Simulation Study

1 Methods Description

1.1 Models

- **Cox proportional hazards model**: A standard semi-parametric model for survival analysis and is widely used as a baseline method.
- **Fair DeepPseudo (FIDP) model**: A model predict the survival function at a set of fixed time points. These time points are a predefined grid at which Kaplan Meier estimated pseudo values are computed and survival probabilities are produced. Model is trained on fairness-regularized loss $\mathcal{L} = \mathcal{L}_{\text{pv}} + \lambda\, \mathcal{L}_{\text{fair}}$, where $\mathcal{L}_{\text{pv}}$ denotes the pseudo-value regression loss, and $\mathcal{L}_{\text{fair}}$ is a fairness penalty computed by fairness definition. This simulation study use individual fairness, because it is used in code provided by paper.
- **Fair PseudoNAM (FIPNAM) model**: A neural additive model offering feature-level interpretability under the same fairness-regularized pseudo value framework.

1.2 Evaluation

- **Performance Metrics:** Evaluate model discrimination using the concordance index (C-index) and the time-dependent AUC across the prediction grid, and assess calibration using the Brier score.
- **Fairness Metrics:**
  - F_I: Individual fairness, comparing predictions between highly similar individuals.
  - F_{CI}: Censoring based individual fairness, comparing predictions between censored and uncensored individuals of similar covariate profiles.
  - F_{CG}: Censoring based group fairness, comparing predictions between censored and uncensored groups.
  - $F_{G\_Prot\_1}$ and $F_{G\_Prot\_2}$: Group fairness with respect to protected attributes. $F_{G\_Prot\_1}$ corresponds to age. $F_{G\_Prot\_2}$ corresponds to sex in the FLChain dataset, and group A_0 and A_1 in the simulated datasets.

2 Datasets Description

2.1 FLChain Dataset

FLChain contains follow up data from a cohort of patients who were tested for serum free light chain levels. The dataset contains 6521 subjects. Overall censoring rate is 70.0%.

2.2 Simulated Datasets

Simulated datasets are generated by following distributions, both of them contains 6994 subjects. Censoring rates are in Table 1, and summary statistics are in Table 2 in Appendix A. They are generated under different assumptions:

- SIM_I: Independent censoring. $C \perp T \mid X, A$
- SIM_II: Informative censoring, with censoring correlated with event time and covariates, true group difference, and unbalanced follow-up. In this dataset, Group A_1 less risk to censor compare to Group A_0.

Data generation process:

- **Sensitive attribute:** $A \sim \text{Bernoulli}(0.5)$
- **Clinical covariates:** $X = (X_1, X_2, X_3, X_4, X_5) \overset{\text{iid}}{\sim} N(\mu, \sigma^2)$. Detailed setting in Table 3 in Appendix A.
- **True event time:**

  $$Z = \frac{X - \mu_X}{\sigma_X}, \qquad \eta = \beta_A A + \beta_X^\top Z, \qquad U \sim \text{Uniform}(0, 1), \qquad T = \left(\frac{-\log U}{\lambda \exp(\eta)}\right)^{1/k}$$

  T is generated by Weibull proportional hazards model.Z is scaled X, scaled in order to produce controllable T distribution, not affect by difference range of X. β_X are coefficients determining the effect of each covariate on the log-risk. Set $k = 3$ and $\lambda = 1 \times 10^{-10}$ to give a shape and range close to real world time.

  - In SIM_I, $\beta_A = 0$, there is no true group difference.
  - In SIM_II, $\beta_A = -0.5$, so that group A_1 has lower risk systemically.

- **Censoring time:**

$$C_{\max}(A) = \begin{cases} C_{\max}(A_0), & A = 0, \\ C_{\max}(A_1), & A = 1. \end{cases}$$

$$U \sim \text{Uniform}(0,1), \qquad \tilde{U} = U^{\exp(\alpha\eta)}, \qquad C = \tilde{U}^{1/p_c} \cdot C_{\max}(A)$$

  Use power function to generate censoring time, censoring increase with time. p_c controls the skewness of the censoring distribution, set $p_c = 3$ to give a reasonable censoring distribution. All data after C_{max} is systematically censored by maximum follow-up. Use $\exp(\alpha\eta)$ to control informative censoring.

