Proceedings of Machine Learning Research LEAVE UNSET:1–8, 2023 Conference on Health, Inference, and Learning (CHIL) 2023

Neural Fine-Gray: Monotonic neural networks for competing risks Supplementary materials

Appendix A. Experiments

A.1. Datasets characteristics

Table 1 presents the times and observed outcomes corresponding to the different quantiles of the uncensored population used for evaluation, differentiated by datasets.

		Quantiles		
		$q_{0.25}$	$q_{0.50}$	$q_{0.75}$
PBC	Time (years)	3.19	4.95	7.45
	Censoring	0.00 %	0.00%	11.54%
	Death	12.82%	23.72%	32.69%
	Transplant	0.64%	3.21%	7.69%
Fram.	Time (years)	5.90	12.57	18.14
	Censoring	0.00 %	0.00%	0.00%
	Death	2.66%	7.40%	12.56%
	CVD	8.30%	14.52%	20.32%
Synth.	Time	3.00	11.00	30.00
	Censoring	20.54%	35.32%	44.65%
	Cause 1	5.28%	12.43%	18.96%
	Cause 2	5.05%	12.21%	18.43%
SEER	Time (years)	1.67	4.00	8.08
	Censoring	10.34%	22.20 %	39.59%
	BC	4.53%	9.32%	13.37%
	CVD	0.80%	1.76%	3.23%

Table 1: Observed outcomes of interest at the different evaluation horizons.

A.2. Evaluation metrics

Time Dependent C-Index Time-dependent C-Index (Antolini et al., 2005) quantifies the model discrimination by comparing the ordering of the predicted survival probability for risk r and the observed survival times, i.e. it is an estimate of:

$$\mathbb{P}(\hat{F}_r(t|x_i) > \hat{F}_r(t|x_j)|d_i = r, t_i < t_j, t_i \leq t)$$

This probability is approximated and weighted by the inverse probability $\omega(t_i)$ of censoring derived from a Kaplan-Meier estimator.

Time Dependent Brier Score Time dependent Brier score (Graf et al., 1999) measures the model calibration for risk r, similarly corrected for censoring:

$$\mathrm{BS}^r(t) = \frac{1}{n}\sum_i \left[\omega(t_i)\mathbb{1}_{i,d_i=r\wedge t_i\leq t}(1-\hat{F}_r(t|x_i))^2 + \omega(t)\mathbb{1}_{t_i>t}\hat{F}_r(t|x_i)^2\right]$$

with $\mathbb{1}$, the indicator function, $\hat{S}(t|x)$, the predicted survival probability at time t.

A.3. Time specific results

Tables 2, 3 and 4 present the performance evaluated at the dataset-specific 0.25, 0.5 and 0.75 quantiles of the uncensored population event times through respectively C-index, ROC-AUC, and Brier score. The table echoes the same conclusions presented in the paper with competing or better than state-of-the-art performance.

A.4. Cumulative evaluation

The cumulative metrics summarise how a model performs over the total distribution. While having the advantage of representing performance in a single number, it is more disconnected from medical applications in which the risk horizon would be discretized to inform patients' treatment. Table 5 displays the time-dependent C-index and cumulative time-dependent Brier score. These results echo the findings from the paper.

A.5. Implementation details

The proposed experiments rely on the `scikit-survival` (Pölsterl, 2020)[1] and `pycox`[2] libraries for evaluation. For baselines' implementations, we used the R library `riskRegression`[3] for CS Cox and Fine-Gray, `pycox` for DeepHit and `auton-survival` (Nagpal et al., 2022)[4] for Deep Survival Machines.

Appendix B. Using R outcomes vs. R networks

In this section, we investigate the impact of using multiple networks – one for each competing risk – instead of one network with multiple outcomes. The model **MonoFG** consists of the same architecture presented in Section 3.3 with only one monotonic network with R outputs. Table 6 shows limited differences between the two architectures. However, we encourage the use of multiple networks when the competing risks present large distributional differences.

Appendix C. DeSurv

C.1. Impact of n

In the upper limit, the Gauss-Legendre quadrature would lead to the exact estimation of the likelihood. However, this requires n forward passes in the neural network with n, the number of point estimation. Fixing the architecture to a 3 hidden layer perceptron with 50 nodes, we measure the model's performances for n in $[1, 15, 100, 1000]$ on the Synthetic dataset as shown in Table 7. For $n = 1$, NeuralFG and DeSurv present the same computational complexity. However, DeSurv benefits from larger n. Note that there is limited gain above the recommended 15-degree quadrature.

