STA497 Final Write up

1 Introduction

Machine learning models are widely used to support clinical decisions nowadays, and these decisions can directly influence how a patient is treated. Because of this impact, the model's fairness becomes essential. From a group-fairness perspective, a model is considered unfair when prediction patterns differ systematically across groups defined by a sensitive attribute ([Castelnovo et al., 2021]) (more definitions on different fairness used in this study are in methodology). One important reason for unfairness is data censoring ([Do et al., 2023]), meaning that their outcomes remain unobserved before the study ends. In real medical datasets, different groups may have different follow-up lengths, which is the period during which their outcomes are observed. They may also experience different censoring rates. Because of this, models may learn biased patterns and make unfair predictions.

Many fairness methods have been developed for this problem. Semi-supervised methods such as Infairness ([Gao et al., 2025]), use both labeled and unlabeled samples to estimate group fairness metrics more reliably. But it is only designed for binary outcomes and doesn't model survival times. Information-theoretic approaches such as FAST ([Do et al., 2023]) aim to minimizing the mutual information between the sensitive attribute and the predicted event-time distribution, preventing the model from making different predictions for different groups. However, its fairness regularization is applied only to the distribution of predicted event times rather than to time-dependent survival probabilities. The FISA paper ([Rahman and Purushotham, 2022]) addresses this by using pseudo-value regression to learn survival probabilities directly from censored data, allowing fairness constraints to be added directly to the training loss. Two models using this framwork is purposed, the Fair DeepPseudo (FIDP) and the Fair PseudoNAM (FIPNAM), which will be the focus of this study.

2 Methodology

2.1 Models

There are three models used in this study. Cox model is a standard model for survival analysis and is widely used as a baseline method. FIDP is a model predict the survival function at a set of fixed time points. These time points are a predefined grid at which Kaplan

Meier (KM) estimated ([Kaplan and Meier, 1958]) 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 ([Rahman and Purushotham, 2022]). λ is the fairness constraint, which controls how much the model treats fairness as a target. FIPNAM uses the same pseudo value and fairness-regularized framework as FIDP, but it is based on a neural additive model, which offers feature-level interpretability. Detailed model training pipelines are in Appendix A.

2.2 Evaluation

In this study, model discrimination is evaluated using the concordance index (C-index) and the time-dependent AUC across the prediction grid, and calibration using the Brier score. Fairness definitions are consistent with the paper ([Rahman and Purushotham, 2022]). F_I is individual fairness, measuring how much the difference in survival predictions deviates from similar subjects. (i.e., subjects with similar covariates). F_{CI} is censoring-based individual fairness, comparing between censored and uncensored similar individuals. F_G is group fairness with respect to protected attributes. It measures the maximum deviation of the groups' average survival predictions from the population's average survival predictions. F_{CG} is censoring-based group fairness, comparing between censored and uncensored groups.

2.3 Data

Three datasets are used in this research. One real dataset, FLChain, contains 6521 subjects, overall censoring rate is 70%. Clinical covariates are age, sex, calendar year of sample collection, free light chain measurements (kappa and lambda), an indicator of abnormal FLC group, creatinine, and an indicator for monoclonal gammopathy, and outcome is time to death. Sensitive attributes here are age and sex. The goal is to estimate survival probabilities over time for each patient. Two simulated datasets both contains 6994 subjects. They are generated under different assumptions, independent censoring and informative censoring. In SIM_I, censoring time is independent. In SIM_II, censoring correlated with event time and covariates. Two groups have a true group difference and an unbalanced follow-up. In this dataset, A_0 and A_1 are sensitive attributes. Group A_1 has less risk to censoring compared to Group A_0. Both simulated datasets have a censoring rate of around 35%. More details on how simulated datasets are generated in Appendix B.

