estimated with a Gaussian Kernel Density Estimator (KDE) (Silverman, 1986) on samples from the model, $t_a \sim p_{h,\Phi}(T_A|X=x)$. Then we have:

$$\text{HR}(t|x) = \frac{\lambda_1(t|x)}{\lambda_0(t|x)} = \frac{S_0(t|x)}{S_1(t|x)} \cdot \frac{S'_1(t|x)}{S'_0(t|x)}, \quad \hat{\text{HR}}(t|x) = \frac{\hat{S}_0^{\text{KDE}}(t|x)}{\hat{S}_1^{\text{KDE}}(t|x)} \cdot \frac{m_1(t|x)}{m_0(t|x)}, \quad (10)$$

where, $S'(t|x) \triangleq dS(t|x)/dt$ is also approximated with fitting a linear function $S(t|x) = m \cdot t + c$, and setting $S'(t|x) \approx m$. Note that for some parametric formulations, $\text{HR}(t|x)$, can be readily evaluated because $f(t_a|x)$ and $S(t_a|x)$ are available in closed form.

2.3 FACTUAL METRICS

Concordance Index C-Index (also related to receiver operating characteristic) is a widely used survival ranking metric which naturally handles censoring. It quantifies the consistency between the order of the predicted times or risk scores relative to ground truth. C-Index is evaluated on point estimates, we summarize individualized predicted samples from CSA and CSA-INFO, *i.e.*, $t_a = \text{median}(\{t_s\}_{s=1}^{200})$, where t_s is a sample from the trained model.

Calibration Slope Calibration quantifies distributional statistical consistency between model predictions relative to ground truth. We measure *population* calibration by comparing population survival curves from model predictions against ground truth according to Chapfuwa et al. (2020). We desire a high calibrated model, with calibration slope of 1, while a slope < 1 and slope > 1 indicates underestimation or overestimation risk, respectively.

Coefficient of Variation The coefficient of variation (COV) $\sigma\mu^{-1}$, the ratio between standard deviation and mean, quantifies distribution dispersion. A COV > 1 and < 1 indicates a high or low variance distribution, in practice, we desire low variance distribution. We use Mean COV $N^{-1} \sum_{i=1}^N \sigma_i \mu_i^{-1}$, where for subject i we compute $\{\mu_i, \sigma_i\}$ from samples $\{t_s\}_{s=1}^{200}$.

3 BASELINES

Cox proportional hazard (CoxPH) CoxPH assumes a semi-parametric linear model $\lambda(t|a) = \lambda_b(t) \exp(a\beta)$, thus the hazard ratio between treatment and control can be obtained without specifying

ing the baseline hazard $\lambda_b(t)$:

$$\text{HR}(t) = \frac{\lambda(t|a=1)}{\lambda(t|a=0)} = \exp(\beta). \quad (11)$$

A simple logistic model $\hat{e}_i = \sigma(x_i; \eta)$, is used to approximate the unknown propensity score $P(A=1|X=x)$. Methods that adjust for selection bias (or confounding) learn β by maximizing a propensity weighted partial likelihood (Schemper et al., 2009; Buchanan et al., 2014; Rosenbaum & Rubin, 1983)

$$\mathcal{L}(\beta) = \prod_{i:\delta_i=1} \left(\frac{\exp(a_i\beta)}{\sum_{j:t_j \geq t_i} \hat{w}_j \cdot \exp(a_j\beta)} \right)^{\hat{w}_i}. \quad (12)$$

We consider three normalized weighting schemes for w , namely, (i) inverse probability weighting (IPW) (Horvitz & Thompson, 1952; Cao et al., 2009), where $\text{IPW}_i = \frac{a_i}{\hat{e}_i} + \frac{1-a_i}{1-\hat{e}_i}$, (ii) overlapping weights (OW) (Crump et al., 2006; Li et al., 2018), where $\text{OW}_i = a_i \cdot (1 - \hat{e}_i) + (1 - a_i) \cdot \hat{e}_i$, and (iii) the standard RCT Uniform assumption. Note that this modeling approach requires fitting over the entire dataset, thus has no inference capability.

