A flexible deep learning framework for survival analysis with medical data

Gabriele Campanella[1,2(✉)], Ida Häggström[3,4(✉)], Lucas Kook[5], Torsten Hothorn[6], and Thomas J. Fuchs[1,2]

[1] Windreich Department of AI and Human Health, Icahn School of Medicine at Mount Sinai, USA
[2] Hasso Plattner Institute at Mount Sinai, Icahn School of Medicine at Mount Sinai, USA
gabriele.campanella@mssm.edu
[3] Dept of Electrical Engineering, Chalmers University of Technology, Sweden
[4] Institute of Clinical Sciences, Gothenburg University, Sweden
idah@chalmers.se
[5] Institute for Statistics and Mathematics, WU Vienna, Austria
[6] Epidemiology, Biostatistics and Prevention Institute, University of Zurich, Switzerland

Abstract. Medical imaging data and electronic health records are an integral part of clinical routine and research for prognostication of patient survival and thus directly inform patient management. However, standard regression models used to derive patient prognoses are ill-equipped to handle such non-tabular data directly. Several neural network architectures based on classification or the Cox model have been proposed. Here, we present deep conditional transformation models (DCTMs) for survival applications with medical imaging data. DCTMs include the Cox model as a special case, but parameterize the log cumulative baseline hazards via Bernstein polynomials and allow the specification of non-linear and non-proportional hazards for both tabular and non-tabular data and extend to all types of uninformative censoring. DCTMs yield moderate to large performance gains over state-of-the-art deep learning approaches to survival analysis on a multitude of publicly available datasets featuring tabular or imaging data from radiology and pathology.

Keywords: Computational pathology · deep learning · radiology · survival analysis · transformation models.

1 Introduction

Arguably one of the most important aspects of health and medical research is being able to understand, prognosticate and predict patient survival in order to improve patient management and ultimately extend their life span or time in remission [9]. Survival analysis is used for these purposes to study time-to-event information relating to for example death, response to treatment, adverse treatment effects, disease relapse, and the development of new treatments [4].

Traditional approaches, such as Cox proportional Hazards [5, CPH], relied on tabular features and are not amenable to analyze high-dimensional non-tabular data such as medical images. With recent advances in computer vision and deep learning (DL), there has been increasingly more interest in performing survival analysis directly from high-dimensional data in order to automatically learn patterns that stratify patients based on their outcome without the need for feature engineering [17,22,14,2].

1.1 State-of-the-art in deep learning for survival analysis

Early deep learning (DL) approaches to survival analysis [20] drew inspiration from the Cox proportional hazards (CPH) model [5], using the partial likelihood as a loss function and extending it to handle piece-wise constant hazards, non-proportional hazards, and non-linear effects. For instance, DeepSurv [13] optimizes the ℓ_2-regularized log partial-likelihood, while [24] introduces a mixture of Cox models using variational inference. In contrast to these semi-parametric methods, [23] proposes a fully parametric variational framework. Other approaches include a generative Weibull model [25] and piece-wise exponential hazard functions within a penalized likelihood loss [6]. Additionally, "DeepHazard" [32] uses a squared error loss based on the counting process, accommodating time-varying covariates, non-linear effects, and non-proportional hazards while accounting for censoring.

Beyond Cox-inspired methods, survival analysis can also be framed as a classification problem. For example, [30] introduce "neural multi-task logistic regression" (N-MTLR), which divides the time axis into intervals and uses a softmax final layer to model event indicators. Similarly, [18] propose "dynamic DeepHit," an extension of "DeepHit" [19], which also employs a softmax final layer and handles competing risks and non-linear effects.

Our proposed class of models belongs to the family of deep conditional transformation models [26] (DCTMs), which have been studied for continuous [1] and discrete [15] outcomes and can be understood as a conditional normalizing flow [16]. DCTMs extend the flexible class of conditional transformation models [10,11] (CTMs) with deep neural networks to handle non-tabular data, such as images.

1.2 Our contribution

In this paper we present DCTMs for survival analysis with medical data as a flexible framework for deep learning based survival analysis rooted in statistical modeling and including the CPH model as a special case.

- Our DCTMs allow the specification of non-linear and non-proportional hazards for both tabular and non-tabular (image or text) data.
- Our DCTMs can be understood as a flexible modular survival head which can be combined with arbitrary feature extractors and foundation models tailored towards specific applications.

