# Neural Fine-Gray: Monotonic neural networks for competing risks

**Vincent Jeanselme** VINCENT.JEANSELME@MRC-BSU.CAM.AC.UK
*The Alan Turing Institute*
*MRC Biostatistics Unit, University of Cambridge*

**Chang Ho Yoon** CHANGHO.YOON@SJC.OX.AC.UK
*The Alan Turing Institute*
*Big Data Institute, University of Oxford*

**Brian Tom** BRIAN.TOM@MRC-BSU.CAM.AC.UK
**Jessica Barrett** JESSICA.BARRETT@MRC-BSU.CAM.AC.UK
*MRC Biostatistics Unit, University of Cambridge*

## Abstract

Time-to-event modelling, known as survival analysis, differs from standard regression as it addresses *censoring* in patients who do not experience the event of interest. Despite competitive performances in tackling this problem, machine learning methods often ignore other *competing risks* that preclude the event of interest. This practice biases the survival estimation. Extensions to address this challenge often rely on parametric assumptions or numerical estimations leading to sub-optimal survival approximations. This paper leverages constrained monotonic neural networks to model each competing survival distribution. This modelling choice ensures the exact likelihood maximisation at a reduced computational cost by using automatic differentiation. The effectiveness of the solution is demonstrated on one synthetic and three medical datasets. Finally, we discuss the implications of considering competing risks when developing risk scores for medical practice.

**Data and Code Availability** Experiments are performed on publicly available datasets: Primary Biliary Cholangitis[1] (Therneau et al., 2000), Framingham[2] (Kannel and McGee, 1979), Synthetic[3] (Lee et al., 2018), and the Surveillance, Epidemiology, and End Results Program[4]. The code to reproduce the proposed model and the presented results is available on GitHub[5].

**Institutional Review Board (IRB)** This research does not require IRB approval as it relies on publicly available datasets from previous studies.

## 1. Introduction

### 1.1. Motivation

Survival analysis involves modelling the time to an event of interest, which plays a critical role in medicine to understand disease manifestation, treatment outcomes, and the influence of different risk factors on patient health (Selvin, 2008). This analysis differs from standard regression settings as patients may not experience the outcome of interest over the study period. These *censored* patients inform this regression as they participate in the study event-free until exiting the study. Multiple approaches have been proposed to take advantage of these patients by maximising the likelihood of the observed data.

Often, in medical data, patients may experience events, known as *competing risks*, that preclude the observation of the event of interest. For instance, in modelling the time to cardiac events, patients who die from another condition during the observation period exit the study because of a competing risk. Competing

1. Available in the R survival package.
2. Available in the R riskCommunicator package.
3. Available at https://github.com/chl8856/DeepHit
4. Available at https://seer.cancer.gov/
5. https://github.com/Jeanselme/NeuralFineGray

risks remain overlooked despite their prevalence in medicine (Koller et al., 2012; Austin et al., 2016b). Particularly, practitioners frequently consider competing risks as censoring (Austin and Fine, 2017a). This practice breaks the common assumption of non-informative censoring, i.e., censored patients must leave the study for reasons independent of the outcome of interest. Considering competing risks as censoring, therefore, results in misestimating the risk of the event of interest (Fisher and Kanarek, 1974; Leung et al., 1997).

To better tackle the problem of competing risks, one can explicitly model them through the marginal probability of observing each risk, known as the Cumulative Incidence Function (CIF). Estimation of these functions often relies on proportional hazards, parametric assumptions, or numerical integration, potentially resulting in the optimisation of a sub-optimal target misrepresenting the true underlying survival distribution.

### 1.2. Contribution

This work introduces a novel machine learning model to tackle the problem of competing risks. This approach generalises Rindt et al. (2022) to competing risks, leveraging monotonic neural networks to model cumulative incidence functions. The proposed method tackles the limitations of existing strategies by an exact computation of the likelihood at a lower computational cost.

First, we explore the existing literature before introducing in detail our proposed model. Subsequently, we demonstrate the advantages and limitations of our approach as applied to one synthetic and three real-world medical datasets. Finally, we further investigate the Framingham dataset to underline the importance of considering competing risks in cardiovascular disease risk estimation.

## 2. Related work

This section summarises the recent progress in machine learning for survival analysis.

### 2.1. Time-to-event modelling

Survival analysis is an active field of research in the statistical community (Kartsonaki, 2016). Non-parametric (Ishwaran et al., 2008) and parametric (Cox, 2008; Royston, 2001; Cox et al., 2007) models have been introduced to model survival outcomes. Despite these multiple alternatives and considerable proposed extensions, the original Cox proportional-hazards model (Cox, 1972) remains widely used in the medical literature (Stensrud and Hernán, 2020). This semi-parametric approach estimates the impact of covariates on the instantaneous risk of observing an event, i.e., hazard. The model assumes the hazard to take the form of the product of a non-parametric estimate of the population survival and a parametric covariate effect. This assumption is known as proportional hazards and renders tractable the model optimisation for covariate effect estimation.

The machine learning community has extended the Cox model for unknown parametric forms of covariate effect. Specifically, DeepSurv (Katzman et al., 2018) replaces this otherwise parametric component with a neural network. However, this model still assumes proportional hazards that may not hold in real-world medical settings (Stensrud and Hernán, 2020). To relax this assumption, DeepCox (Nagpal et al., 2021c) identifies subgroups using independent Cox models. Each subgroup is characterised by its own non-parametric baseline and covariate effect. At the intersection between DeepCox and parametric models, Nagpal et al. (2021b) model each subgroup with a Weibull distribution parameterised by neural networks to allow end-to-end training. Jeanselme et al. (2022) abandon the parametric and proportional hazards assumption with unconstrained distributions learnt through monotonic networks.

With a focus on predictive performance, DeepHit (Lee et al., 2018) approaches survival as a classification problem where survival prediction time is discretised. The associated task is to predict the interval at which a patient experiences the event. The model's training procedure consists of a likelihood and a ranking penalty which favours temporally coherent predictions. Extrapolation of this model to infinite time discretisa-

tion resembles an ordinary differential equation (ODE), as proposed in Danks and Yau (2022).

The models above approximate the underlying survival likelihood either through parametric assumptions, discretisation or numerical integration. Recently, Rindt et al. (2022) proposed to overcome this challenge of likelihood estimation by deploying a constrained neural network with a monotonically increasing outcome to obtain the survival function, and, therefore, the exact likelihood. In addition, to show improved performance, the authors demonstrate that one should prefer likelihood optimisation over discriminative performance as the optimal likelihood is obtained for the true underlying survival distribution, i.e., the likelihood is a proper scoring rule. Our study is a generalisation of this work to competing risks, harnessing monotonic neural networks to directly model CIFs.