Accelerated Failure Time (AFT) We implement IPM regularized neural-based log-Normal and Weibull AFT baselines. Both approaches have a desirable closed form $S_{h,\Phi}(t_a|x)$, thus enabling maximum likelihood based estimation, where

$$-\mathcal{L}_F^{\text{AFT}} \triangleq \mathbb{E}_{(y,\delta,x,a) \sim p(y,\delta,X,A)} [\delta \cdot \log f_{h,\Phi}(t_a|x) + (1 - \delta) \cdot \log S_{h,\Phi}(t_a|x)]. \quad (13)$$

The log-Normal mean and variance parameters are learned such that, $\log t_a = \mu_{h,\Phi}(h(r,a)) + \epsilon$, where $\epsilon \sim \mathcal{N}(0, \sigma_{h,\Phi}^2(h(r,a)))$ and $r = \Phi(x)$. Further, we learn the Weibull scale and shape parameters, where $t_a = \lambda_{h,\Phi}(h(r,a)) \cdot (-\log U)^{(k_{h,\Phi}(h(r,a)))^{-1}}$ and $U \sim \text{Uniform}(0,1)$. We regularize (13) with the IPM loss, for maximum likelihood optimization.

Semi-supervised regression (SR) To demonstrate the effectiveness of our flow-based uncertainty estimation approach we contrast CSA with a deterministic accuracy objective from Chapfuwa et al. (2018), where $t_a = h(r,a)$ and:

$$\mathcal{L}_F^{\text{SR}} \triangleq \mathbb{E}_{(y,\delta,x,a) \sim p(y,\delta,X,A)} [\delta \cdot (|y - t_a|) + (1 - \delta) \cdot (\max(0, y - t_a))], \quad (14)$$

where (14) is regularized according to the IPM loss.

Survival Bayesian additive regression trees (Surv-BART) Surv-BART (Sparapani et al., 2016) is a nonparametric tree-based approach for estimating individualized survivals $\hat{S}(t_a^{(j)}|X=x)$ (defined at pre-specified J time-horizons) from an ensemble of regression trees. Note, Surv-BART does not adjust for both selection bias and informative censoring. While, we fit two separate models based on factual treatment and control data, causal metrics are estimated with both *factual* and *counterfactual* predictions.

4 EXPERIMENTS

4.1 GENERATING ATCG-SYNTHETIC DATASET

The ACTG-SYNTHETIC, is a semi-synthetic dataset based on ACTG covariates (Hammer et al., 1996). We simulate potential outcomes according to a Gompertz-Cox distribution (Bender et al., 2005) with selection bias from a simple logistic model for $P(A=1|X=x)$ and AFT-based censoring

Table 1: Performance comparisons on ACTG data, with 95% $HR(t)$ confidence interval. Test set NN assignment of y_{CF} and δ_{CF} yields unbiased ground truth estimator $HR(t) = 0.54_{(0.51, 0.61)}$, since study is a RCT.

Method	Causal	Factual		
	$HR(t)$	C-Index (A=0, A=1)	Mean COV	C-Slope (A=0, A=1)
CoxPH-Uniform	0.49 _(0.38, 0.64)	NA	NA	NA
CoxPH-IPW	0.49 _(0.36, 0.68)	NA	NA	NA
CoxPH-OW	0.49 _(0.36, 0.68)	NA	NA	NA
Surv-BART	3.93 _(3.93, 4.90)	(0.665, 0.845)	0.001	(0.394, 0.517)
AFT-Weibull	0.53 _(0.53, 0.53)	(0.53, 0.351)	3.088	(0.847, 0.813)
AFT-log-Normal	3.75 _(3.75, 3.75)	(0.717, 0.619)	7.995	(0.847, 0.321)
SR	0.21 _(0.21, 0.28)	(0.628, 0.499)	0	(1.388, 0.442)
CSA (proposed)	0.63 _(0.59, 0.68)	(0.831, 0.814)	0.132	(1.042, 1.129)
CSA-INFO (proposed)	0.6 _(0.54, 0.66)	(0.786, 0.822)	0.13	(0.875, 0.938)

