Supplementary A Survival Dataset

In a survival analysis dataset consisting of N patients, the data for each individual is represented as (x_i, t_i, δ_i), where $x_i \in \mathbb{R}^d$ is the feature set for the i-th patient and $t_i \in \mathbb{R}_+$ is the survival time which is either the censored time or event time. $\delta_i \in \{0, 1\}$ indicates if the patient was censored or not, where $\delta_i = 0$ indicates that the i-th patient was censored and t_i is the censoring time, and $\delta_i = 1$ means that the patient experienced the event (death), and t_i is the time to event. Hence, a survival dataset is represented as $\mathcal{D} = \{(\boldsymbol{x}_i, t_i, \delta_i)\}_{i=1}^{N}$.

Supplementary B Individualized Survival Distribution (ISD)

A patient's ISD curve shows their likelihood of survival as time progresses. This is the probability of survival until time t given the patient's features of x_i and time t, and it is represented as $S(\, t \mid \boldsymbol{x}_i \,) = P(T > t \mid \mathbf{X} = \boldsymbol{x}_i)$. The ISD curve begins with a survival probability of 1 at time zero and gradually declines thereafter. Each ISD is specific to an instance using specific clinical data from that individual patient (x_i), distinguishing them from curves like the KM curve [6], which are derived from an entire population's data.

If one needs a single value, many use time-to-event prediction given by the model's output (ISD curve), either mean (denoted by $\mathbb{E}_t[\, S(\, t \mid \boldsymbol{x}_i \,)\,]$) or median survival time (denoted by median$(S(\, t \mid \boldsymbol{x}_i \,))$. The truncated adaptations of the mean (expected) and median survival time with respect to time τ are defined as follows:

$$\begin{aligned} \hat{t}_{i,\text{T-mean},\tau} = &\quad \min\{\, \mathbb{E}_t[\, S(\, t \mid \boldsymbol{x}_i \,)\,] \,, \quad \tau\} \\ = &\quad \min\{\, \int_0^\infty S(\, t \mid \boldsymbol{x}_i \,)\ dt,\ \tau\}, \end{aligned} \tag{B1}$$

and

$$\begin{aligned} \hat{t}_{i,\text{T-median},\tau} = &\quad \min\{\, \text{median}(S(\, t \mid \boldsymbol{x}_i \,)),\ \tau\} \\ = &\quad \min\{\, S^{-1}(\, P = 0.5 \mid \boldsymbol{x}_i \,),\ \tau\}, \end{aligned} \tag{B2}$$

where τ represents the time point that we truncate, $\boldsymbol{x}_i$ denotes the attributes of patient i, $S(\, t \mid \boldsymbol{x}_i \,)$ is the predicted ISD curve for this patient, and S^{-1} is the inverse function of the survival function S. τ can be set to any time point depending on the application, here in this study, we set it to be the final time point (length of the study). In this study, we use the truncated median time (Equation B2) as the prediction time.

Note that the ISD curve often does not cross the probability of 0.5. In such cases, the common approach for calculating the standard median time is to linearly extrapolate the curve until it reaches the 0.5 probability – we draw a line from the initial time point with a probability of 1 to the final time point, then continue this line until it intersects with the probability of either 0 or 0.5. However, in the case of truncated

median time (Equation B2), extrapolation is not required, as we bound the prediction by τ. For example, in Figure B1 left, the median of the ISD curve is 22 months, which is less than $\tau = 200$ months, which is the end of the study – here, we set the time to event prediction to the median time (22 months). For Figure B1 right, the ISD curve ends before reaching the probability of 0.5, and as a result, we know that the median time is after the end of the study. Since we take the minimum of the median time and the end of the study time (200), we set the truncated prediction time to 200 months. Note this means that we do not need to extrapolate the ISD curve.