  - In SIM_I, $C_{\max}(A_0) = C_{\max}(A_1) = 3000$, set to align with range of T. $\alpha = 0$, it is independent censoring $\tilde{U} = U$
  - In SIM_II, $C_{\max}(A_0) = 3000, C_{\max}(A_1) = 3500$ to achieve unbalanced follow up. $\alpha = 0.5$, then higher η get easier to be censored, which reflects higher risk people are more likely to be censored.

- **Observed outcome:** $Y = \min(T, C, C_{\max}(A)), \quad \Delta = I\{T \leq \min(C, C_{\max}(A))\}$.

3 Results

FIDP and FIPNAM are trained using same setting as the paper, details at Appendix C. All result values are in Table 4 in Appendix A, here using plots to analyze visually.

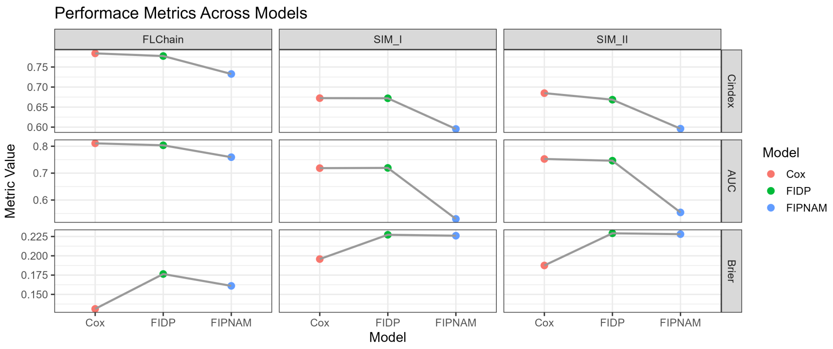

Figure 1: Comparison of C-index, AUC, and Brier score across the Cox, FIDP, and FIPNAM models for the FLChain, SIM_I, and SIM_II datasets. FIDP and FIPNAM models are using $\lambda = 0$ to be comparable with Cox. Each row corresponds to one performance metric and each column to one dataset.

Figure 1 shows Cox model outperform FIDP and FIPNAM, with higher C-index and AUC, and lower Brier score. FIPNAM has lowest C-index and AUC, but it still have similar Brier score compare to FIDP.

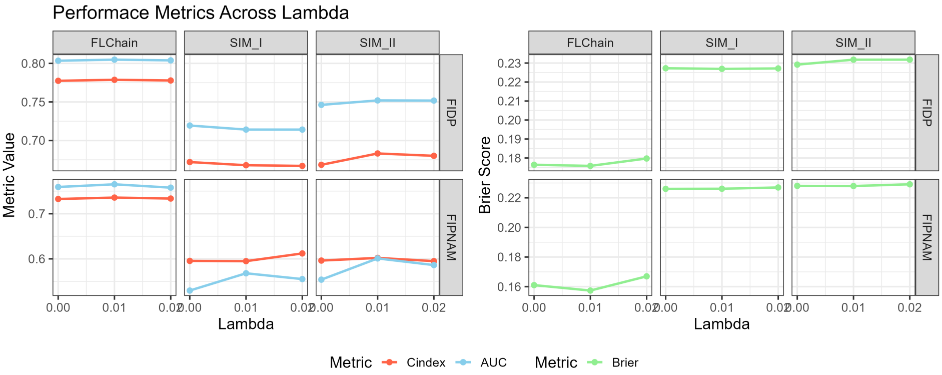

Figure 2: Comparison of performance metrics across $\lambda \in \{0.00, 0.01, 0.02\}$ for the FIDP and FIPNAM models. Results are shown separately for metrics, with C-index and AUC (higher the better), and Brier score (lower the better). Each row corresponds to one model, and each column to one dataset.

Theoretically, λ represent trade of between fairness and accuracy. However, Figure 2 shows that increase λ doesn't necessarily decrease C-index and AUC, or increase brier score. But for most cases, $\lambda = 1$ gives highest C-index and AUC, and lowest brier.