– DCTMs are shown to yield superior prediction and discrimination performance in terms of a time-dependent c-index on several tabular, radiological and histopathological datasets.

To the best of our knowledge, this paper is the first to extend DCTMs to survival regression and present strong empirical evidence for improved performance on relevant medical imaging tasks over established approaches.

2 Background

Survival analysis. Let $(T^*, C, X) \in \mathbb{R}_+ \times \mathbb{R}_+ \times \mathcal{X}$, where T^* denotes the true event time, C denotes the censoring time and X denotes features. We observe n i.i.d. realizations $\{(t_i, x_i, \delta_i)\}_{i=1}^n$ of (T, X, δ), where $T = \min\{T^*, C\}$ and $\delta = \mathbb{1}(T < C)$ is the event indicator. Survival analysis targets the distribution of the event time T conditional on features X, on the scale of the survivor function, $S_{T|X}(t|x) := \mathbb{P}(T > t|X = x)$. One of the most popular choices is the CPH model [5],

$$S_{T|X}(t|x) = \exp(-\Lambda(t|x)) = \exp(-\Lambda_0(t)\exp(x^\top\beta)), \tag{1}$$

where $\Lambda : \mathbb{R}_+ \times \mathcal{X} \to \mathbb{R}_+$ denotes the positive, increasing cumulative hazard function. The hazard function is assumed to be decomposable into a baseline hazard $\Lambda_0 : \mathbb{R}_+ \to \mathbb{R}_+$, independent of the covariates, and multiplicative feature effects $\exp(x^\top\beta)$ with coefficients β representing log-hazard ratios [4]. The CPH model is estimated by maximizing the maximum partial likelihood.

Conditional transformation models. CTMs are distributional regression models of the form [10]

$$S_{T|X}(t|x) = 1 - F_Z(h(t|x)), \tag{2}$$

where the conditional survivor function of the response given the features $T|X = x$ is decomposed into an *a priori* chosen and parameter-free target distribution F_Z and a conditional transformation function h, which depends on the features $X = x$. In order for $S_{T|X}$ to be a valid survivor function, it is sufficient for h to be continuous and monotonically increasing in t for all $x \in \mathcal{X}$ [11]. CTMs can be estimated via maximum likelihood and allow various kinds of responses and uninformative censoring. The model in (2) is closely connected to the CPH model in (1): For $F_Z(z) = 1 - \exp(-\exp(z))$ (the minimum extreme value distribution) and $h(t|x) = \log\Lambda(t|x)$, the two models coincide.

3 DCTMs for survival analysis

We propose DCTMs (Figure 1), which parameterize the transformation function h in (2) via (deep) neural networks. For instance, let $\phi : \mathcal{X} \to \mathbb{R}^d$ denote a feature extractor, which maps the input x to a feature vector of dimension d. The

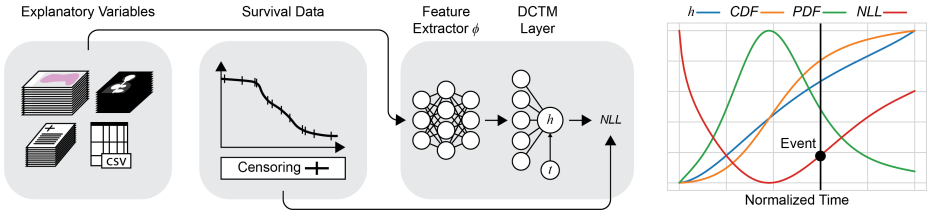

Fig. 1. Proposed deep conditional transformation model (DCTM) architecture. The model processes tabular or non-tabular explanatory variables (e.g., images) for survival data, handling exact or right-censored event times. Non-tabular inputs are mapped to a latent feature space using a feature extractor ϕ (e.g., a neural network), followed by the DCTM head. DCTMs can be trained by minimizing the negative log-likelihood (NLL). The right panel shows the transformation function (h), cumulative distribution function (CDF), probability density function (PDF), and NLL for an exact event.

extracted features are finally passed through a last layer with linear activation and weights w. We propose different parameterizations for the transformation function to control the desired complexity of the model: A *shift*, *scale* and *general* DCTM with a non-linear function in t using Bernstein polynomials. The complexity of DCTMs can be tuned via cross-validation or on a hold-out set. For all experiments, we choose the standard logistic cumulative distribution function (CDF) $\sigma(z) = (1 + \exp(-z))^{-1}$ as the target distribution. Consequently, the transformation function can be interpreted on the log-odds scale.

