A Federated Cox Model with Non-Proportional Hazards

D. Kai Zhang, Francesca Toni, and Matthew Williams

Imperial College London
{dz819,f.toni,matthew.williams}@imperial.ac.uk

Abstract. Recent research has shown the potential for neural networks to improve upon classical survival models such as the Cox model, which is widely used in clinical practice. Neural networks, however, typically rely on data that are centrally available, whereas healthcare data are frequently held in secure silos. We present a federated Cox model that accommodates this data setting and also relaxes the proportional hazards assumption, allowing time-varying covariate effects. In this latter respect, our model does not require explicit specification of the time-varying effects, reducing upfront organisational costs compared to previous works. We experiment with publicly available clinical datasets and demonstrate that the federated model is able to perform as well as a standard model.

1 Introduction

Estimating how long patients might live for is a key task in clinical medicine, and is a common question from patients. Survival analysis is the statistical branch used to perform these estimates, which can range in its application from predicting death following diagnosis to loan defaults or machine part failures. Amongst survival models, the Cox model [6] is one of the most widely used.

Machine learning techniques have received attention for their potential to improve upon the performance of the Cox model. Many recent efforts [13,20,16] have exploited neural networks (NNs) to model more complex relationships as well as enable typically unsupported input data types such as images [27,28,17,4]. Notwithstanding, the Cox model has remained the standard in survival analysis [24]. Indeed, the adoption of machine learning has progressed haltingly in many areas of healthcare [14].

One challenge lies in the distributed nature of healthcare data [25]. In much of machine learning, data are centralised, whereas privacy concerns often result in secure data "silos" in healthcare. Federated learning (FL) accommodates this decentralised data environment and has shown promise in clinical contexts [23].

Despite a fast emerging literature in FL, there has been scant work on federated survival analysis. [1] propose a federated Cox model that is closest to this paper. The standard Cox model is, however, limited in that it can only correctly model proportional hazards. We take an alternative approach allowing us to embed time-varying covariate effects (*non*-proportional hazards) directly

in the architecture, potentially reducing organisational setup costs for federations. Such effects are relevant to adapt models for patients such as those with breast cancer where the proportional hazards assumption has been shown to be violated [3,5,11].

In the following, we briefly discuss relevant background and highlight related work (Section 2), before defining our model (Section 3), instantiating it with different hazards assumptions and presenting our experiments with real-world clinical datasets (Section 4). Section 5 concludes with potential directions for future work.

2 Background and Related Work

The promise of greater control over data ownership and enhanced privacy that FL affords has generated interest in the healthcare community. Few works, however, have investigated the intersection between survival analysis and FL. We present background on each of these areas separately before discussing their intersection relevant for this work.

Background on Federated Learning (FL). FL is a framework for decentralised data that cannot be shared due to their sensitive content or prohibitive communication costs [21]. In the context of healthcare, patient data may be kept in this way by the clinical unit (e.g., the hospital) at which the patient was treated. In the following, we will simply refer to these data-keeping units as federation members or centres. Typically (and in this paper), the federated objective is to minimise $\mathcal{L}_F(X, \phi)$ with respect to ϕ with:

$$\mathcal{L}_F(X, \phi) = \sum_{k \in K} w_k \mathcal{L}_k(X_k, \phi) \tag{1}$$

where $\mathcal{L}_F$ represents the global loss: an average of the local losses $\mathcal{L}_k$ computed by the federation members in K on their own data X_k weighted by w_k, where ϕ represents the model parameters. Typically, each member customises ϕ for a number of local optimisation rounds before aggregating the customised ϕ for a new global consensus model.

Background on Survival Analysis. Survival analysis estimates the time to an event for a population N with data $\mathcal{D} = \{(x_i, t_i, s_i)\}_{i \in N}$ where each person i has covariates $x_i = (x_{i1}, ..., x_{ip})^\top$, a time of observation t_i and an indicator $s_i \in \{0, 1\}$ which equals 1 if i has experienced the event or 0 if not, i.e., if i is *censored*.

The *Cox model* [6] is one of the most widely used survival models. It defines a hazard function h, which expresses the rate of failure at time t subject to survival until then as follows:

$$\begin{aligned} h(t|x_i) &= P(T = t|T > t-1) \\ &= h_0(t)exp[g(x_i)] \text{ with } g(x_i) = \beta^\top x_i \end{aligned} \tag{2}$$

where $h_0(t)$ is some baseline hazard and where $\beta = (\beta_1, ..., \beta_p)^\top$ is a coefficient vector. Later works replace the linear predictor $\beta^\top x_i$ with NNs $g_\phi(x_i)$, demonstrating competitive performance [9,13]. The coefficients are estimated by minimising the negative partial log-likelihood given by:

$$-\sum_{i \in N} s_i[g(x_i) - \log(\sum_{j \in R_i} exp[g(x_j)])] \tag{3}$$

where $R_i = \{j \in N : t_j \geq t_i\}$ denotes the individuals who are still *at risk* when i experiences the event.

In a federated setting, this loss generally cannot be decomposed into local losses due to the logarithmic term, as the risk set R_i can contain individuals from centres other than the one of i. This therefore does not match the formulation of Eq. 1. The hazard function also assumes proportional hazards (PH) – differences in covariates result in constant proportional differences in hazards. Over long time horizons, this can be restrictive [16,3,2].