### 2.2. Modelling competing risks

Using the aforementioned models without consideration of competing risks would lead to a misestimation of the risk associated with the event of interest (Schuster et al., 2020). To tackle this issue, one can independently estimate each competing-risk-specific model and combine them to estimate the risk associated with a specific outcome given the non-observation of the other risks, as formulated in the cause-specific Cox model (Prentice et al., 1978). This independent estimation describes how covariates impact each event risk (Austin and Fine, 2017b) but may misrepresent the relative effect of these covariates on outcomes (Austin et al., 2016b) and lead to sub-optimal predictive performance. Alternatively, Fine and Gray (1999) propose to model the sub-hazards, i.e., the probability of observing a given event if the patient has not experienced this event until $t$, under an assumption of proportionality analogous to the one made in the Cox proportional-hazards model. While providing insights into the link between covariates and risk particularly suitable for prediction (Austin and Fine, 2017b), this model suffers from two shortcomings: (i) the proportionality assumption impairs its real-world applicability; (ii) this approach can result in an ill-defined survival function (Austin et al., 2021).

Machine learning approaches have been extended to jointly model competing risks. DeepHit's time-discretisation results in a straightforward extension in which the output dimension is multiplied by the number of risks (Lee et al., 2018). Similarly, hierarchical discretisation (Tjandra et al., 2021) has been proposed. As parametric distributional assumptions result in a closed-form likelihood, Nagpal et al. (2021b) propose to extend their mixture of Weibull distributions and Bellot and Schaar (2018) introduce a Bayesian mixture of Generalised Gamma distributions to tackle competing risk. Under more complex non-parametric likelihoods, numerical integration (Danks and Yau, 2022; Aastha and Liu, 2020) and pseudo-value approximations (Rahman et al., 2021) have been proposed. Finally, non-likelihood-based approaches have been introduced such as boosted trees (Bellot and van der Schaar, 2018) or survival trees (Schmid and Berger, 2021). However, these methods are optimised towards a Brier-score-like loss.

While survival analysis has received considerable attention in the machine learning community, the problem of competing risks is less well studied (Wang et al., 2019) and even less applied (Monterrubio-Gómez et al., 2022), despite being central to medical applications. The existing methodologies to tackle competing risks rely on parametric assumptions, likelihood approximation, or optimise for a score that may misrepresent the true underlying survival distribution. This paper offers a novel competing risk model relying on constrained networks to obtain CIFs as a derivative instead of an integral. This approach results in the exact maximisation of the likelihood by leveraging automatic differentiation.

## 3. Proposed approach

This section formalises the problem of survival analysis and introduces the proposed model.

### 3.1. Notation

We model a population of the form $\{x_i, t_i, d_i\}_i$ with $x_i$ the covariates for patient $i$, $t_i \in \mathbb{R}^+$ the time of end of follow-up and $d_i \in [\![0, R]\!]$ its associated cause. If $d_i \in [\![1, R]\!]$, the patient left the study due to one of the $R$ considered risks.

Otherwise, the patient is right-censored, i.e., the patient left the study for an *unrelated* reason before any of the events of interest were observed. In this work, we focus on right-censoring, but the model can easily be extended to left-censoring. Note that we assume that experiencing one event precludes the observation of any other.

### 3.2. Survival quantities

**Single risk.** In settings with no competing risk, i.e., $R = 1$, one aims to estimate the *survival function* $S$, the probability of not observing the event of interest before time $t$, i.e.:

$$S(t|x) := \mathbb{P}(T \geq t|x)$$

Equivalently, one aims to estimate the *cumulative hazard function* $\Lambda(t|x)$ related to $S$ as follows:

$$S(t|x) := \exp\left[-\Lambda(t|x)\right] = \exp\left[-\int_0^t \lambda(u|x)du\right]$$

where $\lambda(t|x) = \lim_{\delta t \to 0} \frac{\mathbb{P}(t<T<t+\delta t, |T \geq t, x)}{\delta t}$ is the instantaneous hazard of observing the event of interest, assuming no previous event(s).

Estimating this quantity may rely on maximising the likelihood of the observed data. The assumption of non-informative censoring, i.e., event and censoring times are independent given the covariates, is necessary to express the likelihood. Specifically, each patient $i$ with an observed event contributes to the likelihood, the probability of experiencing the event at $t_i$ without previous events, i.e., $\lambda(t_i|x_i)S(t_i|x_i)$. The likelihood associated with each censored patient is the probability of not experiencing the event until $t_i$, i.e., $S(t_i|x_i)$. This results in the following log-likelihood:

$$l = \sum_{i, d_i \neq 0} \log \lambda(t_i|x_i) - \sum_i \Lambda(t_i|x_i) \tag{1}$$

**Competing risks.** In the context of competing risks $R > 1$, a patient may leave a study for reasons correlated with the event of interest. Practitioners often consider these events as censoring and rely on single-risk models. However, this practice breaks the common assumption of non-informative censoring and results in misestimation of the survival function. When other events may be observed, $S(t \mid x)$ is defined as the probability of observing none of the competing risks before time $t$, i.e.:

$$S(t|x) = 1 - \sum_{r \in [\![1, R]\!]} F_r(t|x)$$

where $F_r$, the *Cumulative Incidence Function* (CIF) for the event $r$ denotes the probability of observing the event $r$ before time $t$ without prior occurrence of any competing event(s), i.e.:

$$F_r(t|x) = \mathbb{P}(T < t, \text{risk} = r|x) \tag{2}$$

with $T$, the random variable denoting the time of observation of any event. Note that the CIF can be expressed as an integral of observing the event in an infinitesimal interval given that no other event was observed until $t$:

$$F_r(t|x) = \int_0^t \lambda_r(u|x) e^{-\int_0^t \sum_r \lambda_r(s)ds} du \tag{3}$$

with $\lambda_r(t|x) = \lim_{\delta t \to 0} \frac{\mathbb{P}(t<T<t+\delta t,\ \text{risk}=r|T \geq t,\ x)}{\delta t}$, the cause-specific hazard, i.e., the instantaneous risk of observing the event $r$, with no other previous event.

A final quantity of interest is the cause-specific survival $S_r(t|x)$ that expresses the probability of not observing a given outcome $r$ by time $t$, i.e.,

$$\begin{aligned} S_r(t|x) &= \mathbb{P}((T \geq t) \cup (T < t, \text{risk} \neq r)|x) \\ &= 1 - F_r(t|x) \end{aligned}$$

Similar to the single-risk settings, we maximise the likelihood to estimate $F_r$. Importantly, we assume non-informative censoring *once controlled* on all identified competing risks. While this assumption is more likely to hold once all competing risks are accounted for, practitioners suspecting its implausibility should perform sensitivity analysis for this assumption (Jackson et al., 2014). Under this assumption, the likelihood can be expressed analogously to (1): patients with an observed event contribute to the likelihood as the probability of observing the event $d_i$ at $t_i$ without observing any events until $t_i$, i.e., $\lambda_r(t_i|x_i)S(t_i|x_i)$. This quantity is the partial derivative of $F_r$ with respect to $t$ evaluated at $t_i$. Remaining censored patients influence the likelihood as the probability of observing no event until $t_i$, i.e., $S(t_i|x_i)$. The competing risks log-likelihood can, therefore, be expressed as:

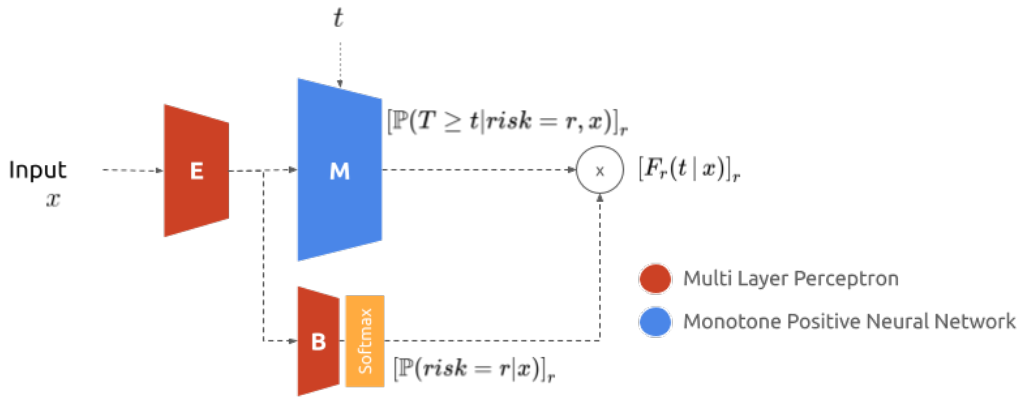


Figure 1: Neural Survival Analysis Architecture. *E embeds the covariate(s) x, which are then inputted in the monotonic networks M and balancing network B to estimate the CIFs.*

$$l = \sum_{r\in[\![1,R]\!]} \sum_{i,d_i=r} \log \frac{\partial F_r(u|x_i)}{\partial u}\Big|_{u=t_i} + \sum_{i,d_i=0} \log[1 - \sum_r F_r(t_i|x_i)] \quad (4)$$

One may extend existing models to the competing risks setting by performing the integration in (3). For instance, the cause-specific Cox model (Prentice et al., 1978) consists of Cox models independently trained on each risk, i.e., treating all other outcomes as censored. Then one evaluates the CIF through (3) using the estimated hazards. However, this staged modelling does not jointly consider the outcomes and may misestimate the covariate effects (Van Der Pas et al., 2018). Fine-Gray (Fine and Gray, 1999) overcomes this issue by directly modelling the sub-distribution hazards $h_r(t|x) = \lim_{\delta t\to 0} \frac{\mathbb{P}(t<T<t+\delta t,\ \text{risk}=r|(T\geq t)\ \cup\ (T<t\ \cap\ \text{risk}\neq r),\ x)}{\delta t}$, relying on a proportionality assumption of these quantities.

Likewise, one can extend machine learning architecture to enable the integration of the CIF and maximise the associated likelihood in (4). However, in the absence of a closed-form expression, this would necessitate numerical integration. This approximation may impact performance with added computational costs for training and predictions. Integration is computationally expensive, whereas derivation can be computed exactly in one backward pass by automatic differentiation – available in most machine learning libraries. Therefore, our approach reduces the computational cost of the likelihood estimation by modelling $F_r$ and differentiating it to obtain $\lambda_r S$, resulting in the exact computation of all the previously described quantities of interest.

### 3.3. Architecture

Neural Fine-Gray, illustrated in Figure 1, aims to model $[F_r]_{r\in[\![1,R]\!]}$ without relying on numerical integration to tackle the problem of competing risks. We decompose $F_r$ as:

$$\begin{aligned} F_r(t|x) &= \mathbb{P}(\text{risk } = r|x) \cdot \mathbb{P}(T \leq t|\text{risk } = r, x) \\ &= B(E(x))_r \cdot [1 - \exp(-t \times M_r(t, E(x)))] \end{aligned}$$

**Embedding network ($E$).** A first multi-layer perceptron $E$ with inter-layer dropout extracts an embedding $\tilde{x}$ from the covariates $x$.

**Sub-distribution networks ($[M_r]_{r\in[\![1,R]\!]}$).** The embedding $\tilde{x}$ is inputted in $R$ positive monotonic networks $[M_r]_{r\in[\![1,R]\!]}$ representing a lifetime distribution conditioned on one risk $r$, through the relation $1-\exp(-t\times M_r(t,\tilde{x})) = \mathbb{P}(T \leq t|x, \text{risk } = r)$. A *positive monotonic neural network* is a network constrained to have its outcome monotonic and positive given its input (see Daniels and Velikova (2010) for theoretical analysis and Lang (2005) for proof of universal approximator). Enforcing these constraints may rely on different transformations of the neural networks' weights (Omi

et al., 2019; Rindt et al., 2022; Chilinski and Silva, 2020). In our work, we enforce all the neural networks' weights to be positive through a square function and use a final *SoftPlus* layer to fulfil these constraints. Enforcing positive weights ensures that the outcome increases with the time dimension $t$. Additionally, enforcing a smooth function ensures a low computational cost and stable optimisation. Note that for model flexibility, we used $R$ monotonic networks. We explore in Appendix B how using one network with $R$ outcomes would impact performance.

**Balancing network ($B$).** A multi-layer perceptron $B$ with a final *SoftMax* layer leverages $\tilde{x}$ to balance the probability of observing each risk $B(\tilde{x}) := [\mathbb{P}(\text{risk} = r|x)]_r$. This weighting ensures that the survival function is correctly specified, i.e., $\sum_{r\in[\![1,R]\!]} F_r(t|x) \leq 1$.

The proposed approach directly models $F_r$ by multiplying the outputs of the distribution and balancing networks. Automatic differentiation of the model's output results in the derivative $\frac{\partial F_r(u|x_i)}{\partial u}\Big|_{u=t_i}$. The model can then be trained end-to-end by maximising the *exact* log-likelihood proposed in Equation (4). By jointly modelling the competing risks, this proposed model is reminiscent of the Fine-Gray approach. The following equation exhibits the link between sub-distribution hazards and CIFs, i.e., between the standard and neural Fine-Gray models:

$$h_r(t|x) = \frac{1}{1 - F_r(t|x)} \cdot \frac{\partial F_r(u|x)}{\partial u}\Big|_{u=t}$$

**Remark 1** *Shchur et al. (2020) raise a limitation of monotonic neural networks that may attribute non-null density to negative times, i.e., $F_r(t = 0|x) \neq 0$. In contrast to Omi et al. (2019); Rindt et al. (2022), we model $\mathbb{P}(T \leq t|risk = r, x)$ as $1 - \exp(-t \times M_r(t, \tilde{x}))$ instead of $M_r(t, \tilde{x})$ to address this issue.*

**Remark 2** *The proposed methodology is a generalisation of the survival model Sumo-Net (Rindt et al., 2022) that estimates $S$ in the single-risk setting. If $R = 1$, then $F_r = 1 - S$ and $B_r = \mathbb{1}$. In this context, the proposed approach results in Sumo-Net. Moreover, the architecture resembles the one proposed in DeSurv (Danks and Yau, 2022) while avoiding numerical integration.*

### 3.4. Computational complexity

Our modelling choices result in the exact computation of the likelihood. However, the other methodologies relying on integral approximation and outcome discretisation converge towards $F_r$ in the upper limit, i.e., when increasing the number of point estimates, or using a finer discretisation. One may therefore question the advantage of the proposed methodology. In this section, we compare the complexity in estimating the CIF and likelihood for DeSurv (Danks and Yau, 2022), the closest method to our proposed model, and NeuralFG.