Assuming T takes values in an interval $[l, r] \subset \mathbb{R}_+$, the *shift* DCTM, DCTM^S, employs Bernstein polynomials of order K and additive feature effects,

$$h(t|x;\phi) = b(t)^\top \vartheta + \phi(x)^\top w, \text{ where } b : [l,r] \to \mathbb{R}^{K+1}, w \in \mathbb{R}^P,$$

where $b_k(t) := \binom{K}{k} \tilde{t}^k (1-\tilde{t})^{K-k}$ and $\tilde{t} = \frac{t-l}{r-l}$. Ensuring $\vartheta_{k+1} > \vartheta_k$, $k = 0, \ldots, K$, for all x, is sufficient to ensure monotonicity of $h(t|x;\phi)$ in t [10]. We enforce monotonicity via $\vartheta = g(\gamma) = \left(\gamma_1, \gamma_1 + \text{softplus}(\gamma_2), \ldots, \gamma_1 + \sum_{k=2}^{K+1} \text{softplus}(\gamma_k)\right)$. For the more flexible *shift scale* DCTM, DCTM^{SS}, we use

$$h(t|x;\phi) = \text{softplus}(\phi(x)^\top \beta) \cdot b(t)^\top \vartheta + \phi(x)^\top w,$$

which allows for non-proportional hazards using a scale term. Allowing the parameters of the Bernstein polynomials to fully flexibly depend on x, we arrive at the most flexible *general* DCTM, DCTM^G:

$$h(t|x;\vartheta) = b(t)^\top \vartheta(\phi(x)), \text{ where } \vartheta : \mathbb{R}^P \to \mathbb{R}^{K+1},$$

The general DCTM^G, allows a flexible baseline hazard function that captures also higher moments of $S_{T|X}$, resulting in a distribution-free model [27,16].

3.1 Training and evaluating DCTMs

DCTMs are fitted by minimizing the empirical NLL: A single observation (t, x, δ) contributes

$$\mathcal{L}(h; t, x, \delta) = \begin{cases} \sigma(h(t|x))(1 - \sigma(h(t|x))h'(t|x) & \delta = 1, \\ 1 - \sigma(h(t|x)) & \delta = 0. \end{cases}$$

to the NLL. The empirical NLL is then given by $\text{NLL} = -\sum_{i=1}^{n_t} \log \mathcal{L}(h; t_i, x_i, \delta_i)$, and minimized over the n_t training samples using any deep learning optimization routine. Interval- and left-censored observations can also be handled with the proposed method see, e.g., [11].

After fitting a DCTM, the conditional survivor function of a test observation can be computed from the estimated parameters via $\widehat{S}_{T|X}(t|x) = 1 - F_Z(\widehat{h}(t|x))$. Now, all evaluation metrics can be computed from (some form of) the predicted conditional distribution. The time-dependent c-index [8] is defined as

$$c(t) = \frac{\sum_{i=1}^{n} \sum_{j=1}^{n} \mathbb{1}(t_j < t_i) \cdot \mathbb{1}(\widehat{\Lambda}(t|x_j) > \Lambda(t|x_i)) \cdot \mathbb{1}(t_j < t)}{\sum_{i=1}^{n} \sum_{j=1}^{n} \mathbb{1}(t_j < t_i) \cdot \mathbb{1}(t_j < t)},$$

and measures the concordance between event times and predicted hazards. The integrated inverse probability of censoring weighted Brier score (IBS)

$$\text{IBS} = \int_0^{\max_i t_i} \frac{1}{n} \sum_{i=1}^{n} \frac{\delta_i \mathbb{1}(t_i \leq t) \widehat{S}_{T|X}(t|x_i)^2 + I(t_i > t)(1 - \widehat{S}_{T|X}(t|x_i))^2}{\widehat{G}(t_i)},$$

measures quality of probabilistic predictions, where $\widehat{G}(t)$ denotes an estimate of the unconditional probability of not being censored until time t.