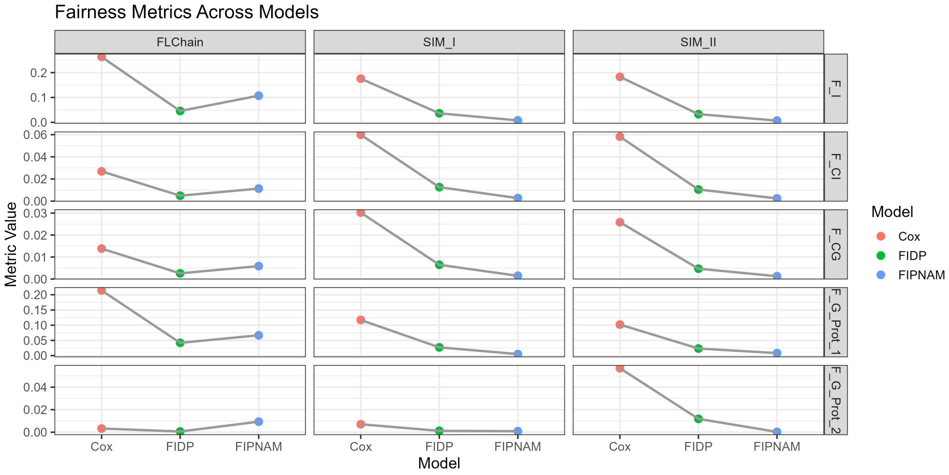

Figure 3: Comparison of fairness metrics (F_I, F_{CI}, F_{CG}, $F_{G\_Prot\_1}$, $F_{G\_Prot\_2}$) across three models for three datasets. FIDP and FIPNAM models are using $\lambda = 0$ to be comparable with Cox. Each row corresponds to one fairness metric, and each column to one dataset.

From Figure 3 we can see although λ are set to 0 for FIDP and FIPNAM, they both have better fairness than Cox model. In FLChain, FIDP has better fairness score than FIPNAM, but it is the opposite for two simulated datasets.

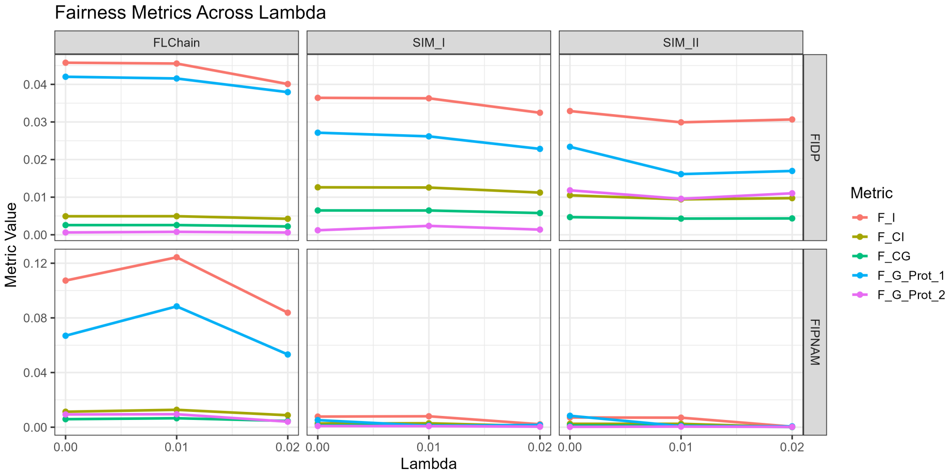

Figure 4: Comparison of fairness metrics across three values of fairness constraints λ for the FIDP and FIPNAM models. Each row corresponds to one model, and each column to one dataset.

Figure 4 shows that in most cases, increase λ from 0 to 0.01 gets better fairness. However, there are special cases like in FLChain, indiviual fairness and fairness of age group become worse. For FIPNAM, all fairness metrics are very low despite λ settings.

In summary, the Cox model consistently provides the strongest predictive performance. Both FIDP and FIPNAM perform reasonably well but underperform the Cox model in discrimination, especially FIPNAM. In terms of fairness, both FIDP and FIPNAM demonstrate substantially better fairness performace than the Cox model even when trained with $\lambda = 0$, indicating that their architectures inherently reduce unfair variability in predicted survival. Although FIPNAM offers better interpretability than FIDP, its poorer performance makes it less useful in practice. Varying the fairness regularization parameter λ shows mixed effects. Increasing λ often improves several fairness metrics from 0 to 0.01, but the improvement does not always continue at $\lambda = 0.02$. Performance metrics generally remain stable across values of λ, indicating that moderate fairness regularization does not impose a significant impact in these settings.