**DeSurv** (Danks and Yau, 2022). This approach models $F_r(t|x)$ as $\text{Tanh}(v(x,t))$ with $v$ being the solution to the ODE defined as $\frac{\partial v(x,u)}{\partial u}\Big|_{u=t} = g(x,t)$ and $v(x,0) = 0$ with $g$, a neural network. For efficiency, the authors propose a Gauss-Legendre quadrature to solve the ODE and obtain $v$. This approximation necessitates $n$ evaluations of $g$ at defined times $[t_j(t)]_{j\in[\![1,n]\!]}$ weighted by the associated $[w_j]_{j\in[\![1,n]\!]}$ (see Press et al. (2007) for a detailed description of Gauss-Legendre quadrature). Each forward pass estimates $\frac{\partial v(x,u)}{\partial u}\Big|_{u=t_j(t)}$ at the points used to approximate the integral, then

$$\hat{F}_r(t|x) = \text{Tanh}\left(\frac{t}{2}\sum_{j\in[\![1,n]\!]} w_j g(x, \frac{t}{2}t_j(t))\right)$$

**DeSurv's computational cost.** Computation of $F_r$ relies on $n$ forward passes through the network. Moreover, the estimation of $\frac{\partial \hat{F}_r(u|x_i)}{\partial u}\Big|_{u=t_i}$ necessary to compute the competing risk likelihood is $g(x,t_i)(1-\text{Tanh}(\hat{F}_r(t_i|x))^2)$, i.e., $n+1$ forward passes. The likelihood has a $\mathcal{O}(nN)$ computational complexity with $N$ the number of patients in the study.

**NeuralFG's computational cost.** $F_r$ and $\frac{\partial \hat{F}_r(u|x_i)}{\partial u}\Big|_{u=t_i}$ are estimated in the same forward

| Dataset | Observations | Features | Primary | Competing risk | Censored |
|---|---|---|---|---|---|
| PBC | 312 | 25 | Death (44.87 % ) | Transplant (9.29 %) | 45.83 % |
| Framingham | 4,434 | 18 | CVD (26.09 %) | Death (17.75 %) | 56.16 % |
| Synthetic | 30,000 | 12 | * (25.33 %) | * (24.67 %) | 50.00 % |
| SEER | 658,354 | 23 | BC (16.51 %) | CVD (5.69 %) | 77.80 % |

Table 1: Datasets characteristics

pass by automatic differentiation. The likelihood estimation has therefore a $\mathcal{O}(N)$ complexity. Our proposed methodology, therefore, presents more than an $n/2$ computational gain compared to DeSurv in estimating the likelihood used for training, and an $n$ gain in inferring $F_r$.

## 4. Experiments

This section introduces the datasets and experimental settings.

### 4.1. Datasets

We explore the model performance on four datasets with competing risks:

- PBC (Therneau et al., 2000) comprises 25 covariates in 312 patients over a 10-year randomised control trial to measure the impact of D-penicillamine on Primary Biliary Cholangitis (PBC). Death on the waiting list is the primary outcome with transplant being a competing risk.
- Framingham (Kannel and McGee, 1979) is a cohort study gathering 18 longitudinal measurements on patients over 20 years. Our analysis focuses on the first observed covariates of 4,434 patients to model cardiovascular disease (CVD) risk. Death from other causes is treated as a competing risk.
- Synthetic (Lee et al., 2018), this dataset consists of 30,000 synthetic patients with 12 covariates following exponential event time distributions, non-linearly dependent on the covariates.
- SEER[6]: the Surveillance, Epidemiology, and End Results Program gathers covariates and outcomes of patients diagnosed with breast cancer between 1992 and 2017. Following the preprocessing proposed by Lee et al. (2018); Danks and Yau (2022), we select 658,354 patients and 23 covariates describing the patient demographics and disease characteristics at diagnosis. Death from breast cancer (BC) is our primary outcome, with CVD, a competing risk.

Table 1 summarises the datasets' characteristics with the respective proportion of outcome and censoring.

### 4.2. Baseline models

The proposed Neural Fine-Gray (**NeuralFG**) was compared against six strategies. First, we considered the well-established cause-specific Cox model (**CS Cox** Prentice et al. (1978)) and **Fine-Gray** model (Fine and Gray, 1999) with a linear parametric form for the covariate effect. The cause-specific Cox model models each cause independently using a Cox proportional-hazards model, while Fine-Gray models the sub-hazard functions assuming proportional sub-hazards.

Thereafter, we compare state-of-the-art competing risk survival neural networks proposed in the machine learning literature. First, Deep Survival Machine (**DSM**, Nagpal et al. (2021b)) consists of a mixture of Weibull distributions parameterised by neural networks. Each point is then assigned to these distributions through an assignment network. Using parametric distributions results in a closed-form likelihood in the competing risks setting. **DeepHit** (Lee et al., 2018) discretises the survival horizon and leverages a multi-head network to associate each patient to the interval corresponding to its observed event time and type. Each head of the network is associated with one cause as in the proposed NeuralFG. The time-discretisation results in a discrete likelihood further penalised by a C-index-like regularisation for model training. Closer to our work,