1. https://github.com/sebp/scikit-survival
2. https://github.com/havakv/pycox
3. https://github.com/tagteam/riskRegression
4. https://github.com/autonlab/auton-survival

	Model	Primary risk			Competing risk		
		$q_{0.25}$	$q_{0.50}$	$q_{0.75}$	$q_{0.25}$	$q_{0.50}$	$q_{0.75}$
PBC	**NeuralFG**	0.810 (0.079)	0.795 (0.114)	0.762 (0.123)	0.799 (0.082)	0.709 (0.309)	*0.788* (0.145)
	DeepHit	0.822 (0.099)	0.844 (0.036)	0.782 (0.033)	0.790 (0.044)	0.614 (0.174)	0.612 (0.095)
	DeSurv	0.821 (0.089)	0.837 (0.050)	0.815 (0.068)	0.802 (0.123)	**0.781** (0.268)	**0.796** (0.153)
	DSM	**0.867** (0.065)	**0.864** (0.037)	**0.828** (0.052)	0.694 (0.224)	0.721 (0.251)	0.703 (0.175)
	Fine-Gray	0.831 (0.136)	*0.852* (0.045)	*0.816* (0.059)	**0.865** (0.087)	0.686 (0.330)	0.741 (0.123)
	CS Cox	*0.833* (0.125)	0.851 (0.040)	0.811 (0.065)	*0.837* (0.022)	*0.734* (0.276)	0.783 (0.118)
Framingham	**NeuralFG**	**0.872** (0.024)	**0.812** (0.029)	**0.782** (0.018)	**0.745** (0.055)	**0.717** (0.038)	*0.713* (0.022)
	DeepHit	0.855 (0.026)	0.781 (0.026)	0.743 (0.014)	0.713 (0.035)	0.690 (0.030)	0.693 (0.015)
	DeSurv	**0.872** (0.027)	*0.807* (0.031)	0.775 (0.022)	0.721 (0.036)	0.706 (0.038)	0.708 (0.028)
	DSM	*0.866* (0.023)	0.806 (0.023)	*0.778* (0.014)	0.717 (0.064)	0.709 (0.034)	0.712 (0.021)
	Fine-Gray	0.842 (0.025)	0.794 (0.024)	0.772 (0.015)	0.729 (0.036)	0.709 (0.040)	0.710 (0.023)
	CS Cox	0.845 (0.020)	0.798 (0.022)	0.774 (0.015)	*0.741* (0.050)	*0.712* (0.044)	**0.715** (0.023)
Synthetic	**NeuralFG**	*0.791* (0.013)	*0.754* (0.013)	**0.715** (0.011)	*0.801* (0.016)	*0.755* (0.018)	*0.714* (0.016)
	DeepHit	0.783 (0.012)	0.747 (0.013)	*0.714* (0.008)	0.792 (0.015)	0.744 (0.015)	**0.715** (0.012)
	DeSurv	**0.793** (0.013)	**0.756** (0.014)	*0.714* (0.014)	**0.803** (0.015)	**0.756** (0.016)	0.713 (0.015)
	DSM	0.776 (0.013)	0.742 (0.013)	0.710 (0.013)	0.785 (0.019)	0.742 (0.019)	0.708 (0.020)
	Fine-Gray	0.611 (0.014)	0.587 (0.007)	0.568 (0.009)	0.633 (0.014)	0.593 (0.015)	0.574 (0.015)
	CS Cox	0.609 (0.015)	0.586 (0.006)	0.568 (0.009)	0.630 (0.013)	0.592 (0.014)	0.573 (0.015)
SEER	**NeuralFG**	*0.893* (0.002)	*0.855* (0.001)	*0.815* (0.001)	0.799 (0.010)	0.782 (0.005)	*0.758* (0.003)
	DeepHit	**0.899** (0.002)	**0.860** (0.001)	**0.818** (0.001)	**0.824** (0.008)	**0.801** (0.005)	**0.770** (0.004)
	DeSurv	0.892 (0.003)	0.852 (0.002)	0.813 (0.001)	0.811 (0.006)	*0.788* (0.006)	0.757 (0.004)
	DSM	0.884 (0.001)	0.842 (0.002)	0.805 (0.002)	*0.813* (0.008)	0.787 (0.004)	0.755 (0.004)
	Fine-Gray	0.836 (0.003)	0.786 (0.003)	0.742 (0.002)	0.757 (0.008)	0.745 (0.005)	0.727 (0.005)
	CS Cox	0.837 (0.003)	0.786 (0.003)	0.742 (0.002)	0.781 (0.010)	0.759 (0.007)	0.734 (0.006)