Overall, with $\lambda = 0.01$, FIDP model offers most balanced performance and fairness, makes it a useful survial model in pratice.

Appendix

A Tables

Table 1: Censoring Rates by Group and Overall

Dataset	Censor_A0	Censor_A1	Censor_Overall
SIM_I	33.3%	33.7%	33.5%
SIM_II	35.1%	31.1%	33.0%

Table 2: Summary Statistics for Two Simulated Datasets

Dataset	Variable	Min	Q1	Median	Mean	Q3	Max
SIM_I	T_true	62.87	1318.02	1820.81	1908.10	2388.30	6301.65
SIM_I	C	174.66	1885.73	2382.23	2247.41	2730.21	2999.93
SIM_I	Y	62.87	1189.88	1624.75	1640.09	2076.36	2997.55
SIM_I	age	3.82	39.57	49.56	49.75	59.93	103.18
SIM_I	LOS	0.10	3.42	5.08	5.15	6.71	14.62
SIM_I	eGFR	1.64	56.88	69.84	69.98	83.52	147.01
SIM_I	Hemoglobin	7.86	12.11	13.22	13.21	14.29	19.80
SIM_I	CCI	0.00	1.00	2.00	2.30	3.00	7.00
SIM_II	T_true	74.27	1423.70	1978.54	2084.86	2618.83	7444.52
SIM_II	C	71.60	2061.52	2612.88	2475.91	2969.33	3499.92
SIM_II	Y	71.60	1299.01	1789.25	1829.43	2340.10	3496.82

Table 3: Clinical Covariate Parameters

Covariate	μ	σ^2	β_x
Age	50	15	0.30
Length of stay (LOS)	5	2.5	0.40
Estimated glomerular filtration rate (eGFR)	70	20	-0.22
Hemoglobin	13.2	1.6	-0.30
Charlson comorbidity index (CCI)	2	1.5	0.35

Table 4: Performace and Fairness Metircs Summary

Model	Data	Lambda	Cindex	Brier	AUC	F_I	F_CI	F_CG	F_G_Prot_1	F_G_Prot_2
Cox	FLChain	/	0.784	0.131	0.811	0.263	0.027	0.014	0.214	0.003
Cox	SIM_I	/	0.672	0.196	0.719	0.176	0.060	0.030	0.117	0.007
Cox	SIM_II	/	0.685	0.187	0.753	0.183	0.058	0.026	0.102	0.057
FIDP	FLChain	0	0.777	0.176	0.803	0.046	0.005	0.003	0.042	0.001
FIDP	FLChain	0.01	0.779	0.176	0.805	0.046	0.005	0.003	0.042	0.001
FIDP	FLChain	0.02	0.778	0.180	0.804	0.040	0.004	0.002	0.038	0.001
FIDP	SIM_I	0	0.672	0.227	0.719	0.036	0.013	0.006	0.027	0.001
FIDP	SIM_I	0.01	0.668	0.227	0.714	0.036	0.013	0.006	0.026	0.002
FIDP	SIM_I	0.02	0.667	0.227	0.714	0.032	0.011	0.006	0.023	0.001
FIDP	SIM_II	0	0.668	0.229	0.746	0.033	0.010	0.005	0.023	0.012
FIDP	SIM_II	0.01	0.683	0.232	0.752	0.030	0.009	0.004	0.016	0.010
FIDP	SIM_II	0.02	0.680	0.232	0.752	0.031	0.010	0.004	0.017	0.011
FIPNAM	FLChain	0	0.733	0.161	0.759	0.107	0.011	0.006	0.067	0.009
FIPNAM	FLChain	0.01	0.736	0.157	0.765	0.124	0.013	0.007	0.088	0.010
FIPNAM	FLChain	0.02	0.734	0.167	0.758	0.084	0.009	0.005	0.053	0.004
FIPNAM	SIM_I	0	0.595	0.226	0.530	0.008	0.003	0.001	0.005	0.001
FIPNAM	SIM_I	0.01	0.595	0.226	0.568	0.008	0.003	0.002	0.001	0.001
FIPNAM	SIM_I	0.02	0.612	0.227	0.555	0.002	0.001	0.000	0.002	0.000
FIPNAM	SIM_II	0	0.596	0.228	0.554	0.007	0.002	0.001	0.008	0.000
FIPNAM	SIM_II	0.01	0.602	0.228	0.601	0.007	0.002	0.001	0.001	0.000
FIPNAM	SIM_II	0.02	0.595	0.229	0.586	0.000	0.000	0.000	0.001	0.000