3 Results

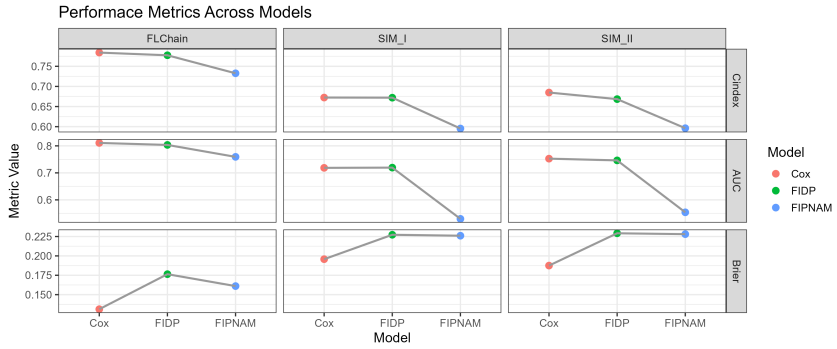

Figure 1: Comparison of C-index, AUC, and Brier score across the Cox, FIDP, and FIPNAM models for the FLChain, SIM_I, and SIM_II datasets. FIDP and FIPNAM models are using $\lambda = 0$ to be comparable with Cox.

Figure 1 compares the predictive performance of the three models. Across all datasets, the Cox model achieves the highest C-index and AUC, and the lowest Brier score. FIDP performs slightly worse but remains close to Cox on most metrics. FIPNAM shows similar Brier score as FIDP, but weaker discrimination. These results indicate that the Cox model remains the strongest baseline in terms of discrimination and calibration. This is expected because the FLChain data has heavy censoring near the end of follow-up, and the simulated datasets are generated from a proportional hazards model. This leads to an increasing hazard trend that closely matches the assumptions of the Cox model.

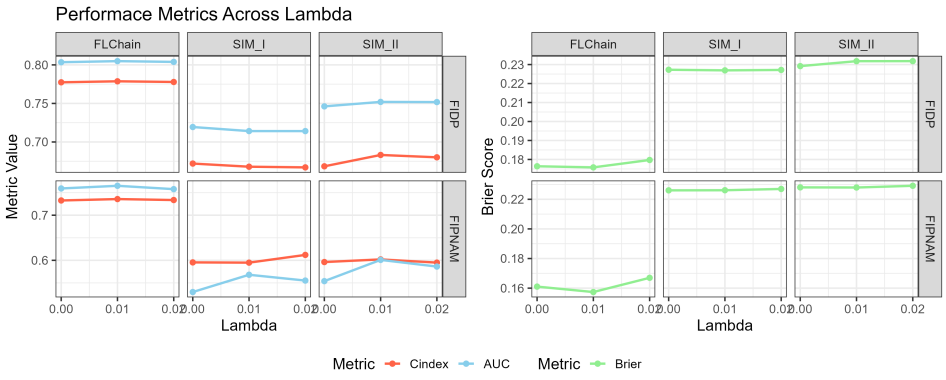

Figure 2: Comparison of performance metrics across $\lambda \in \{0.00, 0.01, 0.02\}$ for the FIDP and FIPNAM models. Results are shown separately for metrics, with C-index and AUC (higher the better), and Brier score (lower the better).

Figure 2 shows how the fairness regularization parameter λ affects performance. In theory,

λ controls the trade-off between fairness and accuracy. However, the results show that C-index or AUC is not always decline when λ increases. In many cases, moderate regularization ($\lambda = 0.01$) slightly improves both C-index and AUC. Brier scores also remain stable. This suggests that mild fairness regularization does not significantly harm predictive accuracy.