4 Experiments and results

We extensively evaluate DCTMs against competing methods on several datasets involving tabular features, computed tomography (CT) images, and histopathology whole slide images. The experiments are carried out as ablation studies to explore the flexible DCTM framework with varying parametrization complexity and order K of the Bernstein polynomial, denoted $\text{DCTM}_K^{\{S,SS,G\}}$. C-indices were calculated at the training cohort event time quantiles 25%, 50%, 75% and 100%, and IBS was integrated over time steps of size 0.01 from 0 to the largest test set event time. All DCTMs and experiments were implemented in Pytorch, unless indicated otherwise. The source code for the DCTM framework is available on GitHub at https://github.com/sinai-computational-pathology/DCTM.

4.1 Tabular data

We first compare our DCTMs to state-of-the-art methods (CPH model, random survival forest [RSF], and DeepSurv) on tabular benchmark datasets. The GBSG

Table 1. Test c-indices (CI, ↑) evaluated at different event time quantiles and IBS (↓) for several models on three tabular datasets. DS=DeepSurv.

	Metric		CPH	RSF	DS	DCTM_1^S	DCTM_{10}^S	DCTM_1^{SS}	DCTM_{10}^{SS}	DCTM_1^G	DCTM_{10}^G
GBSG	CI	25	0.600	0.638	0.655	0.705	0.696	0.686	0.687	0.700	**0.707**
		50	0.605	0.619	0.650	0.691	0.684	0.676	0.682	0.684	**0.691**
		75	0.600	0.603	0.648	0.687	**0.679**	0.674	0.678	0.676	**0.679**
		100	0.600	0.603	0.648	0.687	**0.679**	0.674	0.678	0.676	**0.679**
	IBS		0.125	0.129	0.180	**0.110**	0.115	0.116	0.116	0.115	0.115
METABRIC	CI	25	0.591	0.669	0.563	0.643	0.654	0.684	0.676	**0.694**	0.692
		50	0.601	0.605	0.567	0.648	0.652	0.659	0.645	**0.666**	0.653
		75	0.594	0.566	0.550	0.631	**0.634**	0.633	0.616	**0.634**	0.615
		100	0.589	0.545	0.548	0.629	**0.631**	0.620	0.576	0.615	0.605
	IBS		0.215	0.251	**0.203**	0.227	0.226	0.220	0.229	0.224	0.230
SUPPORT	CI	25	0.569	**0.642**	0.540	0.586	0.603	0.602	0.619	0.602	0.612
		50	0.575	**0.621**	0.520	0.605	0.612	0.611	0.610	0.613	0.616
		75	0.577	0.601	0.517	0.609	0.614	0.614	0.611	**0.616**	**0.616**
		100	0.571	0.588	0.515	0.610	0.614	0.615	0.610	**0.616**	0.610
	IBS		0.251	0.273	0.304	0.253	0.243	0.235	0.235	0.236	**0.231**

dataset contains 686 breast cancer patients with 56% censoring and 7 features. The METABRIC data contains 1980 breast cancer patients with 42% censoring and 9 features. The SUPPORT dataset contains 9105 hospitalized adults, 32% censoring and 14 features. We use the same train/test splits as in [13]. For each dataset, explanatory variables were z-normalized using the training split, while the time to event data was normalized in the range 0–1. Cox and RSF models were trained using the survival and ranger packages in R, respectively. For DeepSurv and DCTM, the encoding portion of the model consisted of a linear layer followed by a ReLU non-linearity. The survival head consisted of a linear layer projecting features to the risk score for DeepSurv, and a DCTM survival head with different parametrizations using $K = 1$ or 10 bases. Models were trained for 15 epochs with a learning rate of 0.001 using the L-BFGS optimizer.

Results are presented in Table 1. In terms of c-index, except for the two earlier time points in the SUPPORT data, DCTMs outperform CPH, RSF and DeepSurv on all other datasets and for all time points. For the GBSG data, the most complex DCTM_{10}^G performs best, whereas the simpler DCTM_1^G and shift-scale ($K = 10$) perform best on the METABRIC and SUPPORT data. In terms of probabilistic predictions (IBS), DCTMs outperform the baseline methods on the GBSG and SUPPORT data, while DeepSurv performs best on METABRIC.