mechanism. Below is our generative scheme:

$$\begin{aligned}
 X &= \text{ACTG covariates} \\
 P(A = 1|X = x) &= \frac{1}{b} \times (a + \sigma(\eta(\text{AGE} - \mu_{\text{AGE}} + \text{CD40} - \mu_{\text{CD40}}))) \\
 T_A &= \frac{1}{\alpha_A} \log \left[1 - \frac{\alpha_A \log U}{\lambda_A \exp(x^T \beta_A)} \right], \quad U \sim \text{Uniform}(0, 1) \\
 \log C &\sim \text{Normal}(\mu_c, \sigma_c^2) \\
 Y &= \min(T_A, C), \quad \delta = 1 \text{ if } T_A < C, \text{ else } \delta = 0,
 \end{aligned}$$

where $\{\beta_A, \alpha_A, \lambda_A, b, a, \eta, \mu_c, \sigma_c\}$ are hyper-parameters and $\{\mu_{\text{AGE}}, \mu_{\text{CD40}}\}$ are the means for age and CD40 respectively. This semi-synthetic dataset will be made publicly available.

4.2 QUANTITATIVE RESULTS

See Table 1 for additional quantitative comparisons on ACTG dataset.

4.3 QUALITATIVE RESULTS

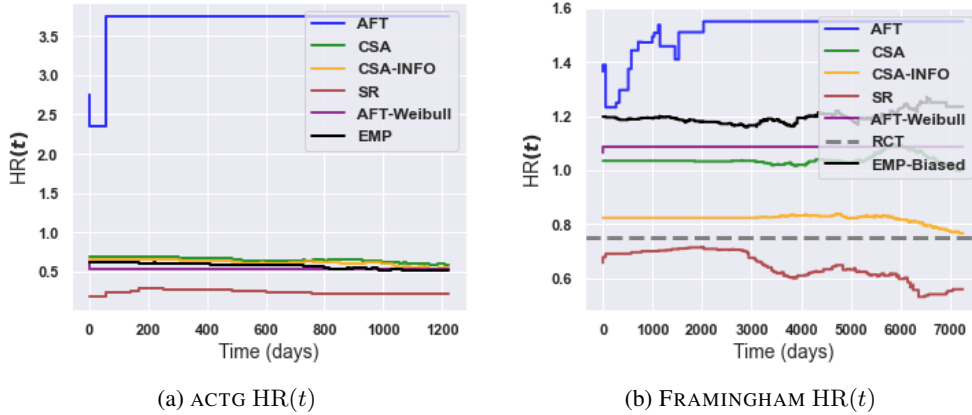


Figure 1: Inferred population $HR(t)$ comparisons on (a) ACTG and (b) FRAMINGHAM datasets.

Figure 1 demonstrates model comparisons across of population hazard, $HR(t)$, on ACTG and FRAMINGHAM datasets. Figure 2, summarizes the positive and negative covariate statistics from the isolated extreme top and bottom quantiles on FRAMINGHAM datasets.

	age6	ascvd_hx6	bmi6	bpmeds6	choi5	dbp6	diab6	female
count	129.000000	129.000000	129.000000	129.000000	129.000000	129.000000	129.000000	129.000000
mean	55.558140	0.069767	29.326328	0.317829	202.147287	75.922481	0.116279	0.465116
std	9.412348	0.255748	4.800124	0.467448	40.368053	7.184632	0.321809	0.500726
min	35.000000	0.000000	20.777429	0.000000	118.000000	55.000000	0.000000	0.000000
25%	50.000000	0.000000	25.995640	0.000000	176.000000	71.000000	0.000000	0.000000
50%	54.000000	0.000000	28.835150	0.000000	196.000000	76.000000	0.000000	0.000000
75%	61.000000	0.000000	31.847777	1.000000	225.000000	80.000000	0.000000	1.000000
max	84.000000	1.000000	45.135681	1.000000	312.000000	100.000000	1.000000	1.000000