Table 2: Comparison of the **C-index** across 5-fold cross-validation. Best performances are in **bold**, second best in *italics*. *NeuralFG is the model introduced in this paper.*

Dataset	Model	Primary risk $q_{0.25}$	Primary risk $q_{0.50}$	Primary risk $q_{0.75}$	Competing risk $q_{0.25}$	Competing risk $q_{0.50}$	Competing risk $q_{0.75}$
PBC	**NeuralFG**	0.822 (0.088)	0.825 (0.145)	0.809 (0.161)	0.804 (0.097)	0.741 (0.316)	**0.842** (0.169)
	DeepHit	0.831 (0.104)	0.876 (0.054)	0.803 (0.057)	0.786 (0.045)	0.610 (0.175)	0.623 (0.102)
	DeSurv	0.823 (0.088)	0.866 (0.065)	**0.855** (0.100)	0.807 (0.113)	**0.795** (0.269)	*0.831* (0.174)
	DSM	**0.876** (0.067)	**0.900** (0.043)	*0.854* (0.062)	0.707 (0.228)	0.728 (0.242)	0.695 (0.175)
	Fine-Gray	0.835 (0.136)	*0.887* (0.059)	0.844 (0.089)	**0.871** (0.075)	0.706 (0.336)	0.754 (0.133)
	CS Cox	*0.839* (0.127)	0.886 (0.056)	0.843 (0.097)	*0.843* (0.009)	*0.750* (0.272)	0.798 (0.123)
Framingham	**NeuralFG**	**0.877** (0.025)	**0.827** (0.028)	**0.810** (0.016)	**0.752** (0.056)	**0.736** (0.042)	*0.742* (0.024)
	DeepHit	0.860 (0.026)	0.796 (0.024)	0.770 (0.015)	0.720 (0.034)	0.708 (0.037)	0.723 (0.018)
	DeSurv	*0.876* (0.028)	*0.821* (0.030)	0.803 (0.020)	0.728 (0.035)	0.724 (0.043)	0.736 (0.032)
	DSM	0.870 (0.023)	0.819 (0.022)	0.803 (0.015)	0.722 (0.065)	0.727 (0.039)	0.741 (0.024)
	Fine-Gray	0.849 (0.027)	0.812 (0.023)	0.802 (0.015)	0.736 (0.036)	0.727 (0.044)	0.739 (0.027)
	CS Cox	0.852 (0.022)	0.816 (0.021)	*0.804* (0.015)	*0.748* (0.051)	*0.730* (0.047)	**0.745** (0.025)
Synthetic	**NeuralFG**	*0.814* (0.015)	*0.806* (0.012)	**0.790** (0.015)	*0.821* (0.021)	*0.804* (0.017)	*0.785* (0.015)
	DeepHit	0.806 (0.016)	0.803 (0.013)	*0.788* (0.015)	0.814 (0.020)	0.796 (0.015)	**0.790** (0.013)
	DeSurv	**0.817** (0.016)	**0.809** (0.013)	0.787 (0.017)	**0.824** (0.020)	**0.805** (0.016)	0.780 (0.013)
	DSM	0.800 (0.016)	0.794 (0.013)	0.782 (0.013)	0.807 (0.023)	0.790 (0.020)	0.776 (0.022)
	Fine-Gray	0.603 (0.018)	0.583 (0.008)	0.562 (0.005)	0.624 (0.014)	0.585 (0.018)	0.565 (0.018)
	CS Cox	0.601 (0.018)	0.583 (0.008)	0.562 (0.005)	0.621 (0.013)	0.584 (0.018)	0.565 (0.018)
SEER	**NeuralFG**	*0.901* (0.001)	*0.868* (0.002)	*0.826* (0.001)	0.804 (0.010)	0.789 (0.006)	0.761 (0.003)
	DeepHit	**0.907** (0.002)	**0.874** (0.001)	**0.835** (0.002)	**0.832** (0.008)	**0.814** (0.006)	**0.783** (0.004)
	DeSurv	0.899 (0.003)	0.866 (0.002)	0.825 (0.001)	0.818 (0.006)	*0.798* (0.007)	*0.764* (0.004)
	DSM	0.891 (0.001)	0.855 (0.002)	0.815 (0.002)	*0.821* (0.009)	0.796 (0.004)	0.758 (0.005)
	Fine-Gray	0.840 (0.003)	0.799 (0.003)	0.757 (0.002)	0.760 (0.009)	0.749 (0.005)	0.736 (0.005)
	CS Cox	0.841 (0.003)	0.799 (0.003)	0.758 (0.002)	0.785 (0.010)	0.766 (0.007)	0.745 (0.006)