4.2 Radiology data

The radiology data comprised 3346 volume CT scans of patients with head & neck cancer from the open RADCURE cohort with curated disease-free survival time [29]. 780 (23%) subjects had an event and the rest were censored. The

Table 2. Cross-validation c-indices (CI, ↑) of the test set for the models evaluated at different event time quantiles of the CT-image dataset, and IBS (↓). Values are average±std (%). DS=DeepSurv, DH=DeepHit.

	Metric		DS	DH	DCTM_1^S	DCTM_{10}^S	DCTM_1^{SS}	DCTM_{10}^{SS}	DCTM_1^G	DCTM_{10}^G
RADCURE	CI	25	66.2±1.2	73.0±1.1	**75.0**±0.7	74.8±0.9	74.2±0.8	74.8±0.9	74.4±0.6	74.2±0.4
		50	65.8±1.1	72.0±1.0	**73.4**±0.7	73.3±0.7	72.9±0.7	73.3±0.9	72.9±0.6	72.8±0.5
		75	65.8±1.1	72.0±1.0	**73.4**±0.7	73.3±0.7	72.9±0.7	73.2±0.9	72.9±0.6	72.8±0.5
		100	65.8±1.1	50.8±21.7	**73.4**±0.7	73.3±0.7	73.0±0.6	73.0±1.0	73.0±0.6	72.7±0.5
	IBS		13.2±0.1	16.7±0.0	10.7±0.2	10.2±0.5	10.5±0.2	9.9±0.3	10.3±0.3	**9.7**±0.3

planned target volumes (i.e. tumor ROIs as RTstructs) for radiotherapy were also available, which were converted to binary masks. All CTs and tumor masks were resampled to an isotropic voxel size of $3 \times 3 \times 3$ mm. Of the cohort, 750 (22%) patients were held out for testing according to RADCURE splits, and the remaining 2596 (78%) were randomly divided for 5-fold cross-validation. The event times were normalized to range 0–1, and the CT images were clipped to $[-1000, +1000]$ HU and min-max normalized to 0–1. The tumor ROI masks were added as a second image channel. A 3d ResNet34 backbone [7] was used for feature extraction (size 512) of the image volumes, followed by a survival head based on DeepSurv, DeepHit or variants of our DCTM. We used 3000 time thresholds in DeepHit. The DCTM variants were trained by minimizing NLL, DeepSurv by minimizing the log partial-likelihood [13] and DeepHit by minimizing cross entropy (CE) [19] (no ranking loss). For DeepSurv and DeepHit we calculated a risk score (log cumulative hazard for DeepSurv and survival probability for DeepHit) for each patient by including an additional trainable linear layer in the model with image features as input and output size 1 for DeepSurv, and 3000 for DeepHit. Models were trained for 20 epochs with a learning rate of 0.001 using the SGD optimizer. The final results were calculated as the average across the five cross-validation models, applied to the test set.

Results are shown in Table 2. The DCTM models outperform all competitors at each quantile by a considerable margin. While the simplest DCTM_1^S achieves the best discrimination performance, the IBS indicates that the most complex DCTM_{10}^G produces the best probabilistic predictions.

4.3 Histopathology data

For the histopathology experiments we leveraged three cohorts from the TCGA [28] dataset containing overall survival information and whole slide images: BRCA—1023 breast cancer patients with 86% censoring, LUAD—55 lung adenocarcinoma patients with 65% censoring, and UCEC—548 patients with uterine corpus endometrial carcinoma with 83% censoring.

For the analysis of the digital slides we extracted features using the state-of-the-art foundation model UNI [3]. The models then consisted of a gated MIL attention model [12] to aggregate features over the slide and a survival head. We

Table 3. Cross-validation c-indices (CI, ↑) for models evaluated at different event time quantiles of three histopathology datasets, and IBS (↓). Values are average±std (%). DH=DeepHit.