	gluc5	hdl5	pad_hx6	sbp6	smoke6	stk_hx6	mi_hx6	trigly5
count	129.000000	129.000000	129.000000	129.000000	129.000000	129.000000	129.000000	129.000000
mean	102.201550	43.953488	0.007752	123.782946	0.170543	0.015504	0.038760	164.139535
std	34.450912	11.543979	0.088045	13.923879	0.377575	0.124027	0.193774	78.358625
min	75.000000	26.000000	0.000000	99.000000	0.000000	0.000000	0.000000	46.000000
25%	90.000000	35.000000	0.000000	114.000000	0.000000	0.000000	0.000000	119.000000
50%	95.000000	43.000000	0.000000	122.000000	0.000000	0.000000	0.000000	143.000000
75%	103.000000	50.000000	0.000000	131.000000	0.000000	0.000000	0.000000	200.000000
max	289.000000	95.000000	1.000000	170.000000	1.000000	1.000000	1.000000	468.000000

	age6	ascvd_hx6	bmi6	bpmeds6	choi5	dbp6	diab6	female
count	129.000000	129.000000	129.000000	129.000000	129.000000	129.000000	129.000000	129.000000
mean	60.666667	0.139535	26.375823	0.286822	197.581395	71.627907	0.046512	0.604651
std	10.185263	0.347855	5.325558	0.454041	28.029197	11.657260	0.211411	0.490832
min	37.000000	0.000000	17.676632	0.000000	118.000000	49.000000	0.000000	0.000000
25%	53.000000	0.000000	22.687889	0.000000	180.000000	62.000000	0.000000	0.000000
50%	60.000000	0.000000	25.285077	0.000000	198.000000	70.000000	0.000000	1.000000
75%	69.000000	0.000000	29.230393	1.000000	217.000000	81.000000	0.000000	1.000000
max	78.000000	1.000000	45.992112	1.000000	290.000000	105.000000	1.000000	1.000000

	gluc5	hdl5	pad_hx6	sbp6	smoke6	stk_hx6	mi_hx6	trigly5
count	129.000000	129.000000	129.000000	129.000000	129.000000	129.000000	129.000000	129.000000
mean	96.023256	58.550388	0.054264	128.007752	0.217054	0.046512	0.054264	119.937984
std	16.391912	16.147253	0.227420	22.184417	0.413847	0.211411	0.227420	133.107261
min	48.000000	22.000000	0.000000	88.000000	0.000000	0.000000	0.000000	33.000000
25%	88.000000	49.000000	0.000000	111.000000	0.000000	0.000000	0.000000	63.000000
50%	95.000000	59.000000	0.000000	126.000000	0.000000	0.000000	0.000000	87.000000
75%	101.000000	69.000000	0.000000	140.000000	0.000000	0.000000	0.000000	109.000000
max	228.000000	101.000000	1.000000	214.000000	1.000000	1.000000	1.000000	1149.000000

(a) FRAMINGHAM $HR(t|x) < 0.024$ (b) FRAMINGHAM $HR(t|x) > 1.916$ Figure 2: Covariate statistics for top (a) and bottom (b) quantiles, of the median log $HR(t|x)$ values for the test set of FRAMINGHAM.