Table 3: Comparison of the **time-dependent AUC** across 5-fold cross-validation. Best performances are in **bold**, second best in *italics*. *NeuralFG is the model introduced in this paper.*

	Model	Primary risk $q_{0.25}$	Primary risk $q_{0.50}$	Primary risk $q_{0.75}$	Competing risk $q_{0.25}$	Competing risk $q_{0.50}$	Competing risk $q_{0.75}$
PBC	**NeuralFG**	0.099 (0.028)	0.140 (0.020)	0.169 (0.050)	*0.018* (0.001)	0.036 (0.015)	0.092 (0.017)
	DeepHit	*0.090* (0.030)	0.132 (0.013)	0.180 (0.021)	*0.018* (0.002)	0.039 (0.020)	0.100 (0.010)
	DeSurv	**0.088** (0.022)	0.113 (0.011)	**0.136** (0.047)	0.019 (0.001)	**0.031** (0.011)	**0.087** (0.020)
	DSM	0.091 (0.039)	0.124 (0.015)	0.161 (0.022)	**0.017** (0.000)	*0.035* (0.018)	0.099 (0.017)
	Fine-Gray	0.091 (0.042)	*0.103* (0.009)	0.150 (0.038)	**0.017** (0.000)	0.041 (0.017)	*0.091* (0.017)
	CS Cox	0.091 (0.038)	**0.102** (0.008)	*0.148* (0.038)	*0.018* (0.000)	0.038 (0.018)	**0.087** (0.018)
Framingham	**NeuralFG**	*0.050* (0.003)	**0.095** (0.010)	**0.128** (0.004)	*0.027* (0.003)	**0.070** (0.004)	*0.112* (0.005)
	DeepHit	0.053 (0.003)	0.102 (0.007)	0.141 (0.002)	*0.027* (0.003)	0.072 (0.005)	0.115 (0.005)
	DeSurv	**0.049** (0.005)	**0.095** (0.009)	*0.129* (0.003)	*0.027* (0.003)	**0.070** (0.005)	0.113 (0.004)
	DSM	0.057 (0.005)	0.104 (0.006)	0.141 (0.002)	*0.027* (0.003)	*0.071* (0.004)	**0.111** (0.004)
	Fine-Gray	0.057 (0.006)	0.099 (0.007)	0.131 (0.003)	*0.027* (0.003)	*0.071* (0.005)	*0.112* (0.005)
	CS Cox	0.056 (0.006)	*0.098* (0.007)	0.131 (0.003)	**0.026** (0.003)	**0.070** (0.005)	**0.111** (0.005)
Synthetic	**NeuralFG**	**0.068** (0.003)	*0.125* (0.004)	**0.192** (0.005)	**0.064** (0.003)	*0.125* (0.002)	**0.191** (0.005)
	DeepHit	0.079 (0.003)	0.136 (0.002)	*0.212* (0.003)	0.075 (0.003)	0.132 (0.003)	*0.204* (0.005)
	DeSurv	**0.068** (0.002)	**0.124** (0.004)	**0.192** (0.004)	**0.064** (0.003)	**0.124** (0.003)	**0.191** (0.005)
	DSM	*0.073* (0.002)	0.139 (0.002)	*0.220* (0.003)	*0.069* (0.002)	0.138 (0.002)	0.217 (0.004)
	Fine-Gray	0.078 (0.002)	0.159 (0.003)	0.241 (0.002)	0.074 (0.003)	0.159 (0.003)	0.238 (0.004)
	CS Cox	0.078 (0.002)	0.159 (0.003)	0.240 (0.002)	0.074 (0.003)	0.159 (0.003)	0.238 (0.004)
SEER	**NeuralFG**	**0.038** (0.000)	**0.069** (0.001)	**0.101** (0.000)	**0.009** (0.000)	*0.021* (0.000)	**0.043** (0.000)
	DeepHit	**0.038** (0.000)	*0.070* (0.000)	*0.102* (0.001)	**0.009** (0.000)	**0.020** (0.000)	**0.043** (0.000)
	DeSurv	**0.038** (0.000)	*0.070* (0.000)	*0.102* (0.001)	**0.009** (0.000)	*0.021* (0.000)	**0.043** (0.000)
	DSM	*0.039* (0.000)	0.076 (0.001)	0.112 (0.000)	**0.009** (0.000)	**0.020** (0.000)	**0.043** (0.000)
	Fine-Gray	0.043 (0.001)	0.081 (0.000)	0.118 (0.000)	**0.009** (0.000)	*0.021* (0.000)	*0.044* (0.000)
	CS Cox	0.042 (0.001)	0.081 (0.000)	0.118 (0.000)	**0.009** (0.000)	*0.021* (0.000)	*0.044* (0.000)