	Metric		DH [21]	DCTM_1^S	DCTM_{10}^S	DCTM_1^{SS}	DCTM_{10}^{SS}	DCTM_1^G	DCTM_{10}^G
BRCA	CI	25	68.9±9.1	70.5±10.7	66.4±6.0	70.0±9.4	73.3±0.9	**74.2**±7.4	70.3±11.2
		50	64.2±8.2	69.0±8.3	65.6±9.9	67.8±7.7	68.7±3.7	**71.0**±5.0	66.6±9.5
		75	63.3±8.1	67.1±7.5	64.6±7.8	66.6±8.8	67.6±4.0	**69.7**±6.0	64.7±7.8
		100	62.5±7.5	66.9±6.4	64.3±6.9	65.3±5.3	66.0±6.0	**67.5**±6.3	63.2±8.7
	IBS		22.2±4.2	10.8±0.6	10.0±1.7	**9.3**±1.1	9.7±2.8	9.5±1.4	9.6±1.3
LUAD	CI	25	64.5±7.7	**65.4**±6.9	62.2±3.8	61.1±7.3	62.9±6.6	60.7±8.1	61.2±5.9
		50	**62.3**±8.0	62.2±7.9	61.2±5.9	60.5±4.1	60.9±7.3	59.0±5.2	60.1±5.3
		75	61.1±6.1	**61.8**±5.0	61.0±5.2	61.0±5.2	61.1±5.0	59.0±6.5	59.4±5.3
		100	**62.1**±5.0	61.7±4.7	61.0±4.5	58.7±8.0	60.0±5.9	57.6±7.1	57.4±5.0
	IBS		29.9±9.1	19.2±3.1	17.1±4.9	**14.9**±3.5	16.1±4.3	15.4±4.4	16.6±4.4
UCEC	CI	25	64.5±5.1	66.0±6.1	65.3±3.3	68.6±3.2	**69.9**±6.0	68.0±9.1	68.2±7.0
		50	62.6±4.9	65.8±4.6	65.8±8.7	67.3±4.7	68.3±6.2	**68.4**±6.5	67.1±5.1
		75	64.5±8.1	67.0±5.5	66.8±7.5	67.7±4.7	67.9±5.5	**68.3**±8.3	67.7±5.7
		100	66.9±7.9	67.1±5.6	66.9±7.7	67.7±6.7	**68.2**±6.1	68.1±9.1	67.3±8.5
	IBS		21.0±8.9	9.4±2.6	9.6±4.2	**8.4**±2.5	8.7±3.1	8.5±2.6	8.8±2.9

compared our proposed DCTM head with the DeepHit model as implemented in a recent benchmark study [21]. We omitted DeepSurv as it is incompatible with batch sizes of one as used in histopathology experiments due to size. The models were trained using splits from [21] in a 5-fold cross-validation scheme for 100 epochs following a cosine annealing learning rate schedule with maximum learning rate of 0.0002 using the AdamW optimizer.

Results are shown in Table 3. For the BRCA and UCEC data, the general DCTM ($K = 1$) or shift-scale DCTM ($K = 10$) perform best in terms of c-index. For the LUAD dataset, the simplest shift DCTM ($K = 1$) performs best at the 25th and 75th percentile in terms of c-index, while DeepHit performs best for the 50th and 100th percentile. In terms of IBS, the shift-scale DCTM with $K = 1$ performs best on all three datasets.

5 Discussion and conclusions

We have proposed DCTMs for survival analysis and demonstrated their effectiveness on a wide range of datasets featuring both tabular and imaging data, as well as various neural network architectures, feature extractors and foundation models. In terms of time-dependent c-index, DCTMs outperform state-of-the-art models (DeepSurv, DeepHit, RSF, and CPH) across most data modalities and perform on par otherwise. Our ablations show that the various degrees of flexibility of DCTMs aids in trading off flexibility and prediction performance. In particular, our results on the radiology data indicate that lower order Bernstein polynomials ($K = 1$) yield better discrimination but lower probabilistic

prediction performance than higher order ($K = 10$). Furthermore, the most flexible model, DCTM^G, does not typically perform best in terms of discrimination. The type of explanatory data and their relationship with the survival information likely govern the optimal DCTM complexity. In practice, this complexity can be chosen based on a validation split or cross-validation and thereby reduce the risk of overfitting. Lastly, our parameterization of the NLL in terms of Bernstein polynomials and the sigmoid CDF yielded more stable training curves - a common problem when optimizing the partial likelihood (e.g. DeepSurv) - and avoids introducing bias as in CE losses [31] (e.g. DeepHit). Using the time-dependent c-index for evaluation is necessary due to the non-linear nature of some of the DCTMs and allows a more fine-grained view on model performance.