6. https://seer.cancer.gov/

| | Risk | Model | C-Index *(Larger is better)* | | | Brier Score *(Smaller is better)* | | |
|---|---|---|---|---|---|---|---|---|
| | | | $q_{0.25}$ | $q_{0.50}$ | $q_{0.75}$ | $q_{0.25}$ | $q_{0.50}$ | $q_{0.75}$ |
| PBC | Death | **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) |
| | | DeepHit | 0.822 (0.099) | 0.844 (0.036) | 0.782 (0.033) | *0.090* (0.030) | 0.132 (0.013) | 0.180 (0.021) |
| | | DeSurv | 0.821 (0.089) | 0.837 (0.050) | 0.815 (0.068) | **0.088** (0.022) | 0.113 (0.011) | **0.136** (0.047) |
| | | DSM | **0.867** (0.065) | **0.864** (0.037) | **0.828** (0.052) | 0.091 (0.039) | 0.124 (0.015) | 0.161 (0.022) |
| | | Fine-Gray | 0.831 (0.136) | *0.852* (0.045) | *0.816* (0.059) | 0.091 (0.042) | *0.103* (0.009) | 0.150 (0.038) |
| | | CS Cox | *0.833* (0.125) | 0.851 (0.040) | 0.811 (0.065) | 0.091 (0.038) | **0.102** (0.008) | *0.148* (0.038) |
| Framingham | 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) |
| | | DeepHit | 0.855 (0.026) | 0.781 (0.026) | 0.743 (0.014) | 0.053 (0.003) | 0.102 (0.007) | 0.141 (0.002) |
| | | DeSurv | **0.872** (0.027) | *0.807* (0.031) | 0.775 (0.022) | **0.049** (0.005) | **0.095** (0.009) | *0.129* (0.003) |
| | | DSM | *0.866* (0.023) | 0.806 (0.023) | *0.778* (0.014) | 0.057 (0.005) | 0.104 (0.006) | 0.141 (0.002) |
| | | Fine-Gray | 0.842 (0.025) | 0.794 (0.024) | 0.772 (0.015) | 0.057 (0.006) | 0.099 (0.007) | 0.131 (0.003) |
| | | CS Cox | 0.845 (0.020) | 0.798 (0.022) | 0.774 (0.015) | 0.056 (0.006) | *0.098* (0.007) | 0.131 (0.003) |
| 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) |
| | | DeepHit | 0.783 (0.012) | 0.747 (0.013) | *0.714* (0.008) | 0.079 (0.003) | 0.136 (0.002) | *0.212* (0.003) |
| | | DeSurv | **0.793** (0.013) | **0.756** (0.014) | *0.714* (0.014) | **0.068** (0.002) | **0.124** (0.004) | **0.192** (0.004) |
| | | DSM | 0.776 (0.013) | 0.742 (0.013) | 0.710 (0.013) | *0.073* (0.002) | 0.139 (0.002) | 0.220 (0.003) |
| | | Fine-Gray | 0.611 (0.014) | 0.587 (0.007) | 0.568 (0.009) | 0.078 (0.002) | 0.159 (0.003) | 0.241 (0.002) |
| | | CS Cox | 0.609 (0.015) | 0.586 (0.006) | 0.568 (0.009) | 0.078 (0.002) | 0.159 (0.003) | 0.240 (0.002) |
| 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) |
| | | DeepHit | **0.899** (0.002) | **0.860** (0.001) | **0.818** (0.001) | **0.038** (0.000) | *0.070* (0.000) | *0.102* (0.001) |
| | | DeSurv | 0.892 (0.003) | 0.852 (0.002) | 0.813 (0.001) | **0.038** (0.000) | *0.070* (0.000) | *0.102* (0.001) |
| | | DSM | 0.884 (0.001) | 0.842 (0.002) | 0.805 (0.002) | *0.039* (0.000) | 0.076 (0.001) | 0.112 (0.000) |
| | | Fine-Gray | 0.836 (0.003) | 0.786 (0.003) | 0.742 (0.002) | 0.043 (0.001) | 0.081 (0.000) | 0.118 (0.000) |
| | | CS Cox | 0.837 (0.003) | 0.786 (0.003) | 0.742 (0.002) | 0.042 (0.001) | 0.081 (0.000) | 0.118 (0.000) |

Table 2: Comparison of model performance by means (standard deviations) across 5-fold cross-validation. Best performances are in **bold**, second best in *italics*. *NeuralFG is the model introduced in this paper.*

**DeSurv** (Danks and Yau, 2022) approaches $F_r$ as the solution to an ODE.

### 4.3. Experimental settings

The analysis relies on 5-fold cross-validation with 10% of each training set left aside for hyper-parameter tuning. Random search is used on the following grid over 100 iterations: learning rate ($10^{-3}$ or $10^{-4}$), batch size (100, 250, except for SEER: $1,000$ or $5,000$), dropout rate (0, 0.25, 0.5 or 0.75), number of layers ($[\![1, 4]\!]$) and nodes (25 or 50). All activation functions are fixed to *Tanh* to ensure a properly defined derivative – note that any $\mathcal{C}^1$ activation would work. All models are optimised using an Adam optimiser (Kingma and Ba, 2015) over $1,000$ epochs, with an early stopping criterion computed on a 10% left-aside subset of the training set.

Other methods are optimised over the same grid (if applicable). Additionally, we explore both Log-Normal and Weibull distributions for DSM and use $10,000$ warm-up iterations to estimate the parametric form closest to the average survival as proposed in the original paper (Nagpal et al., 2021b). For DeSurv, we followed the original paper's recommendation of a 15-degree Gauss-Legendre quadrature to estimate the CIFs. In Appendix C.1, we further investigate how increasing the number of point estimates impacts performance. We use a similar approximation for DeepHit with a 15-split time discretisation. Finally, for a fair compari-

| Death | Model | C-Index *(Larger is better)* | | | Brier Score *(Smaller is better)* | | |
|---|---|---|---|---|---|---|---|
| | | $q_{0.25}$ | $q_{0.50}$ | $q_{0.75}$ | $q_{0.25}$ | $q_{0.50}$ | $q_{0.75}$ |
| CVD | **Competing** | **0.872** (0.024) | **0.812** (0.029) | **0.782** (0.018) | **0.050** (0.003) | **0.095** (0.010) | **0.128** (0.004) |
| | Non-Competing | 0.862 (0.029) | 0.807 (0.032) | 0.780 (0.020) | 0.053 (0.004) | 0.099 (0.011) | 0.129 (0.005) |
| Death | **Competing** | **0.745** (0.055) | 0.717 (0.038) | 0.713 (0.022) | **0.027** (0.003) | **0.070** (0.004) | 0.112 (0.005) |
| | Non-Competing | 0.741 (0.053) | **0.718** (0.045) | **0.719** (0.025) | **0.027** (0.003) | 0.071 (0.002) | **0.109** (0.004) |

Table 3: Modelling competing risk - means (standard deviations) across the 5-fold cross-validation.

son, we double the number of possible layers for architectures without embedding networks.

### 4.4. Evaluation metrics

As per current practice in survival literature, we used the time-dependent Brier score (Graf et al., 1999) to quantify calibration, and the C-index (Antolini et al., 2005) for discrimination at the dataset-specific 0.25, 0.5 and 0.75 quantiles of the uncensored population event times (See Appendix A.1 for data characteristics, A.2 for further description of the metrics and A.4 for the cumulative version of these metrics). Means and standard deviations are computed over the 5 folds of cross-validation.

## 5. Results

Table 2 summarises the calibration and discriminative performance of the analysed models on the primary outcome (see Appendix A.3 for the performances on the competing risk).

### 5.1. Model's strengths

NeuralFG demonstrates lower or equal Brier scores than other state-of-the-art machine learning models across the majority of datasets and time horizons.

While DSM presents good discriminative performances, this edge is not reflected in its calibration. This observation indicates that parametric assumptions may result in estimated survival functions discriminative of the outcome but further from the underlying survival distribution. Deep-Hit penalisation results in better C-Index values but hurts model calibration, with misaligned discrimination and calibration throughout the different datasets.

Finally, performances are comparable to DeSurv. However, DeSurv's likelihood approximation multiplies its computational cost by the numerical integration complexity (see Appendix C.2 for a comparison of training speed on the Framingham dataset). NeuralFG, therefore, achieves state-of-the-art performance while avoiding computationally-expensive approximations.

### 5.2. Model's limitations

The proposed methodology has lower performance on the PBC dataset, which notably comprises a limited amount of data. In small-data settings, practitioners should prefer simpler models to avoid overfitting.

For instance, the linear Fine-Gray and CS Cox models result in competitive performances on PBC. However, this linearity assumption hurts performance under more complex covariate effects as in the SEER and Synthetic datasets. Note that leveraging domain expertise could enhance performance through the addition of interactions and the use of alternative models. However, these approaches deviate from the automated discovery of interactions facilitated by neural networks. Similarly, the parametric assumption of DSM results in the best discrimination in PBC, but it under-performs under more complex survival distributions.