State-of-the-Art in Federated Survival Analysis. Our work is situated in the intersection of federated learning and survival analysis and proposes a novel framework. A handful of works have already proposed such frameworks of which we provide a brief overview here. The works of [19] and [8] embody one approach which relies on substantial sharing of summary statistics over the local datasets in every training iteration. This differs in spirit from FL where more abstract parameters are shared and, often, infrequently so. Moreover, their models are based on linear predictors and do not address integration with NNs.

Recent work by [1] is closest to our approach. Their model exploits a discretisation of the Cox model (also by [6]) with an NN-based predictor:

$$\frac{h(t|x)}{1 - h(t|x)} = \frac{h_0(t)}{1 - h_0(t)} exp[g_\phi(x_i)] \tag{4}$$

which can be rewritten in a sigmoid form:

$$h(t|x_i) = \frac{1}{1 + exp[-(\alpha_t + g_\phi(x_i))]} \tag{5}$$

where $\alpha_t = \log(\frac{h_0(t)}{1-h_0(t)})$.

They follow [7] in estimating this function like a logistic regression with negative log-likelihood:

$$-\sum_{i \in N} \sum_{k=1}^{t_i} [y_{ik} \log[h(k|x_i)] + (1 - y_{ik}) \log[(1 - h(k|x_i))]] \tag{6}$$

where $y_{ij} = \mathbb{1}\{t_j = t_i, s_i = 1\}$. Importantly, this loss does not depend on risk sets and is therefore separable – each centre's loss only depends on local data – recovering the federated objective (Eq. 1).

[1] demonstrate that a federation of this model can draw even in performance with a model trained on pooled data. This is, however, only shown with aggregation after every local optimisation round – a setup that may need to differ in practice [23] – and assuming PH, as their predictor $g_\phi(x_i)$ is time-invariant.

We note that non-PH can be admitted to their model by including time interactions (giving $g_\phi(x_i, f(t))$) – an approach sometimes taken in standard Cox models – as demonstrated on pooled data by [7]. This could, however, introduce a dependency on the specification of $f(t)$ and its interactions. Crucially, this may add to the organisational setup costs of a federation: even though interactions could be learned, $f(t)$ needs to be fully specified and agreed upon in advance. In contrast, we follow [10] by making the choice between PH and non-PH a binary decision over the architecture of the output layer.

3 Model

We build upon the discretised Cox model and detail how the PH assumption is relaxed and formulate the federated objective. We describe a discretisation procedure and an optional interpolation scheme for smooth predictions. Lastly, we outline two complementary performance metrics.

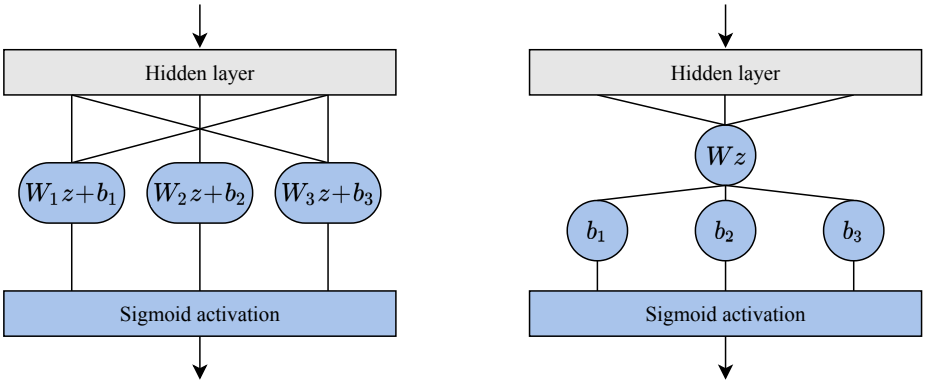
Non-Proportional Hazards. We use a discretised Cox model (Eq. 5) but parameterised with a time-varying, NN-based predictor $g_{\phi,t}(x)$. Following [10], we allow for non-PH by fully connecting the output layer to the previous layer. The output layer thus encodes time-varying covariate effects in time-specific weights. A sigmoid is used to retrieve the hazard rates.

For PH, the output component is split into a first layer with a single neuron and no bias. The output of this neuron is passed into a second layer with as many neurons as time steps. This captures time-varying baseline hazards in the second layer and time-invariant covariate effects in the first. The difference in components is illustrated in Figure 1.

Federated Objective. To conform to a federated formulation (Eq. 1), we split the objective (Eq. 6):

$$\begin{aligned} \mathcal{L}_F &= \sum_{k \in K} w_k \mathcal{L}_k(X_k, \phi) \\ &= -\sum_{k \in K} \frac{|N_k|}{|N|} \sum_{i \in N_k} \sum_{j=1}^{t_i} \Big[y_{ij} \log[h(j|x_i)] + (1 - y_{ij}) \log[(1 - h(j|x_i))] \Big] \end{aligned} \tag{7}$$

where $y_{ij} = \mathbb{1}\{t_j = t_i, s_i = 1\}$. Each centre calculates $\mathcal{L}_k(X_k, \phi)$ on its own subset of the population N_k. We adapt the FedAvg algorithm of [21] to minimise this loss (Algorithm 1).



(a) Non-PH (b) PH

Fig. 1: Output components for 3 time steps.