Taken together, DCTM for survival analysis have been shown to be a versatile neural network survival head which can be combined with various input data modalities, feature extraction methods and foundation models, and yield prediction and discrimination performance superior or at least similar to state-of-the-art deep learning methods for survival outcomes.

Acknowledgments. GC, IH, and TF were supported in part through the NIH/NCI Cancer Center Support Grant (P30 CA008748). LK was supported by Novartis Research Foundation (FreeNovation 2019) and the Swiss NSF (S-86013-01-01; S-42344-04-01). TH was supported by the Swiss NSF (200021_184603). **Disclosure of Interests.** TF is a founder and equity owner of Paige.AI, and Chief AI Officer at Eli Lilly.

References

1. Baumann, P.F.M., Hothorn, T., Rügamer, D.: Deep conditional transformation models. In: Machine Learning and Knowledge Discovery in Databases. Research Track. pp. 3–18. Springer-Verlag (2021)
2. Byun, S.S., Heo, T.S., Choi, J.M., Jeong, Y.S., Kim, Y.S., Lee, W.K., Kim, C.: Deep learning based prediction of prognosis in nonmetastatic clear cell renal cell carcinoma. Scientific Reports **11**(1) (2021)
3. Chen, R.J., Ding, T., Lu, M.Y., Williamson, D.F., Jaume, G., Chen, B., Zhang, A., Shao, D., Song, A.H., Shaban, M., et al.: Towards a general-purpose foundation model for computational pathology. Nature Medicine **30**, 850—862 (2024)
4. Collett, D.: Modelling Survival Data in Medical Research. CRC press (2015)
5. Cox, D.R.: Regression Models and Life-Tables. Journal of the Royal Statistical Society B **34**(2), 187–202 (1972)
6. Fornili, M., Ambrogi, F., Boracchi, P., Biganzoli, E.: Piecewise Exponential Artificial Neural Networks (PEANN) for Modeling Hazard Function with Right Censored Data (2014), lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)
7. Hara, K., Kataoka, H., Satoh, Y.: Can spatiotemporal 3d cnns retrace the history of 2d cnns and imagenet? In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 6546–6555 (2018), GitHub: https://github.com/kenshohara/3D-ResNets-PyTorch
8. Harrell, F.E.: Evaluating the yield of medical tests. JAMA: The Journal of the American Medical Association **247**(18), 2543 (1982)