Furthermore, the DeSurv model performs better than the proposed methodology on PBC. This may reflect that approximating the likelihood can regularise model training, which is beneficial in the context of small data.

### 5.3. Modelling vs ignoring competing risks

This last section explores the importance of modelling competing risks in the Framingham dataset. First, we present the performance differences between the proposed model in comparison to the same architecture maximising the cause-specific likelihoods. Then, we explore which subgroups of the population most benefit from this modelling. Finally, we study how guidelines would differ under the proposed NeuralFG and its non-competing alternative.

**Why account for competing risks?** To measure how modelling competing risks impacts performance, while ensuring the *same number of parameters*, we propose to use the same architecture presented in Section 3.3 whilst maximising the sum of the cause-specific likelihoods, i.e.:

$$l = \sum_r \left[ \sum_{i, d_i = r} \log \lambda_r(t_i | \tilde{x}_i) - \sum_i \Lambda_r(t_i | \tilde{x}_i) \right]$$

Each monotonic network, therefore, models the cumulative hazard function for risk $r$, $\Lambda_r$, by maximising the likelihood of one cause whilst considering the rest of the population as censored, relying on a shared embedding $\tilde{x}$. Automatic differentiation outputs $[\lambda_r]_{r \in [\![1, R]\!]}$. Table 3 summarises the discrimination and calibration differences in the non-competing survival $e^{-\Lambda_r(t|x)}$ obtained with this model and the previously described NFG's cause-specific survival $1 - F_r(t|x)$. Note how modelling competing risks significantly improves performance for the primary outcome of interest, CVD, without significant differences for the competing risk. Since patients who die from other causes during the study period do not present the same risk of CVD as patients remaining in the study, not accounting for all-cause mortality results in a misestimation of CVD risk.

**Who may benefit?** One can explore which subgroups benefit the most from modelling competing risks. Intuitively, patients who are the most likely to suffer from competing risks may benefit the most from this modelling. Table 4 illustrates this with older patients benefiting the most from modelling death as a competing risk.

| Age | Brier Score Difference | | |
|---|---|---|---|
| | $q_{0.25}$ | $q_{0.50}$ | $q_{0.75}$ |
| $< 40$ | -0.000 (0.000) | -0.001 (0.002) | 0.000 (0.005) |
| 40-50 | -0.001 (0.001) | -0.002 (0.003) | -0.002 (0.001) |
| 50-60 | *-0.003* (0.005) | *-0.004* (0.003) | *-0.006* (0.007) |
| 60+ | **-0.013** (0.011) | **-0.022 (0.018)** | **-0.007** (0.024) |

Table 4: Calibration differences - Means and standard deviations over 5-fold cross-validation. *Larger negative values correspond to better calibration for the competing risk model.*

**What is the impact on medical practice?** The Framingham dataset was used to model the eponymous 10-year cardiovascular disease (CVD) risk score (Wilson et al., 1998). This score guides clinical practice in preventatively treating patients, usually with a combination of cholesterol-lowering therapy, e.g., statins, and holistic treatment of other CVD risk factors (Bosomworth, 2011). To minimise overtreatment and adverse side effects, accurate risk estimates are critical for targeting the population most at risk so as to maximise the benefit-risk ratio (Mangione et al., 2022). However, the original Framingham score relies on a non-competing risk model (Mangione et al., 2022; van Kempen et al., 2014).

Clinical treatment often relies on a discretisation of this risk (Bosomworth, 2011): low, intermediate and high risk, at $< 10\%$, $10 - 20\%$ and $> 20\%$ chance, respectively, of observing a CVD event in the following 10 years. Current guidelines in the United States suggest placing all patients with $\geq 10\%$ risk on cholesterol-lowering drugs (Mangione et al., 2022). Furthermore, in the US alone, several million patients are on these medications (Wall et al., 2018). Therefore, even modest shifts in patient risk classification could, at scale, amount to considerable numbers either inappropriately receiving preventative treatment or inappropriately receiving none. To demonstrate how considering competing risks can fundamentally alter such risk profiling, we present in Table 5 the reclassification matrices of risk levels given competing and non-competing NeuralFG differentiated by observed outcomes for patients aged 50 or over. For instance, note that 251 deemed intermediate-to-

Table 5: Reclassification matrices between competing and non-competing risk scores for patients older than 50. *Red (resp. blue) shows when the competing risks score is less aligned with the 10-year observed outcome than the non-competing model (resp. more aligned). Note that censored patients are ignored.*

(*a*) Patients *with no event* in the 10-year follow-up.

| | | Non - Competing | | | |
|---|---|---|---|---|---|
| | Risk | Low | Inter. | High | Total |
| Comp. | Low | 502 | 228 | 23 | 753 |
| | Intermediate | 2 | 189 | 229 | 420 |
| | High | 0 | 9 | 314 | 323 |
| | Total | 504 | 426 | 566 | 1496 |

(*b*) Patients *with* an observed event during the 10-year follow-up.

| | | Non - Competing | | | |
|---|---|---|---|---|---|
| | Risk | Low | Inter. | High | Total |
| Comp. | Low | 23 | 28 | 5 | 56 |
| | Intermediate | 1 | 37 | 41 | 79 |
| | High | 2 | 4 | 248 | 254 |
| | Total | 26 | 69 | 294 | 389 |

high risk by the non-competing risks model are reclassified as lower risk by the competing-risks model, who, in turn, could have avoided the initiation of therapy. These results echo the medical literature's findings of risk misestimation due to the non-consideration of competing risks in this risk score (Lloyd-Jones et al., 2004; van Kempen et al., 2014). More accurate simulations to estimate the potential lives saved and harmed through such reclassification is beyond the scope of this article but could provide insight into the possible consequences of considering competing risks. In summary, using a non-competing risk score would have important clinical consequences of over- and under-treatment (Schuster et al., 2020). More predictive models accounting for competing risks must be preferred to ensure better care.

## 6. Conclusion

This work provides a solution to address competing risks that preclude the observation of the outcome of interest, often present in medical applications. We introduce Neural Fine-Gray, a monotonic neural network architecture, to tackle the problem of competing risks in survival modelling. The model outputs the cumulative incidence functions and, consequently, allows the exact likelihood computation. Importantly, this architecture choice achieves competitive performance while avoiding the parametric assumptions or computationally expensive approximations made by state-of-the-art survival neural networks. Further analyses of the Framingham dataset contribute to the literature, inviting practitioners to use competing-risk modelling in risk score development for improved care (Abdel-Qadir et al., 2018; Austin et al., 2016a; Lloyd-Jones et al., 2004; Schuster et al., 2020).

Our future work will (i) extend this architecture to model other modalities such as time series as in Nagpal et al. (2021a) and, (ii) explore medically interpretable survival clusters as presented in Jeanselme et al. (2022); Nagpal et al. (2022a).

## Acknowledgments

This work was supported by The Alan Turing Institute's Enrichment Scheme and partially funded by UKRI Medical Research Council (MC_UU_00002/5 and MC_UU_00002/2).