4.4 ARCHITECTURE OF THE NEURAL NETWORK

We detail the architecture of neural-based methods, namely, baselines (AFT-log-Normal, AFT-Weibull, SR) and our proposed methods (CSA and CSA-INFO). All methods are trained using one NVIDIA P100 GPU with 16GB memory. In all experiments we set the minibatch size $M = 200$, Adam optimizer with the following hyperparameters: learning rate 3×10^{-4} , first moment 0.9, second moment 0.99, and epsilon 1×10^{-8} . Further, all network weights are initialed according to $\text{Uniform}(-0.01, 0.01)$. Datasets are split into training, validation and test sets according to 70%, 15% and 15% partitions, respectively, stratified by event and treatment proportions. The validation set is used for hyperparameter search and early stopping. All hidden units in $\{h_A(\cdot), \nu_A(\cdot)\}$, are characterized by Leaky Rectified Linear Unit (ReLU) activation functions, batch normalization and dropout probability of $p = 0.2$ on all layers. The output layers of predicted times $\{T_A, C_A\}$ have an additional exponential transformation.

Encoder The encoding function $\Phi(\cdot)$ for mapping $r = \Phi(x)$ is shared among all the neural based methods (proposed and baselines) and specified in terms of two-layer MLPs of 100 hidden units.

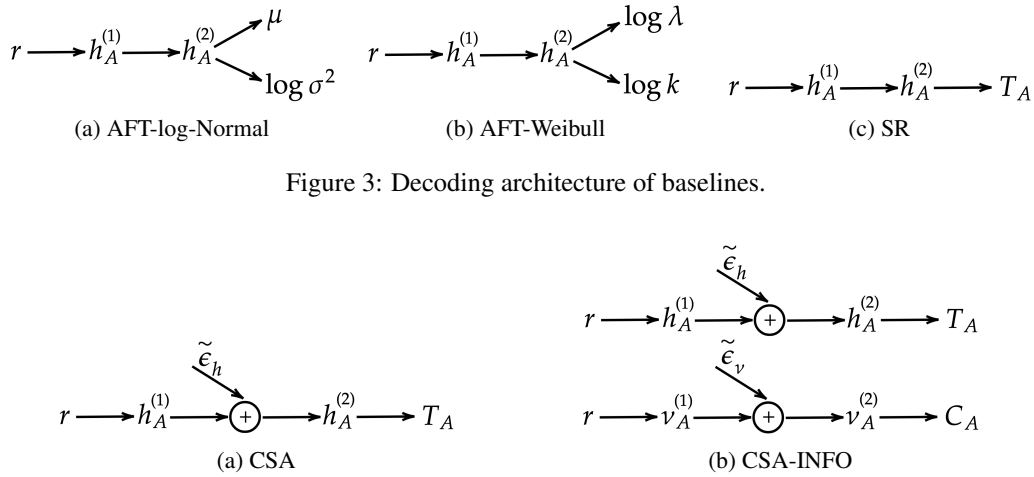


Figure 3: Decoding architecture of baselines.

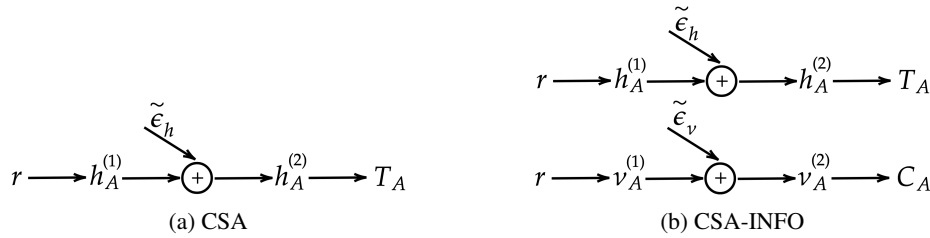


Figure 4: Decoding architecture of proposed methods.

Decoder Figure 3 shows the architectural details of the baselines, where the decoding function $h_A(\cdot)$ is specified in terms of two-layer MLPs of 100 hidden units. Further, the proposed *planar* flow based methods shown in Figure 4, are comprised of two-layer MLPs for $\{h_A(\cdot), \nu_A(\cdot)\}$ of

dimensions $[100, 200]$. Moreover, the hidden layers $\{h_A^{(2)}, \nu_A^{(2)}\}$, take as input the concatenated $[h_A^{(1)}, \tilde{\epsilon}_h]$ and $[\nu_A^{(1)}, \tilde{\epsilon}_\nu]$ respectively. Finally, we set the planar flow dimensions for both $\{\tilde{\epsilon}_\nu, \tilde{\epsilon}_h\}$ to 100.