Table 4: Comparison of the **Brier Score** across 5-fold cross-validation. Best performances are in **bold**, second best in *italics*. *NeuralFG is the model introduced in this paper.*

	Model	Primary risk		Competing Risk	
		C^{td}-Index	Brier Score	C^{td}-Index	Brier Score
PBC	**NeuralFG**	0.746 (0.116)	0.166 (0.024)	*0.785* (0.166)	*0.154* (0.035)
	DeepHit	0.733 (0.069)	*0.157* (0.013)	0.627 (0.088)	*0.154* (0.013)
	DeSurv	*0.804* (0.059)	*0.157* (0.033)	**0.819** (0.123)	**0.153** (0.049)
	DSM	**0.812** (0.050)	**0.152** (0.019)	0.707 (0.152)	0.164 (0.028)
	Fine-Gray	0.797 (0.057)	0.182 (0.170)	0.732 (0.138)	0.177 (0.064)
	CS Cox	0.796 (0.056)	-	0.769 (0.120)	0.160 (0.072)
Framingham	**NeuralFG**	**0.775** (0.018)	*0.089* (0.004)	*0.716* (0.022)	0.072 (0.002)
	DeepHit	0.760 (0.022)	0.157 (0.141)	0.698 (0.011)	0.081 (0.003)
	DeSurv	*0.771* (0.021)	**0.082** (0.041)	0.712 (0.021)	0.072 (0.003)
	DSM	0.767 (0.016)	0.099 (0.002)	0.701 (0.014)	**0.069** (0.002)
	Fine-Gray	0.765 (0.016)	0.152 (0.036)	*0.716* (0.022)	0.072 (0.003)
	CS Cox	0.767 (0.015)	-	**0.718** (0.028)	*0.071* (0.002)
Synthetic	**NeuralFG**	**0.735** (0.010)	**0.228** (0.004)	**0.738** (0.014)	**0.233** (0.003)
	DeepHit	0.722 (0.009)	0.245 (0.004)	0.725 (0.010)	0.240 (0.004)
	DeSurv	*0.734* (0.010)	*0.231* (0.005)	*0.737* (0.014)	*0.237* (0.006)
	DSM	0.719 (0.010)	0.286 (0.005)	0.722 (0.017)	0.287 (0.007)
	Fine-Gray	0.583 (0.007)	0.257 (0.002)	0.591 (0.014)	0.265 (0.002)
	CS Cox	0.582 (0.007)	0.254 (0.002)	0.590 (0.013)	0.262 (0.002)
SEER	**NeuralFG**	**0.819** (0.001)	**0.079** (0.000)	0.755 (0.004)	**0.032** (0.000)
	DeepHit	0.803 (0.002)	0.198 (0.004)	**0.763** (0.003)	*0.148* (0.002)
	DeSurv	*0.818* (0.001)	0.176 (0.002)	*0.756* (0.004)	0.183 (0.002)
	DSM	0.801 (0.001)	0.193 (0.002)	0.745 (0.004)	0.184 (0.001)
	Fine-Gray	0.750 (0.002)	*0.160* (0.033)	0.723 (0.004)	0.179 (0.001)
	CS Cox	0.750 (0.002)	0.200 (0.003)	0.733 (0.005)	0.180 (0.001)