B Change of codes

- **CPU Compatibility:** Oringal script only runs with GPU, now it canrun with CPU.
- **Dependency Updates:** The deprecated `scipy.integrate.simps` function was replaced with `scipy.integrate.simpson`, with a compatibility layer added in `Experiment.py`.
- **Simulated Dataset Support:** A new dataset type `SIMULATED` was added to `data_preprocess.py`, including column renaming (`Y` → `time`, `Delta` → `status`), removal of unused columns,

one-hot encoding for protected attribute `A`. The experiment runner automatically distinguishes between `SIMULATED_I` and `SIMULATED_II` based on the input filename.

- **Cox Model Support:** Add Cox model in script, with almost same pipeline as FIDP and FIPNAM, but no epoch training.
- **Output Management:** Results are now saved in CSV format instead of xls, and result files including the lambda parameter in their filenames. Training loss curves are saved as PNG files for FIDP and FIPNAM.
- **Evaluation During Training:** A new `compute_metrics` function calculates C-index, Brier score, and AUC during training every 3 epochs to monitor training progress.
- **Sigmoid Activation Function Modifications:** For simulated dataset, Pseudo values fall outside the [0,1] range. To allow the network to fit these unconstrained targets, FIDP model removes the sigmoid activation in the final layer during training to allow unbounded outputs and correctly fit these targets. A new `forward_prob` method was added to apply sigmoid only at evaluation time to produce valid survival probabilities. For the FIPNAM model, it retains the sigmoid in the forward pass because its additive linear structure becomes numerically unstable without output constraints; it also includes a `forward_prob` method for consistent probability-based evaluation.

C Result comparison with paper

Table 5: FLChain results for FIDP and FIPNAM: paper, original code, and modified code

	Paper Results (FLChain)									**Original Code Results (FLChain)**									**Modified Code Results (FLChain)**								
Model	Epoch	Cindex	Brier	AUC	F_I	F_{CI}	F_{CG}	$F_{G\_\text{Prot\_1}}$	$F_{G\_\text{Prot\_2}}$	Epoch	Cindex	Brier	AUC	F_I	F_{CI}	F_{CG}	$F_{G\_\text{Prot\_1}}$	$F_{G\_\text{Prot\_2}}$	Epoch	Cindex	Brier	AUC	F_I	F_{CI}	F_{CG}	$F_{G\_\text{Prot\_1}}$	$F_{G\_\text{Prot\_2}}$
FIDP	–	0.773	0.137	0.799	0.164	0.017	0.009	0.122	0.002	6	0.779	0.143	0.804	0.161	0.018	0.009	0.132	0.010	6	0.779	0.176	0.805	0.046	0.005	0.003	0.042	0.001
FIPNAM	–	0.768	0.153	0.798	0.189	0.022	0.011	0.139	0.022	2	0.748	0.154	0.775	0.119	0.014	0.007	0.081	0.014	11	0.736	0.157	0.765	0.124	0.013	0.007	0.088	0.010

In original code, I only change CPU compatibility and dependency update, so that it can run in HPC. No change to other code.

They all use Train 64%, Valid 16%, Test 20%, same as paper original setting.

They're all trained with 128 batch size, learning rate 0.01, 40 epoch, early stopping patience = 8. Paper states they use 100, but since they all encounter early stopping, maximum epoch size does not have impact on result. First column in the table is the epoch which model achieves best validation loss, and used to evaluated performance.

I think the main reason modified code performs a bit different is because modify on sigmoid activation function. Original code can not run on simulated data without this modification, c-index will always be around 0.5 and loss does not decrease, as stated in previous update.