9. Hosny, A., Parmar, C., Coroller, T.P., Grossmann, P., Zeleznik, R., Kumar, A., Bussink, J., Gillies, R.J., Mak, R.H., Aerts, H.J.W.L.: Deep learning for lung cancer prognostication: A retrospective multi-cohort radiomics study. PLOS Medicine **15**(11), e1002711 (2018)
10. Hothorn, T., Kneib, T., Bühlmann, P.: Conditional Transformation Models. Journal of the Royal Statistical Society. Series B: Statistical Methodology **76**(1), 3–27 (2014)
11. Hothorn, T., Möst, L., Bühlmann, P.: Most Likely Transformations. Scandinavian Journal of Statistics **45**(1), 110–134 (2018)
12. Ilse, M., Tomczak, J., Welling, M.: Attention-based deep multiple instance learning. In: The International Conference on Machine Learning (ICML). pp. 2132–2141 (2018)
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) (2018), GitHub: https://github.com/jaredleekatzman/DeepSurv
14. Kim, D.W., Lee, S., Kwon, S., Nam, W., Cha, I.H., Kim, H.J.: Deep learning-based survival prediction of oral cancer patients. Scientific Reports **9**(1) (2019)
15. Kook, L., Herzog, L., Hothorn, T., Dürr, O., Sick, B.: Deep and interpretable regression models for ordinal outcomes. Pattern Recognition **122**, 108263 (2022)
16. Kook, L., Kolb, C., Schiele, P., Dold, D., Arpogaus, M., Fritz, C., Baumann, P.F.M., Kopper, P., Pielok, T., Dorigatti, E., Rügamer, D.: How inverse conditional flows can serve as a substitute for distributional regression. In: The 40th Conference on Uncertainty in Artificial Intelligence (2024)
17. Lao, J., Chen, Y., Li, Z.C., Li, Q., Zhang, J., Liu, J., Zhai, G.: A Deep Learning-Based Radiomics Model for Prediction of Survival in Glioblastoma Multiforme. Scientific Reports **7**(1), 1–8 (2017)
18. Lee, C., Yoon, J., van der Schaar, M., Van Der Schaar, M.: Dynamic-Deephit: A Deep Learning Approach for Dynamic Survival Analysis with Competing Risks Based on Longitudinal Data. IEEE Transactions on Biomedical Engineering **67**(1), 122–133 (2019)
19. Lee, C., Zame, W., Yoon, J., van der Schaar, M.: Deephit: A deep learning approach to survival analysis with competing risks. Proceedings of the AAAI Conference on Artificial Intelligence **32**(1) (2018), GitHub: https://github.com/chl8856/DeepHit
20. Liestbl, K., Andersen, P.K., Andersen, U.: Survival Analysis and Neural Nets. Statistics in Medicine **13**(12), 1189–1200 (1994)
21. Ma, J., Guo, Z., Zhou, F., Wang, Y., Xu, Y., Cai, Y., Zhu, Z., Jin, C., Lin, Y., Jiang, X., Han, A., Liang, L., Chan, R.C.K., Wang, J., Cheng, K.T., Chen, H.: Towards a generalizable pathology foundation model via unified knowledge distillation. arXiv preprint arXiv:2407.18449 (2024)
22. Matsuo, K., Purushotham, S., Jiang, B., Mandelbaum, R.S., Takiuchi, T., Liu, Y., Roman, L.D.: Survival Outcome Prediction in Cervical Cancer: Cox Models vs Deep-Learning Model. American Journal of Obstetrics and Gynecology **220**(4), 381.e1–381.e14 (2019)
23. Nagpal, C., Li, X., Dubrawski, A.: Deep survival machines: Fully parametric survival regression and representation learning for censored data with competing risks. IEEE Journal of Biomedical and Health Informatics **25**(8), 3163–3175 (2021)
24. Nagpal, C., Yadlowsky, S., Rostamzadeh, N., Heller, K.: Deep cox mixtures for survival regression. In: Machine Learning for Healthcare Conference. pp. 674–708. PMLR (2021)

25. Ranganath, R., Perotte, A., Elhadad, N., Blei, D.: Deep survival analysis. In: Doshi-Velez, F., Fackler, J., Kale, D., Wallace, B., Wiens, J. (eds.) Proceedings of the 1st Machine Learning for Healthcare Conference. Proceedings of Machine Learning Research, vol. 56, pp. 101–114. PMLR, Northeastern University, Boston, MA, USA (2016)
26. Sick, B., Hothorn, T., Durr, O.: Deep transformation models: Tackling complex regression problems with neural network based transformation models. In: 2020 25th International Conference on Pattern Recognition (ICPR). IEEE (2021)
27. Siegfried, S., Kook, L., Hothorn, T.: Distribution-Free Location-Scale Regression. The American Statistician **0**, 1–18 (2023)
28. TCGA: https://www.cancer.gov/ccg/research/genome-sequencing/tcga
29. Welch, M.L., Kim, S., Hope, A.J., Huang, S.H., Lu, Z., Marsilla, J., Kazmierski, M., Rey-McIntyre, K., Patel, T., O'Sullivan, B., Waldron, J., Bratman, S., Haibe-Kains, B., Tadic, T., Princess Margaret Head and Neck Site Group: RADCURE: An open-source head and neck cancer CT dataset for clinical radiation therapy insights (version 4) [dataset]. Medical Physics **51**(4), 3101–3109 (2024), https://cancerimagingarchive.net/collection/radcure
30. Yun, S., Du, B., Mao, Y.: Robust deep multi-task learning framework for cancer survival analysis. In: 2021 International Joint Conference on Neural Networks (IJCNN). pp. 1–8 (2021)
31. Zadeh, S.G., Schmid, M.: Bias in cross-entropy-based training of deep survival networks. IEEE Transactions on Pattern Analysis and Machine Intelligence **43**(9), 3126–3137 (2020)
32. Zhang, Y.N., Peng, H.F.: Zhang neural network for linear time-varying equation solving and its robotic application. In: 2007 International Conference on Machine Learning and Cybernetics. vol. 6, pp. 3543–3548 (2007)