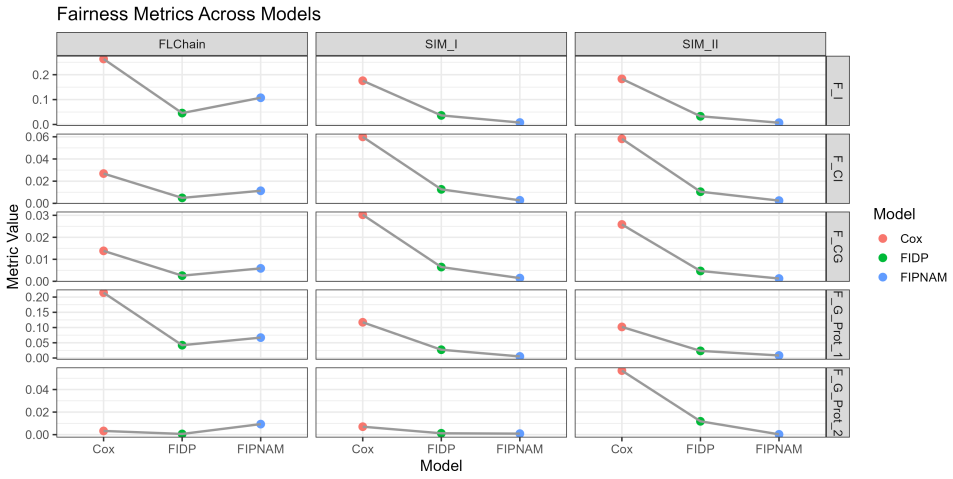

Figure 3: Comparison of fairness metrics (F_I, F_{CI}, F_{CG}, $F_{G\_Prot\_1}$, $F_{G\_Prot\_2}$) across three models for three datasets. $F_{G\_Prot\_1}$ treats age as a sensitive attribute in FLChain, and group A_0 in SIM_I. $F_{G\_Prot\_1}$ treats sex as a sensitive attribute in FLChain, and group A_1 in two simulated datasets. FIDP and FIPNAM models are using $\lambda = 0$.

Figure 3 compares fairness metrics at $\lambda = 0$. Even without fairness penalties, both FIDP and FIPNAM produce lower fairness scores than the Cox model, indicating better fairness. This behavior is consistent with the model design. The Cox model predicts risk through an exponential hazard function, which can strongly enlarge small covariate differences. In FIDP and FIPNAM, covariate effects enter the prediction through pseudo-value regression rather than an exponential hazard term, resulting in smaller changes in predicted survival for comparable covariate differences. In FLChain, FIDP achieves better fairness than FIPNAM, while in the simulated datasets the pattern is reversed. This implies that FIDP and FIPNAM are not inherently one fairer than the other, but rather depend more on the data structure and censoring conditions.

Figure 4 examines fairness across different values of λ. Increasing λ from 0 to 0.01 generally improves most fairness metrics, but doesn't always continue when increasing to 0.02. However, FIPNAM doesn't follow this pattern. In FLChain, it shows the opposite trend. In the two simulated datasets, all metrics don't change significantly. This might be caused by the smoothness constraint used in FIPNAM's additive structure. It shows that the effect of λ can

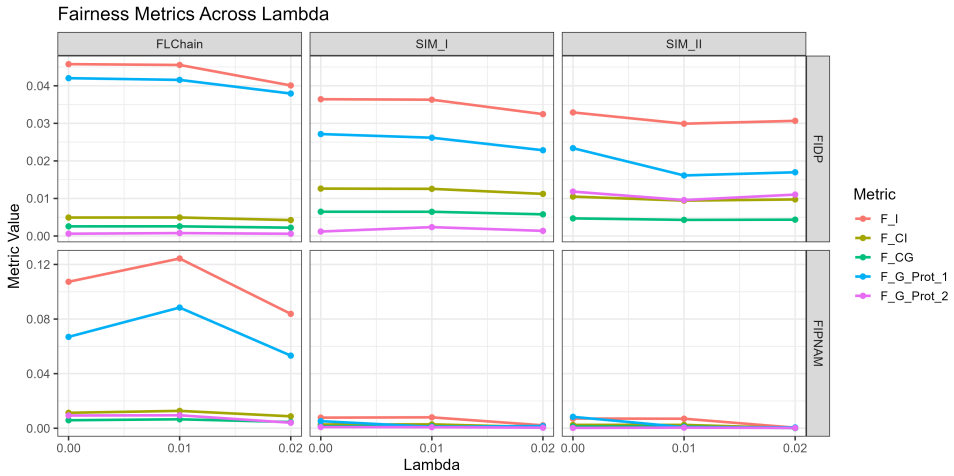

Figure 4: Comparison of fairness metrics across $\lambda \in \{0.00, 0.01, 0.02\}$ for the FIDP and FIPNAM models. Each row corresponds to one model, and each column to one dataset.

be model and dataset-dependent.