REFERENCES

- Ralf Bender, Thomas Augustin, and Maria Blettner. Generating survival times to simulate cox proportional hazards models. *Statistics in medicine*, 2005.
- Ashley L Buchanan, Michael G Hudgens, Stephen R Cole, Bryan Lau, Adaora A Adimora, and Women’s Interagency HIV Study. Worth the weight: using inverse probability weighted cox models in aids research. *AIDS research and human retroviruses*, 2014.
- Weihua Cao, Anastasios A Tsiatis, and Marie Davidian. Improving efficiency and robustness of the doubly robust estimator for a population mean with incomplete data. *Biometrika*, 2009.
- P. Chapfuwa, C. Tao, C. Li, I. Khan, K. J. Chandross, M. J. Pencina, L. Carin, and R. Henao. Calibration and uncertainty in neural time-to-event modeling. *IEEE Transactions on Neural Networks and Learning Systems*, 2020.
- Paidamoyo Chapfuwa, Chenyang Tao, Chunyuan Li, Courtney Page, Benjamin Goldstein, Lawrence Carin, and Ricardo Henao. Adversarial time-to-event modeling. In *ICML*, 2018.
- Richard K Crump, V Joseph Hotz, Guido W Imbens, and Oscar A Mitnik. Moving the goalposts: Addressing limited overlap in the estimation of average treatment effects by changing the estimand. Technical report, National Bureau of Economic Research, 2006.
- Scott M Hammer, David A Katzenstein, Michael D Hughes, Holly Gundacker, Robert T Schooley, Richard H Haubrich, W Keith Henry, Michael M Lederman, John P Phair, Manette Niu, et al. A trial comparing nucleoside monotherapy with combination therapy in hiv-infected adults with cd4 cell counts from 200 to 500 per cubic millimeter. *New England Journal of Medicine*, 1996.
- Daniel G Horvitz and Donovan J Thompson. A generalization of sampling without replacement from a finite universe. *Journal of the American statistical Association*, 1952.
- Edward L Kaplan and Paul Meier. Nonparametric estimation from incomplete observations. *Journal of the American statistical association*, 1958.
- David G Kleinbaum and Mitchel Klein. *Survival analysis*. Springer, 2010.
- Fan Li, Kari Lock Morgan, and Alan M Zaslavsky. Balancing covariates via propensity score weighting. *Journal of the American Statistical Association*, 2018.
- Paul R Rosenbaum and Donald B Rubin. The central role of the propensity score in observational studies for causal effects. *Biometrika*, 1983.
- Michael Schemper, Samo Wakounig, and Georg Heinze. The estimation of average hazard ratios by weighted cox regression. *Statistics in medicine*, 2009.
- Bernard W Silverman. *Density estimation for statistics and data analysis*. CRC press, 1986.
- Rodney A Sparapani, Brent R Logan, Robert E McCulloch, and Purushottam W Laud. Nonparametric survival analysis using bayesian additive regression trees (bart). *Statistics in medicine*, 2016.
- Ludovic Trinquart, Justine Jacot, Sarah C Conner, and Raphaël Porcher. Comparison of treatment effects measured by the hazard ratio and by the ratio of restricted mean survival times in oncology randomized controlled trials. *Journal of Clinical Oncology*, 2016.
- Lihui Zhao, Lu Tian, Hajime Uno, Scott D Solomon, Marc A Pfeffer, Jerald S Schindler, and Lee Jen Wei. Utilizing the integrated difference of two survival functions to quantify the treatment contrast for designing, monitoring, and analyzing a comparative clinical study. *Clinical trials*, 2012.