Algorithm 1 Procedure to optimise the federated objective.

Initialise global model with ϕ_0
for each round $t = 1, ..., T$ **do**
  for each centre $k = 1, ..., K$ in parallel **do**
    Send ϕ_{t-1} to centre k
    for each local round $b = 1, ..., B$ **do**
      Local update $\phi_t^k \leftarrow \phi_t^k - \lambda \nabla \mathcal{L}_k(X_k, \phi_t^k)$
    end for
    Receive ϕ_t^k from centre k
  end for
  Aggregate $\phi_t \leftarrow \sum_{k \in K} \frac{|N_k|}{|N|} \phi_t^k$
end for
return ϕ_T

Discretisation. The model operates on discretised time, so that t indexes into a set of intervals $[\tau_{t-1}, \tau_t)$. Following [15] we discretise on Kaplan-Meier quantiles. Defining the survival curve $S(\tau) = S(\tau - 1)(1 - h(\tau))$, the quantiles $\{\tau_1, \tau_2, ..., \tau_m\}$ can be obtained as:

$$S(T = \tau_j) - S(T = \tau_{j+1}) = \frac{1 - S(T = \tau_{max})}{m} \tag{8}$$

for $j = \{0, 1, ..., m - 1\}$. This discretisation procedure yields a set of steps $\{\tau_1, \tau_2, ..., \tau_m\}$ where each step results in the same decrease in survival (an illustration is provided in Appendix A).

Interpolation. To smooth step-wise predictions, we use constant density interpolation [15]. Letting $\widetilde{S}(\tau)$ denote the interpolation of the survival curve $S(\tau)$, we then have:

$$\widetilde{S}(\tau) = S(\tau_{j-1}) + [S(\tau_j) - S(\tau_{j-1})] \frac{\tau - \tau_{j-1}}{\tau_j - \tau_{j-1}} \tag{9}$$

for a given time $\tau \in (\tau_{j-1}, \tau_j]$. Intuitively, the step survival curve is linearly interpolated between any adjacent steps, resulting in constant densities in the corresponding interval (an illustration is provided in Appendix A).

Performance Metrics – Concordance. We use the time-dependent concordance index [2], or c-index, which is a discriminative measure for how well the model ranks the relative survival between patient pairs, expressed as:

$$P(S(t_i|x_i) < S(t_i|x_j) \;\&\; t_i < t_j \;\&\; s_i = 1) \tag{10}$$

which is estimated as follows:

$$\widehat{c} = \frac{\sum_{i \in N} \sum_{j \in N, j \neq i} conc_{ij}}{\sum_{i \in N} \sum_{j \in N, j \neq i} comp_{ij}} \tag{11}$$

$$comp_{ij} = \mathbb{1}\{t_i < t_j \;\&\; s_i = 1\} + \mathbb{1}\{t_i = t_j \;\&\; s_i = 1 \;\&\; s_j = 0\} \tag{12}$$

$$conc_{ij} = \mathbb{1}\{S(t_i|x_i) < S(t_i|x_j)\}\, comp_{ij} \tag{13}$$

Performance Metrics – Calibration. While the c-index measures the discriminative performance of the model, it does not measure how well *calibrated* these estimates are (an illustration is provided in Appendix A).

As a measure of calibration, we follow [12] who propose a Brier score for use with censored data defined as follows:

$$BS(t) = \frac{1}{|N|} \sum_{i \in N} w_i(t) \Big(y_i(t) - h(t|x_i) \Big)^2 \tag{14}$$

$$w_i(t) = \begin{cases} s_i / G(t), & \text{if } t_i \leq t \\ 1 / G(t), & \text{if } t_i > t \end{cases} \tag{15}$$

where $y_i(t) = \mathbb{1}\{t_i = t\}$ and $G(t)$ is the Kaplan-Meier estimate of the censoring distribution (i.e., estimated on $\{(x_i, t_i, 1 - s_i)\}_{i \in N}$). To measure calibration across the entire time horizon, we numerically integrate the Brier score using 100 time points [16].

4 Experiments

We introduce the datasets we experiment on, describe the setup of a simulated federation and instantiate our model with three different linearity and hazards assumptions, and present our results.[1]

Dataset	Size	Features	Prop. censored	Last event
METABRIC	1,904	9	42%	355 days
SUPPORT	8,873	14	32%	1,944 days
GBSG	2,232	7	43%	83 days

Table 1: Overview of datasets.

4.1 Datasets

We experiment on three clinical datasets (Table 1; for Kaplan-Meier curves see Appendix A) made available by [13], namely the Molecular Taxonomy of Breast Cancer International Consortium (METABRIC), the Study to Understand Prognoses Preferences Outcomes and Risks of Treatment (SUPPORT), and the Rotterdam tumour bank and German Breast Cancer Study Group (GBSG). METABRIC and GBSG both relate to breast cancer patients, a group for whom non-PH have been noted [3,5], while SUPPORT presents serious hospitalisations for a second application area.