## References


Pengyu Huang Aastha and Yan Liu. Deepcompete: A deep learning approach to competing risks in continuous time domain. In *AMIA Annual Symposium Proceedings*, volume 2020, page 177. American Medical Informatics Association, 2020.

Husam Abdel-Qadir, Jiming Fang, Douglas S Lee, Jack V Tu, Eitan Amir, Peter C Austin, and Geoffrey M Anderson. Importance of considering competing risks in time-to-event analyses: application to stroke risk in a retrospective cohort study of elderly patients with atrial fibrillation. *Circulation: Cardiovascular Quality and Outcomes*, 11(7):e004580, 2018.

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

Peter C Austin and Jason P Fine. Accounting for competing risks in randomized controlled trials: a review and recommendations for improvement. *Statistics in medicine*, 36(8):1203–1209, 2017a.

Peter C Austin and Jason P Fine. Practical recommendations for reporting f ine-g ray model analyses for competing risk data. *Statistics in medicine*, 36(27):4391–4400, 2017b.

Peter C Austin, Douglas S Lee, Ralph B D'Agostino, and Jason P Fine. Developing points-based risk-scoring systems in the presence of competing risks. *Statistics in medicine*, 35(22):4056–4072, 2016a.

Peter C Austin, Douglas S Lee, and Jason P Fine. Introduction to the analysis of survival data in the presence of competing risks. *Circulation*, 133(6):601–609, 2016b.

Peter C Austin, Ewout W Steyerberg, and Hein Putter. Fine-gray subdistribution hazard models to simultaneously estimate the absolute risk of different event types: cumulative total failure probability may exceed 1. *Statistics in Medicine*, 40(19):4200–4212, 2021.

Alexis Bellot and Mihaela Schaar. Tree-based bayesian mixture model for competing risks. In *International Conference on Artificial Intelligence and Statistics*, pages 910–918. PMLR, 2018.

Alexis Bellot and Mihaela van der Schaar. Multitask boosting for survival analysis with competing risks. *Advances in Neural Information Processing Systems*, 31, 2018.

N John Bosomworth. Practical use of the framingham risk score in primary prevention: Canadian perspective. *Canadian Family Physician*, 57(4):417–423, 2011.

Pawel Chilinski and Ricardo Silva. Neural likelihoods via cumulative distribution functions. In *Conference on Uncertainty in Artificial Intelligence*, pages 420–429. PMLR, 2020.

Christopher Cox. The generalized f distribution: an umbrella for parametric survival analysis. *Statistics in medicine*, 27(21):4301–4312, 2008.

Christopher Cox, Haitao Chu, Michael F Schneider, and Alvaro Munoz. Parametric survival analysis and taxonomy of hazard functions for the generalized gamma distribution. *Statistics in medicine*, 26(23):4352–4374, 2007.

David R Cox. Regression models and life-tables. *Journal of the Royal Statistical Society: Series B (Methodological)*, 34(2):187–202, 1972.

Hennie Daniels and Marina Velikova. Monotone and partially monotone neural networks. *IEEE Transactions on Neural Networks*, 21(6):906–917, 2010.

Dominic Danks and Christopher Yau. Derivative-based neural modelling of cumulative distribution functions for survival analysis. In *International Conference on Artificial Intelligence and Statistics*, pages 7240–7256. PMLR, 2022.

Jason P Fine and Robert J Gray. A proportional hazards model for the subdistribution of a competing risk. *Journal of the American statistical association*, 94(446):496–509, 1999.

Lloyd Fisher and Paula Kanarek. Presenting censored survival data when censoring and survival times may not be independent. *Reliability and Biometry*, pages 303–326, 1974.

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.

Hemant Ishwaran, Udaya B Kogalur, Eugene H Blackstone, and Michael S Lauer. Random survival forests. 2008.

Dan Jackson, Ian R White, Shaun Seaman, Hannah Evans, Kathy Baisley, and James Carpenter. Relaxing the independent censoring assumption in the cox proportional hazards model using multiple imputation. *Statistics in medicine*, 33(27):4681–4694, 2014.

Vincent Jeanselme, Brian Tom, and Jessica Barrett. Neural survival clustering: Non-parametric mixture of neural networks for survival clustering. In *Conference on Health, Inference, and Learning*, pages 92–102. PMLR, 2022.

William B Kannel and Daniel L McGee. Diabetes and cardiovascular disease: the framingham study. *Jama*, 241(19):2035–2038, 1979.

Christiana Kartsonaki. Survival analysis. *Diagnostic Histopathology*, 22(7):263–270, 2016.

Jared L Katzman, Uri Shaham, Alexander Cloninger, Jonathan Bates, Tingting Jiang, and Yuval Kluger. Deepsurv: personalized treatment recommender system using a cox proportional hazards deep neural network. *BMC medical research methodology*, 18(1):1–12, 2018.

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *3rd International Conference on Learning Representations, ICLR 2015*, 2015.

Michael T Koller, Heike Raatz, Ewout W Steyerberg, and Marcel Wolbers. Competing risks and the clinical community: irrelevance or ignorance? *Statistics in medicine*, 31(11-12):1089–1097, 2012.

Bernhard Lang. Monotonic multi-layer perceptron networks as universal approximators. In *Artificial Neural Networks: Formal Models and Their Applications–ICANN 2005: 15th International Conference, Warsaw, Poland, September 11-15, 2005. Proceedings, Part II 15*, pages 31–37. Springer, 2005.

Changhee Lee, William Zame, Jinsung Yoon, and Mihaela van der Schaar. Deephit: A deep learning approach to survival analysis with competing risks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018.

Kwan-Moon Leung, Robert M Elashoff, and Abdelmonem A Afifi. Censoring issues in survival analysis. *Annual review of public health*, 18(1):83–104, 1997.

Donald M Lloyd-Jones, Peter WF Wilson, Martin G Larson, Alexa Beiser, Eric P Leip, Ralph B D'Agostino, and Daniel Levy. Framingham risk score and prediction of lifetime risk for coronary heart disease. *The American journal of cardiology*, 94(1):20–24, 2004.

Carol M Mangione, Michael J Barry, Wanda K Nicholson, Michael Cabana, David Chelmow, Tumaini Rucker Coker, Esa M Davis, Katrina E Donahue, Carlos Roberto Jaén, Martha Kubik, et al. Statin use for the primary prevention of cardiovascular disease in adults: Us preventive services task force recommendation statement. *JAMA*, 328(8):746–753, 2022.

Karla Monterrubio-Gómez, Nathan Constantine-Cooke, and Catalina A Vallejos. A review on competing risks methods for survival analysis. *arXiv preprint arXiv:2212.05157*, 2022.

Chirag Nagpal, Vincent Jeanselme, and Artur Dubrawski. Deep parametric time-to-event regression with time-varying covariates. In Russell Greiner, Neeraj Kumar, Thomas Alexander Gerds, and Mihaela van der Schaar, editors, *Proceedings of AAAI Spring Symposium on Survival Prediction - Algorithms, Challenges, and Applications 2021*, volume 146 of *Proceedings of Machine Learning Research*, pages 184–193. PMLR, 22–24 Mar 2021a.