4 Conclusion

The Cox model delivers the strongest predictive performance but the poorest fairness, especially when informative censoring. FIDP and FIPNAM achieve better fairness even with $\lambda = 0$. This indicates that pseudo-value modeling can reduce unfairness. FIDP delivers a better balance compared to FIPNAM, which shows weaker accuracy.

Fairness regularization has mixed effects. Increasing λ from 0 to 0.01 improves fairness in most settings, while performance remains largely unchanged. Higher regularization values do not consistently yield further gains in this study setting. Among the two pseudo-value models, FIDP offers stronger accuracy than FIPNAM while still improving fairness, making it the most practical option in this study.

Future work can dig deeper into the effect of a smaller change in λ when models train on a larger dataset. For a training loss regularization term, its effect can be more stable as the data gets larger. There is also ipcw loss function ([Rahman and Purushotham, 2022]) integrated into FIDP and FIPNAM, which is designed for informative censoring. It is worth exploring whether it can further improve fairness in this setting. Also, good fairness metrics performance can still mask substantial inequalities in clinical utility ([Chen et al., 2024]), to make models truly useful and practical, evaluation on clinical utility can be addressed on these two models.

References

[Castelnovo et al., 2021] Castelnovo, A., Crupi, R., Greco, G., Regoli, D., Penco, I. G., and Cosentini, A. C. (2021). A clarification of the nuances in the fairness metrics landscape. Later published in Scientific Reports, 12(1):4209, 2022.

[Chen et al., 2024] Chen, I. Y., Zhang, H., Beutel, A., et al. (2024). Epsilon-fairness: Unifying fairness and utility in machine learning. *arXiv preprint arXiv:2405.09360*.

[Do et al., 2023] Do, H., Chang, Y., Cho, Y. S., Smyth, P., and Zhong, J. (2023). Fair survival time prediction via mutual information minimization. In *Proceedings of Machine Learning for Healthcare (MLHC 2023)*, volume 219, pages 1–45. PMLR.

[Gao et al., 2025] Gao, Y., Zeng, W., Tang, J., and Zou, J. (2025). Infairness: Semi-supervised fairness auditing under limited labels. *arXiv preprint arXiv:2505.12181*.

[Kaplan and Meier, 1958] Kaplan, E. L. and Meier, P. (1958). Nonparametric estimation from incomplete observations. *Journal of the American Statistical Association*, 53(282):457–481.

[Rahman and Purushotham, 2022] Rahman, M. M. and Purushotham, S. (2022). Fair and interpretable models for survival analysis. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD '22)*, pages 1–11, Washington, DC, USA. ACM.

Appendix

A Model Training Piplines

All three models use a 64%, 16%, 20% train–validation–test split. FIDP and FIPNAM are trained with 128 batch size, 0.01 learning rate, and 40 epochs, with early stopping if the validation loss is not smaller than the best model so far for 8 consecutive epochs.

For results of this study, FIDP achieves the best validation loss at epoch 6, and FIPNAM achieves the best validation loss at epoch 11.

B Simulated Data Generation Process

- **Sensitive attribute:** $A \sim \text{Bernoulli}(0.5)$
- **Clinical covariates:** $X = (X_1, X_2, X_3, X_4, X_5) \overset{\text{iid}}{\sim} N(\mu, \sigma^2)$.

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

- **True event time:**

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

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. T is generated by the Weibull proportional hazards model. With λ (baseline rate) and k (shape parameter) controlling time scale and hazard monotonicity. Set $k = 3$ to get an increasing hazard over time, close to realistic time-to-event dynamics for clinical settings. Set $\lambda = 1 \times 10^{-10}$ so that event times fall within a range comparable to the FLChain dataset.

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

Simulated datasets are generated with 8000 subjects. After filtering unrealistic covariates (`age > 1, C > 1, LOS > 0.1, eGFR > 0.1, Hemoglobin > 0.1, CCI > 0.1`), there are 6994 subjects left in both datasets. 8000 is chosen because the final datasets contain a similar number of subjects as FLChain, so that they're comparable. Table 1 provides summary statistics for two simulated datasets.

Table 1: 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

C Models Performance

Table 2 provides detailed performance metrics values used in the Result section.

Table 2: 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