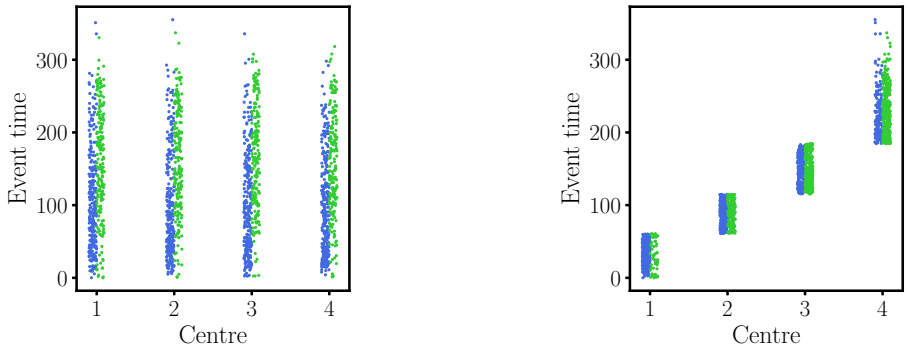
4.2 Setup

We simulate two federated data cases: In the first, data are randomly distributed ("IID"), simulating the case of each centre seeing a similar sample of the patient population. In the second, data are stratified on the time to event ("Non-IID"), simulating the case that each centre sees a non-overlapping quantile of the population – from centre 1 seeing only the shortest survivals leading to centre 4 with the longest survivals (Figure 2). For comparability, we maintain the total number of local training rounds at 100. A pooled data baseline is provided ("Pooled" – no distinction between local and global rounds). In all cases, 80% of the overall data are split, if federated, and used for training, while 20% are held out for evaluation.

We instantiate the model with different choices for $g(x)$ – with a linear predictor or with an NN, with and without PH (Table 2). For baselines, we considered the works of [1] and [7]. The former assumes PH, however, while the latter is a pooled data model. Both require upfront agreement on a specification of $f(t)$ to include non-PH, adding to the setup costs of a federation. We further note that no implementations of these models are available. We therefore provide the NN PH model to approximate the model of [1] – a federated NN-based Cox model with PH – and the Linear PH model as a standard baseline.

Architectures (Figure 1) are implemented in PyTorch 1.8.0 [22] with two hidden layers of 32 neurons for NN models and none for the linear model. Optimisation uses Adam with grid-sought learning rates (10^{-1} to 10^{-5} on 20% of the training data) and a batch size of 256. Base case discretisation uses 10 time steps.

[1] For source code, see https://github.com/dkaizhang/federated-survival.



(a) IID (b) Non-IID

Fig. 2: Event time distribution for IID and non-IID data using stratification by event time on METABRIC.

Model	Predictor
Linear PH	$\beta^\top x$
NN PH	$g_\phi(x)$
NN nonPH	$g_{\phi,t}(x)$

Table 2: Model choices.

4.3 Results

In this section, we first compare the performance of the three models trained in a centralised fashion on pooled data against their federated performance on decentralised data. We next provide additional experiments exploring the impact of the discretisation grid chosen for the base case. We report averaged 5-fold cross-validation performance throughout.

Federated Performance. On pooled data, the NN nonPH model outperforms or ties in concordance (Table 3) and ties with the best in calibration (Table 4), indicating a gain from the relaxation of the PH assumption. For METABRIC and GBSG this aligns with [3,5,11] who find non-PH amongst this patient group.

Comparing this to the federated setting with IID data, all three models maintain their performance (within one standard deviation) in concordance and calibration when aggregation is frequent. First hints of performance degradation amongst the NN-based models occur as aggregation becomes very infrequent (rightmost columns) while the Linear PH model appears largely unaffected. This observation is noteworthy, as infrequent aggregation will be a likely feature in practice given communication costs. This indicates a potential trade-off between model complexity and achievable aggregation frequency to support its training.

In practice, data are likely to be non-IID across centres. The results show that the performances of all three models suffer when this is the case. Generally, the NN-based models experience the most severe losses in performance and are largely outperformed by the Linear PH model. When aggregation is infrequent, the NN-based models on SUPPORT and GBSG effectively approach a no-skill predictor in concordance (average c-index of 0.5). Further, performance losses under infrequent aggregation of the NN-based models are, as would be expected, worse than under IID data. On SUPPORT, the NN-based models exhibit much greater performance differences than on the other datasets. In this respect, we note that SUPPORT has much longer survival times than METABRIC or GBSG (Table 1), so that stratification by event time likely results in a more significantly different partition of the data for the former than for the latter two.

Impact of Discretisation Fineness. We re-train models on finer time grids using 100 global and 1 local rounds. A finer grid on METABRIC (Figure 3 upper panel) and GBSG (Figure 3 lower panel) did not improve performance and, in fact, appears to degrade performance. Notably, the Linear PH model becomes a no-skill predictor in terms of concordance on the non-IID GBSG case. The results are less conclusive on SUPPORT (Figure 3 middle panel), as a finer time grid appears to result in a minor to no increase in concordance at the expense of a loss in calibration.

A finer time grid can be expected to result in a trade-off between closer approximation of true (smooth) survival and a reduction in data available in any given time step. An increase from 10 to 20 time steps, for instance, halves the number of available data points to estimate a given step. The latter effect appears to dominate on the smaller METABRIC and GBSG datasets, and less so for the approximately 4-times larger SUPPORT dataset.

5 Conclusion

We present a federated Cox model that relaxes the proportional hazards (PH) assumption and demonstrate its ability to maintain concordance and calibration relative to a pooled baseline under various linearity and PH assumptions. Compared to prior work, this federation scheme encodes the decision between PH and non-PH in a binary choice over the output layer, rather than requiring upfront agreement on a specification of $f(t)$. We note that our model is not restricted to a particular data type or network architecture excepting the output component. Future work could adapt the model for image-based federated survival predictions.