Table 5: Comparison of model performance by means (standard deviations) across 5-fold cross-validation. Best performances are in **bold**, second best in *italics*. '-' indicates the divergence of the estimated Brier score. *NeuralFG is the model introduced in this paper.*

	Risk	Model	C-Index *(Larger is better)* $q_{0.25}$	$q_{0.50}$	$q_{0.75}$	Brier Score *(Smaller is better)* $q_{0.25}$	$q_{0.50}$	$q_{0.75}$
PBC	Dea.	**NeuralFG**	0.810 (0.079)	0.795 (0.114)	0.762 (0.123)	0.099 (0.028)	0.140 (0.020)	0.169 (0.050)
		MonoFG	**0.815** (0.086)	**0.797** (0.097)	**0.773** (0.114)	**0.095** (0.026)	**0.135** (0.026)	**0.155** (0.060)
	Tra.	**NeuralFG**	**0.799** (0.082)	**0.709** (0.309)	**0.788** (0.145)	**0.018** (0.001)	**0.036** (0.015)	**0.092** (0.017)
		MonoFG	0.699 (0.072)	0.632 (0.272)	0.709 (0.097)	**0.018** (0.001)	0.040 (0.019)	0.098 (0.019)
Fram.	CVD	**NeuralFG**	**0.872** (0.024)	**0.812** (0.029)	**0.782** (0.018)	0.050 (0.003)	**0.095** (0.010)	**0.128** (0.004)
		MonoFG	0.870 (0.024)	0.807 (0.028)	0.778 (0.020)	**0.049** (0.003)	**0.095** (0.009)	**0.128** (0.005)
	Dea.	**NeuralFG**	**0.745** (0.055)	**0.717** (0.038)	**0.713** (0.022)	**0.027** (0.003)	**0.070** (0.004)	**0.112** (0.005))
		MonoFG	0.735 (0.047)	**0.717** (0.037)	**0.713** (0.018)	**0.027** (0.003)	0.071 (0.005)	0.113 (0.005)
Synthetic	1	**NeuralFG**	0.791 (0.013)	0.754 (0.013)	**0.715** (0.011)	**0.068** (0.003)	**0.125** (0.004)	**0.192** (0.005)
		MonoFG	**0.792** (0.012)	**0.755** (0.013)	**0.715** (0.011)	**0.068** (0.003)	**0.125** (0.004)	**0.192** (0.006)
	2	**NeuralFG**	**0.801** (0.016)	**0.755** (0.018)	**0.714** (0.016)	**0.064** (0.003)	**0.125** (0.002)	**0.191** (0.005)
		MonoFG	**0.801** (0.015)	**0.755** (0.016)	0.713 (0.013)	**0.064** (0.003)	**0.125** (0.002)	**0.191** (0.004)
SEER	BC	**NeuralFG**	0.893 (0.002)	**0.855** (0.001)	**0.815** (0.001)	**0.038** (0.000)	**0.069** (0.001)	**0.101** (0.000)
		MonoFG	**0.894** (0.001)	**0.855** (0.001)	**0.815** (0.001)	**0.038** (0.000)	**0.069** (0.000)	**0.101** (0.001)
	CVD	**NeuralFG**	0.799 (0.010)	0.782 (0.005)	**0.758** (0.003)	**0.009** (0.000)	**0.021** (0.000)	**0.043** (0.000)
		MonoFG	**0.804** (0.010)	**0.785** (0.005)	**0.758** (0.004)	**0.009** (0.000)	**0.021** (0.000)	**0.043** (0.000)

Table 6: Comparison of model performance by means (standard deviations) across 5-fold cross-validation. Best performances are in **bold**.