Chirag Nagpal, Xinyu Li, and Artur Dubrawski. Deep survival machines: Fully parametric survival regression and representation learning for censored data with competing risks. *IEEE Journal of Biomedical and Health Informatics*, 2021b.

Chirag Nagpal, Steve Yadlowsky, Negar Rostamzadeh, and Katherine Heller. Deep cox mixtures for survival regression. *Machine Learning for Healthcare Conference*, 2021c.

Chirag Nagpal, Mononito Goswami, Keith Dufendach, and Artur Dubrawski. Counterfactual phenotyping with censored time-to-events. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, KDD '22, page 3634–3644, 2022a.

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 2022b.

Takahiro Omi, Kazuyuki Aihara, et al. Fully neural network based model for general temporal point processes. In *Advances in Neural Information Processing Systems*, pages 2122–2132, 2019.

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.

Ross L Prentice, John D Kalbfleisch, Arthur V Peterson Jr, Nancy Flournoy, Vernon T Farewell, and Norman E Breslow. The analysis of failure times in the presence of competing risks. *Biometrics*, pages 541–554, 1978.

William H Press, Saul A Teukolsky, William T Vetterling, and Brian P Flannery. *Numerical recipes 3rd edition: The art of scientific computing*. Cambridge university press, 2007.

Md Mahmudur Rahman, Koji Matsuo, Shinya Matsuzaki, and Sanjay Purushotham. Deeppseudo: pseudo value based deep learning models for competing risk analysis. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 479–487, 2021.

David Rindt, Robert Hu, David Steinsaltz, and Dino Sejdinovic. Survival regression with proper scoring rules and monotonic neural networks. In *International Conference on Artificial Intelligence and Statistics*, pages 1190–1205. PMLR, 2022.

Patrick Royston. Flexible parametric alternatives to the cox model, and more. *The Stata Journal*, 1(1):1–28, 2001.

Matthias Schmid and Moritz Berger. Competing risks analysis for discrete time-to-event data. *Wiley Interdisciplinary Reviews: Computational Statistics*, 13(5):e1529, 2021.

Noah A Schuster, Emiel O Hoogendijk, Almar AL Kok, Jos WR Twisk, and Martijn W Heymans. Ignoring competing events in the analysis of survival data may lead to biased results: a nonmathematical illustration of competing risk analysis. *Journal of clinical epidemiology*, 122:42–48, 2020.

Steve Selvin. *Survival analysis for epidemiologic and medical research*. Cambridge University Press, 2008.

Oleksandr Shchur, Marin Biloš, and Stephan Günnemann. Intensity-free learning of temporal point processes. In *International Conference on Learning Representations*, 2020.

Mats J Stensrud and Miguel A Hernán. Why test for proportional hazards? *Jama*, 323(14): 1401–1402, 2020.

Terry M Therneau, Patricia M Grambsch, Terry M Therneau, and Patricia M Grambsch. *The cox model*. Springer, 2000.

Donna Tjandra, Yifei He, and Jenna Wiens. A hierarchical approach to multi-event survival analysis. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 591–599, 2021.

Stéphanie Van Der Pas, Rob Nelissen, and Marta Fiocco. Different competing risks models for different questions may give similar results in arthroplasty registers in the presence of few events: illustrated with 138,234 hip (124,560 patients) and 139,070 knee (125,213 patients) replacements from the dutch arthroplasty register. *Acta Orthopaedica*, 89(2):145–151, 2018.

Bob JH van Kempen, Bart S Ferket, Maryam Kavousi, Maarten JG Leening, Ewout W Steyerberg, M Arfan Ikram, Jacqueline CM Witteman, Albert Hofman, Oscar H Franco, and MG Myriam Hunink. Performance of framingham cardiovascular disease (cvd) predictions in the rotterdam study taking into account competing risks and disentangling cvd into coronary heart disease (chd) and stroke. *International journal of cardiology*, 171(3): 413–418, 2014.

Hilary K Wall, Matthew D Ritchey, Cathleen Gillespie, John D Omura, Ahmed Jamal, and Mary G George. Vital signs: prevalence of key cardiovascular disease risk factors for million hearts 2022—united states, 2011–2016. *Morbidity and Mortality Weekly Report*, 67(35): 983, 2018.

Ping Wang, Yan Li, and Chandan K Reddy. Machine learning for survival analysis: A survey. *ACM Computing Surveys (CSUR)*, 51(6):1–36, 2019.

Peter WF Wilson, Ralph B D'Agostino, Daniel Levy, Albert M Belanger, Halit Silbershatz, and William B Kannel. Prediction of coronary heart disease using risk factor categories. *Circulation*, 97(18):1837–1847, 1998.

# Appendix A. Experiments

## A.1. Datasets characteristics

Table 6 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 6: 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:

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

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

## A.3. Time specific results

Tables 7, 8 and 9 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 10 displays the time-dependent C-index and cumulative time-dependent Brier score. These results echo the findings from Table 2.

## A.5. Implementation details

The proposed experiments rely on the `scikit-survival` (Pölsterl, 2020)[7] and `pycox`[8] libraries for evaluation. For baselines' implementations, we used the R library `riskRegression`[9] for CS Cox and Fine-Gray, `pycox` for DeepHit and `auton-survival` (Nagpal et al., 2022b)[10] 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 –

7. https://github.com/sebp/scikit-survival
8. https://github.com/havakv/pycox
9. https://github.com/tagteam/riskRegression
10. 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 7: 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.*

| | 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** | 0.822 (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 8: 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 | | | Competing risk | | |
|---|---|---|---|---|---|---|---|
| | | $q_{0.25}$ | $q_{0.50}$ | $q_{0.75}$ | $q_{0.25}$ | $q_{0.50}$ | $q_{0.75}$ |
| 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 9: 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 10: 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)* | | | Brier Score *(Smaller is better)* | | |
|---|---|---|---|---|---|---|---|---|
| | | | $q_{0.25}$ | $q_{0.50}$ | $q_{0.75}$ | $q_{0.25}$ | $q_{0.50}$ | $q_{0.75}$ |
| 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. | Dea. 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) |
| | | **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 11: Comparison of model performance by means (standard deviations) across 5-fold cross-validation. Best performances are in **bold**.

instead of one network with multiple outcomes. The model **MonoFG** consists of the same architecture presented in Figure 1 with only one monotonic network with $R$ outputs. Table 11 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 12. 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.

### 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 13 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 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

| | Risk | Model | C-Index *(Larger is better)* | | | Brier Score *(Smaller is better)* | | |
|---|---|---|---|---|---|---|---|---|
| | | | $q_{0.25}$ | $q_{0.50}$ | $q_{0.75}$ | $q_{0.25}$ | $q_{0.50}$ | $q_{0.75}$ |
| 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 12: Impact of increasing $n$ on DeSurv performances. Performance measured by means (standard deviations) across 5-fold cross-validation. Best performances are in **bold**.

| | 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 13: Training speed comparison on the Framingham dataset. Performance measured by means (standard deviations) across 100-fold *Monte Carlo* cross-validation.

number of iterations required by our proposed methods in comparison to DeSurv reflects the more complex convergence of *constrained* monotonic neural networks.