The decrease in performance on non-IID data (even if pathologically derived in this paper) represents a challenge to the application of federated learning in practice. Extensions could include exploring methods accounting for statistical heterogeneity [26,18] or other federation topologies which maintain locally specialised models trained in a peer-to-peer fashion [23]. While the heterogeneity in this paper was derived from label stratification, other types of heterogeneity,

		METABRIC			SUPPORT			GBSG		
		Global / local rounds			Global / local rounds			Global / local rounds		
Data	Model	100 / 1	20 / 5	1 / 100	100 / 1	20 / 5	1 / 100	100 / 1	20 / 5	1 / 100
Pooled	Linear PH	63.5 ± 1.4			57.2 ± 1.0			66.5 ± 2.1		
	NN PH	64.0 ± 0.6			60.8 ± 0.6			66.2 ± 2.6		
	NN nonPH	$\mathbf{66.7 \pm 2.1}$			$\mathbf{61.5 \pm 1.2}$			$\mathbf{66.6 \pm 1.9}$		
IID	Linear PH	63.9 ± 0.8	64.0 ± 2.2	$\mathbf{63.7 \pm 1.7}$	57.2 ± 0.8	57.2 ± 0.8	57.2 ± 0.4	66.5 ± 1.5	66.3 ± 0.6	$\mathbf{66.3 \pm 1.4}$
	NN PH	63.8 ± 1.6	62.5 ± 1.9	63.3 ± 1.2	60.6 ± 1.0	60.9 ± 0.9	$\mathbf{58.3 \pm 1.4}$	$\mathbf{67.4 \pm 1.7}$	$\mathbf{67.1 \pm 1.2}$	63.5 ± 2.6
	NN nonPH	$\mathbf{65.4 \pm 1.9}$	$\mathbf{65.7 \pm 1.4}$	61.5 ± 2.1	$\mathbf{62.1 \pm 0.7}$	$\mathbf{62.4 \pm 0.4}$	56.7 ± 2.3	66.5 ± 1.0	66.4 ± 0.9	62.8 ± 1.9
Non-IID	Linear PH	59.2 ± 3.0	$\mathbf{59.8 \pm 2.1}$	$\mathbf{61.0 \pm 1.3}$	55.4 ± 1.5	56.1 ± 0.9	$\mathbf{56.2 \pm 0.5}$	61.1 ± 0.9	$\mathbf{61.9 \pm 3.0}$	$\mathbf{56.5 \pm 6.9}$
	NN PH	$\mathbf{60.9 \pm 1.1}$	59.4 ± 2.5	57.3 ± 4.2	$\mathbf{57.0 \pm 1.4}$	$\mathbf{56.7 \pm 1.1}$	52.9 ± 0.8	$\mathbf{61.3 \pm 4.3}$	61.8 ± 2.9	53.0 ± 6.2
	NN nonPH	57.9 ± 2.9	59.6 ± 3.9	54.6 ± 5.3	50.8 ± 0.6	50.9 ± 0.9	50.8 ± 1.2	57.6 ± 1.9	55.8 ± 2.0	52.0 ± 2.5

Table 3: C-index (rebased to 100) – mean and standard deviation. Higher values are better.

		METABRIC			SUPPORT			GBSG		
		Global / local rounds			Global / local rounds			Global / local rounds		
Data	Model	100 / 1	20 / 5	1 / 100	100 / 1	20 / 5	1 / 100	100 / 1	20 / 5	1 / 100
Pooled	Linear PH	$\mathbf{16.4 \pm 0.6}$			20.9 ± 0.5			$\mathbf{18.0 \pm 0.5}$		
	NN PH	16.8 ± 0.7			$\mathbf{19.6 \pm 0.4}$			18.2 ± 0.8		
	NN nonPH	$\mathbf{16.4 \pm 0.8}$			$\mathbf{19.6 \pm 0.4}$			$\mathbf{18.0 \pm 0.4}$		
IID	Linear PH	$\mathbf{16.3 \pm 1.1}$	16.7 ± 0.9	$\mathbf{16.5 \pm 0.7}$	20.9 ± 0.6	20.9 ± 0.3	$\mathbf{20.9 \pm 0.6}$	18.1 ± 0.2	18.1 ± 0.2	$\mathbf{18.1 \pm 0.3}$
	NN PH	17.2 ± 1.1	18.1 ± 0.5	18.1 ± 1.1	19.7 ± 0.7	19.8 ± 0.2	22.6 ± 1.3	$\mathbf{17.7 \pm 0.3}$	$\mathbf{17.8 \pm 0.4}$	19.8 ± 1.9
	NN nonPH	$\mathbf{16.3 \pm 1.3}$	$\mathbf{16.5 \pm 0.6}$	19.1 ± 0.8	$\mathbf{19.5 \pm 0.7}$	$\mathbf{19.4 \pm 0.4}$	21.1 ± 0.6	18.3 ± 0.5	18.2 ± 0.6	21.7 ± 1.5
Non-IID	Linear PH	$\mathbf{18.3 \pm 0.8}$	$\mathbf{18.0 \pm 0.6}$	$\mathbf{19.3 \pm 0.7}$	22.9 ± 0.7	22.1 ± 0.3	$\mathbf{21.4 \pm 0.3}$	$\mathbf{20.4 \pm 0.5}$	$\mathbf{20.1 \pm 0.5}$	$\mathbf{22.1 \pm 0.3}$
	NN PH	$\mathbf{18.3 \pm 0.5}$	18.1 ± 0.2	19.7 ± 1.0	$\mathbf{22.6 \pm 0.5}$	$\mathbf{21.9 \pm 0.2}$	24.0 ± 1.8	20.5 ± 0.6	20.6 ± 0.4	22.3 ± 0.5
	NN nonPH	20.9 ± 0.3	20.1 ± 0.5	21.1 ± 1.0	26.2 ± 0.4	25.3 ± 0.5	25.3 ± 0.7	23.6 ± 0.9	22.4 ± 0.8	22.7 ± 1.0