	Risk	Model	C-Index *(Larger is better)* $q_{0.25}$	$q_{0.50}$	$q_{0.75}$	Brier Score *(Smaller is better)* $q_{0.25}$	$q_{0.50}$	$q_{0.75}$
Synthetic	1	$n = 1$	0.779 (0.012)	0.743 (0.013)	0.705 (0.010)	**0.076** (0.003)	**0.180** (0.004)	**0.344** (0.004)
		$\mathbf{n = 15}$	**0.792** (0.011)	**0.758** (0.014)	**0.724** (0.012)	0.079 (0.003)	0.186 (0.004)	0.355 (0.004)
		$n = 100$	0.791 (0.013)	**0.758** (0.014)	0.723 (0.012)	0.079 (0.003)	0.186 (0.004)	0.354 (0.004)
		$n = 1,000$	**0.792** (0.011)	**0.758** (0.013)	0.723 (0.011)	0.079 (0.003)	0.186 (0.004)	0.355 (0.004)
	2	$n = 1$	0.788 (0.016)	0.737 (0.021)	0.702 (0.017)	**0.073** (0.003)	**0.180** (0.005)	**0.338** (0.009)
		$\mathbf{n = 15}$	0.800 (0.014)	**0.754** (0.017)	**0.721** (0.016)	0.074 (0.003)	0.185 (0.005)	0.346 (0.009)
		$n = 100$	0.800 (0.013)	0.753 (0.016)	0.720 (0.015)	0.074 (0.003)	0.185 (0.004)	0.346 (0.008)
		$n = 1,000$	**0.801** (0.014)	0.753 (0.017)	0.720 (0.017)	0.075 (0.003)	0.185 (0.004)	0.347 (0.008)

Table 7: Impact of increasing n on DeSurv performances. Performance measured by means (standard deviations) across 5-fold cross-validation. Best performances are in **bold**.

C.2. Training speed

Finally, we examine the training and convergence speed for both DeSurv and NeuralFG on the Framingham dataset. We trained a fixed architecture with a total depth of 3 hidden layers composed of 50 nodes each. The learning rate was fixed at 0.001 and the batch size at 100. Table 8 presents the number of training iterations required to converge and the training time over 100 random splits of the data. We parallelised DeSurv's n forward passes following the original paper's recommendation. This set of experiments is performed on an Apple M1 Pro chip with 32 GB of memory.

The Desurv's results highlight that a coarser approximation ($n = 1$) requires more iterations to converge due to the lower-quality target loss, but each iteration is faster. Conversely, increasing n results in fewer iterations for convergence, but slower training. Echoing the theoretical computational

	Convergence Speed (in number of iterations)	Total Training Time (in seconds)
NeuralFG	91.98 (43.33)	13.60 (6.03)
MonoFG	66.26 (28.08)	**6.66** (2.90)
DeSurv ($n = 1$)	151.88 (123.50)	13.93 (11.07)
DeSurv ($\mathbf{n = 15}$)	55.09 (43.56)	56.68 (47.35)
DeSurv ($n = 100$)	**52.02** (24.45)	363.95 (172.55)

Table 8: Training speed comparison on the Framingham dataset. Performance measured by means (standard deviations) across 100-fold *Monte Carlo* cross-validation.

cost introduced in Section 3.4, our proposed methodology results in faster iterations, especially when considering a single network architecture for competing risks as shown by MonoFG's training time. However, the larger number of iterations required by our proposed methods in comparison to DeSurv reflects the more complex convergence of *constrained* monotonic neural networks.

References

Laura Antolini, Patrizia Boracchi, and Elia Biganzoli. A time-dependent discrimination index for survival data. *Statistics in medicine*, 24(24):3927–3944, 2005.

Erika Graf, Claudia Schmoor, Willi Sauerbrei, and Martin Schumacher. Assessment and comparison of prognostic classification schemes for survival data. *Statistics in medicine*, 18(17-18):2529–2545, 1999.

Chirag Nagpal, Willa Potosnak, and Artur Dubrawski. auton-survival: an open-source package for regression, counterfactual estimation, evaluation and phenotyping with censored time-to-event data. In Zachary Lipton, Rajesh Ranganath, Mark Sendak, Michael Sjoding, and Serena Yeung, editors, *Proceedings of the 7th Machine Learning for Healthcare Conference*, volume 182 of *Proceedings of Machine Learning Research*, pages 585–608. PMLR, 05–06 Aug 2022.

Sebastian Pölsterl. scikit-survival: A library for time-to-event analysis built on top of scikit-learn. *The Journal of Machine Learning Research*, 21(1):8747–8752, 2020.