Table 4: Integrated Brier scores (rebased to 100) – mean and standard deviation. Lower values are better.

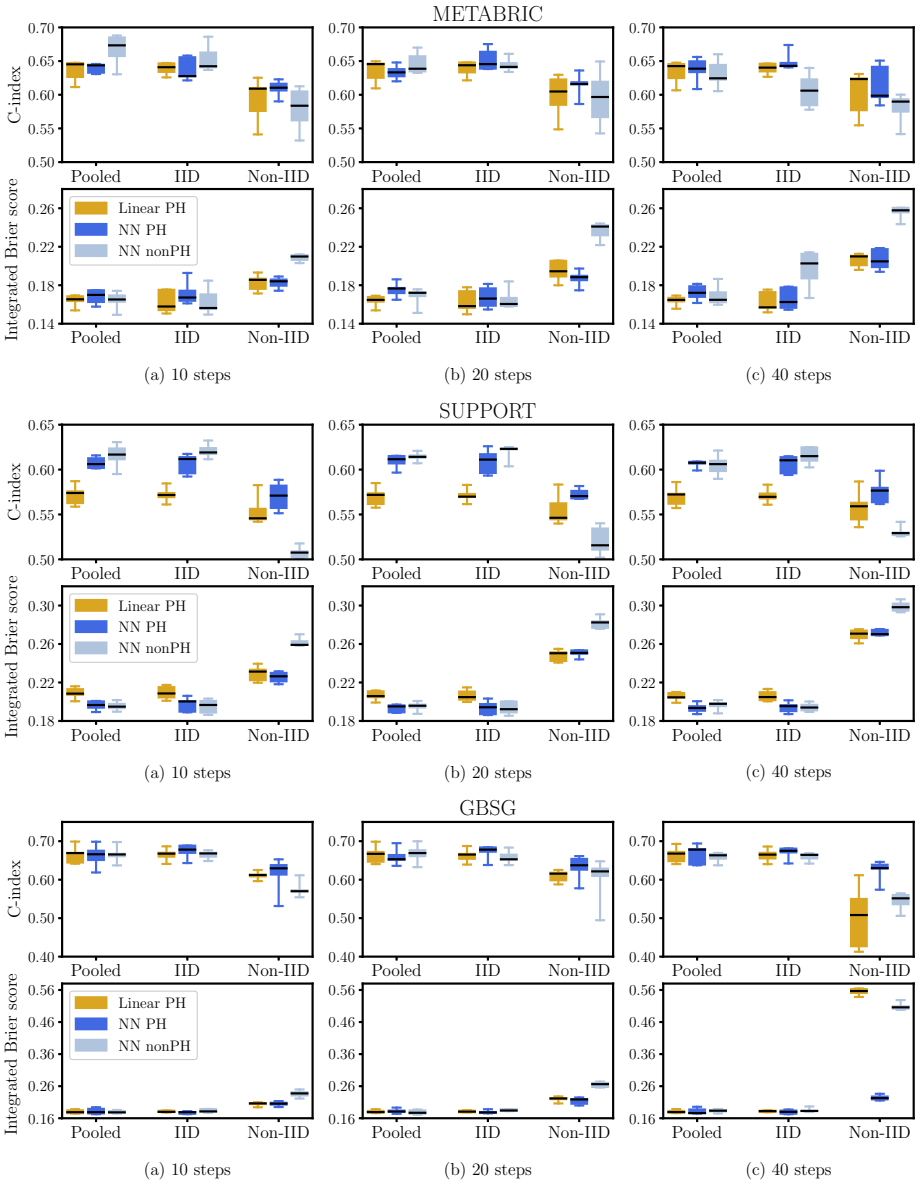
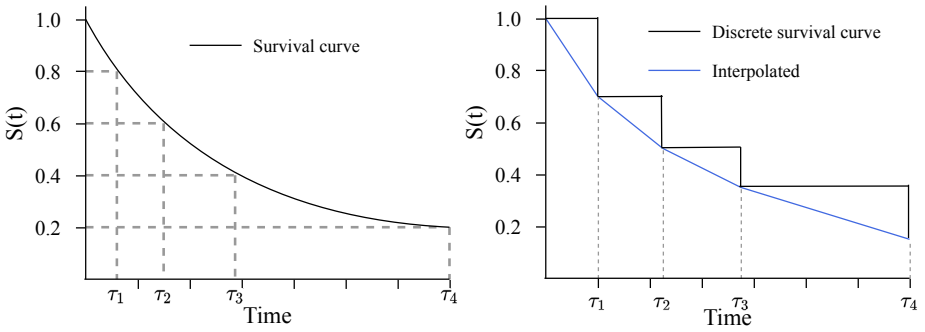


Fig. 3: Model performance with increasing discretisation fineness. Federated models were trained with 100 global and 1 local rounds. Performance decreases on smaller METABRIC and GBSG datasets with mixed results on the larger SUPPORT dataset.

such as covariate shifts, could be explored: for image-based survival predictions, differences in acquisition protocols could provide one such avenue.

Acknowledgements. This work was supported by the UKRI CDT in AI for Healthcare http://ai4health.io (Grant No. EP/S023283/1)

A Additional Figures



(a) Kaplan-Meier-based discretisation. (b) Constant density interpolation.

Fig. 4: Discretisation and interpolation.

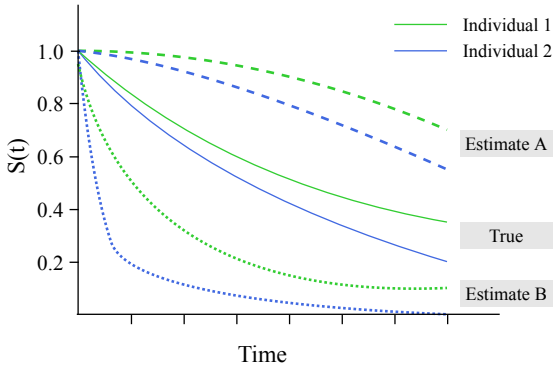


Fig. 5: Two sets of survival estimates with correct ranking (green above blue) but poor calibration given under- / overestimation of true survival curves.

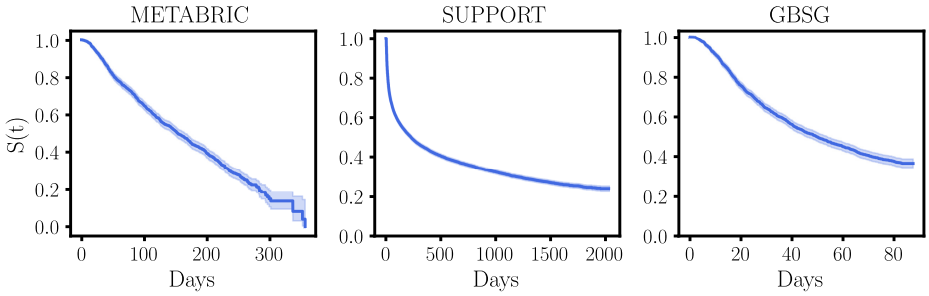


Fig. 6: Kaplan-Meier estimates with 95% confidence interval.

References

1. Andreux, M., Manoel, A., Menuet, R., Saillard, C., Simpson, C.: Federated Survival Analysis with Discrete-Time Cox Models. arXiv pp. 1–21 (2020)
2. Antolini, L., Boracchi, P., Biganzoli, E.: A time-dependent discrimination index for survival data. Statistics in Medicine **24**(24), 3927–3944 (2005). https://doi.org/10.1002/sim.2427
3. Bellera, C.A., MacGrogan, G., Debled, M., De Lara, C.T., Brouste, V., Mathoulin-Pélissier, S.: Variables with time-varying effects and the Cox model: Some statistical concepts illustrated with a prognostic factor study in breast cancer. BMC Medical Research Methodology **10** (2010). https://doi.org/10.1186/1471-2288-10-20
4. Bello, G.A., Dawes, T.J., Duan, J., Biffi, C., de Marvao, A., Howard, L.S., Gibbs, J.S.R., Wilkins, M.R., Cook, S.A., Rueckert, D., O'Regan, D.P.: Deep-learning cardiac motion analysis for human survival prediction. Nature Machine Intelligence **1**(2), 95–104 (2019). https://doi.org/10.1038/s42256-019-0019-2
5. Coradini, D., Daidone, M.G., Boracchi, P., Biganzoli, E., Oriana, S., Bresciani, G., Pellizzaro, C., Tomasic, G., Di Fronzo, G., Marubini, E.: Time-dependent relevance of steroid receptors in breast cancer. Journal of Clinical Oncology **18**(14), 2702–2709 (2000). https://doi.org/10.1200/JCO.2000.18.14.2702
6. Cox, D.R.: Regression Models and Life-Tables. Journal of the Royal Statistical Society. Series B (Methodological) **34**(2), 187–202 (1972). https://doi.org/10.1111/j.2517-6161.1972.tb00899.x, https://www.jstor.org/stable/2985181
7. Craig, E., Zhong, C., Tibshirani, R.: Survival stacking: casting survival analysis as a classification problem. arXiv pp. 1–17 (2021), http://arxiv.org/abs/2107.13480
8. Dai, W., Jiang, X., Bonomi, L., Li, Y., Xiong, H., Ohno-Machado, L.: VERTICOX: Vertically Distributed Cox Proportional Hazards Model Using the Alternating Direction Method of Multipliers. IEEE Transactions on Knowledge and Data Engineering **4347**(c), 1–1 (2020). https://doi.org/10.1109/tkde.2020.2989301
9. Faraggi, D., Simon, R.: A neural network model for survival data. Statistics in Medicine **14**(1), 73–82 (1995). https://doi.org/10.1002/sim.4780140108
10. Gensheimer, M.F., Narasimhan, B.: A scalable discrete-time survival model for neural networks. PeerJ pp. 1–19 (2019). https://doi.org/10.7717/peerj.6257
11. Gore, S.M., Pocock, S.J., Kerr, G.R.: Regression Models and Non-Proportional Hazards in the Analysis of Breast Cancer Survival Author. Journal of the Royal Statistical Society. Series C (Applied Statistics) **33**(2), 176–195 (1984)

12. Graf, E., Schmoor, C., Sauerbrei, W., Schumacher, M.: Assessment and comparison of prognostic classification schemes for survival data. Statistics in Medicine **18**(17-18), 2529–2545 (1999). https://doi.org/10.1002/(sici)1097-0258(19990915/30)18:17/18¡2529::aid-sim274¿3.0.co;2-5
13. Katzman, J.L., Shaham, U., Cloninger, A., Bates, J., Jiang, T., Kluger, Y.: DeepSurv: Personalized treatment recommender system using a Cox proportional hazards deep neural network. BMC Medical Research Methodology **18**(1), 1–11 (2018). https://doi.org/10.1186/s12874-018-0482-1
14. Kelly, C.J., Karthikesalingam, A., Suleyman, M., Corrado, G., King, D.: Key challenges for delivering clinical impact with artificial intelligence. BMC Medicine **17**(1), 1–9 (2019). https://doi.org/10.1186/s12916-019-1426-2
15. Kvamme, H., Borgan, O.: Continuous and discrete-time survival prediction with neural networks. arXiv (2019)
16. Kvamme, H., Borgan, O., Scheel, I.: Time-to-event prediction with neural networks and cox regression. Journal of Machine Learning Research **20**, 1–30 (2019)
17. Li, H., Boimel, P., Janopaul-Naylor, J., Zhong, H., Xiao, Y., Ben-Josef, E., Fan, Y.: Deep convolutional neural networks for imaging data based survival analysis of rectal cancer. arXiv pp. 1–4 (2019)
18. Li, T., Sahu, A.K., Zaheer, M., Sanjabi, M., Talwalkar, A., Smith, V.: Federated Optimization in Heterogeneous Networks (2018), http://arxiv.org/abs/1812.06127
19. Lu, C.L., Wang, S., Ji, Z., Wu, Y., Xiong, L., Jiang, X., Ohno-Machado, L.: WebDISCO: A web service for distributed cox model learning without patient-level data sharing. Journal of the American Medical Informatics Association **22**(6), 1212–1219 (2015). https://doi.org/10.1093/jamia/ocv083
20. Luck, M., Sylvain, T., Cardinal, H., Lodi, A., Bengio, Y.: Deep Learning for Patient-Specific Kidney Graft Survival Analysis (Nips 2017) (2017), http://arxiv.org/abs/1705.10245
21. McMahan, H.B., Moore, E., Ramage, D., Hampson, S., Agüera y Arcas, B.: Communication-efficient learning of deep networks from decentralized data. Proceedings of the 20th International Conference on Artificial Intelligence and Statistics, AISTATS 2017 **54** (2017)
22. Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Köpf, A., Yang, E., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., Bai, J., Chintala, S.: PyTorch: An imperative style, high-performance deep learning library. Advances in Neural Information Processing Systems **32**(NeurIPS) (2019)
23. Rieke, N., Hancox, J., Li, W., Milletarì, F., Roth, H.R., Albarqouni, S., Bakas, S., Galtier, M.N., Landman, B.A., Maier-Hein, K., Ourselin, S., Sheller, M., Summers, R.M., Trask, A., Xu, D., Baust, M., Cardoso, M.J.: The future of digital health with federated learning. npj Digital Medicine **3**(1), 1–7 (2020). https://doi.org/10.1038/s41746-020-00323-1, http://dx.doi.org/10.1038/s41746-020-00323-1
24. Wang, P., Li, Y., Reddy, C.K.: Machine learning for survival analysis: A survey. ACM Computing Surveys **51**(6) (2019). https://doi.org/10.1145/3214306
25. Wiens, J., Saria, S., Sendak, M., Ghassemi, M., Liu, V.X., Doshi-Velez, F., Jung, K., Heller, K., Kale, D., Saeed, M., Ossorio, P.N., Thadaney-Israni, S., Goldenberg, A.: Do no harm: a roadmap for responsible machine learning for health care. Nature Medicine **25**(September) (2019). https://doi.org/10.1038/s41591-019-0548-6, http://dx.doi.org/10.1038/s41591-019-0548-6
26. Yang, L., Beliard, C., Rossi, D.: Heterogeneous data-aware federated learning. arXiv (1) (2020)

27. Zhu, X., Yao, J., Huang, J.: Deep convolutional neural network for survival analysis with pathological images. Proceedings - 2016 IEEE International Conference on Bioinformatics and Biomedicine, BIBM 2016 (2), 544–547 (2017). https://doi.org/10.1109/BIBM.2016.7822579
28. Zhu, X., Yao, J., Zhu, F., Huang, J.: WSISA: Making survival prediction from whole slide histopathological images. Proceedings - 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017 **2017-Janua**, 6855–6863 (2017). https://doi.org/10.1109/CVPR.2017.725