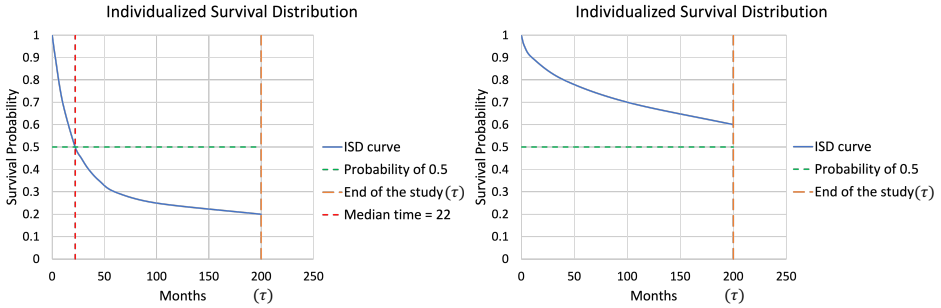

Fig. B1: ISD curves for two patients, for a study that ended at $\tau = 200$ months. The truncated time to event prediction using the median time on the left side is 22 months and on the right one is 200 months.

Supplementary C Features

Table C1 lists the features we used for each type of cancer. Note that we chose not to do the feature selection step as: (1) the number of included features was less than 20, and (2) we viewed this as a distraction from the primary objective of our research.

Supplementary D Motivation for Truncated MAE

In this study, we proposed the truncated variation of MAE-PO; this section motivates this variation one step further. In Section 3, we discussed truncating the predicted time-to-event, which is the median time of the ISD curve, and the same issue is raised in the context of KM curves. When dealing with the KM curve of datasets with high censorship, this curve often fails to descend to zero and might not even cross the 0.5 survival probability threshold. Consequently, the median time, typically employed as a time-to-event prediction, is unknown. Among our included datasets, as illustrated in Figure 2, for cancers of breast, kidney and renal pelvis, prostate, thyroid, and urinary bladder, the blue KM curve does not intersect the green line (representing 0.5 probability) by the study's conclusion.

Some prior studies have attempted to address this matter, proposing: (1) dropping the curve vertically to zero post-study conclusion, (2) employing linear extrapolation

Table C1: List of features included in each type of cancer dataset. In this table, prostate refers to prostate # 2.

Feature	Brain	Breast	Kidney	Liver	Lung	Prostate	Stomach	Thyroid	Urinary
Age	✓	✓	✓	✓	✓	✓	✓	✓	✓
Sex	✓	-	✓	✓	✓	-	✓	✓	✓
Behavior recode for analysis	✓	-	-	-	✓	-	-	-	-
Combined Summary Stage	✓	✓	✓	✓	✓	✓	✓	✓	✓
Grade	✓	✓	✓	✓	✓	✓	✓	-	✓
RX Summ–Scope Reg LN Sur	✓	✓	✓	✓	✓	✓	✓	✓	✓
RX Summ–Surg Oth Reg/Dis	✓	✓	✓	✓	✓	✓	✓	✓	✓
RX Summ–Surg Prim Site	✓	✓	✓	✓	✓	✓	✓	✓	✓
Summary stage 2000	✓	✓	✓	✓	✓	✓	✓	✓	✓
SEER historic stage A	✓	✓	✓	✓	✓	✓	✓	✓	✓
Derived AJCC T, 6th ed (2004-2015)	-	✓	✓	✓	✓	✓	✓	✓	✓
Derived AJCC N, 6th ed (2004-2015)	-	✓	✓	✓	✓	✓	✓	✓	✓
Derived AJCC M, 6th ed (2004-2015)	-	✓	✓	✓	✓	✓	✓	✓	✓
Breast - Adjusted AJCC 6th Stage	-	✓	-	-	-	-	-	-	-
Derived AJCC Stage Group, 6th ed	-	✓	✓	✓	✓	✓	✓	✓	✓
Derived AJCC Stage Group, 7th ed	-	✓	✓	✓	✓	✓	✓	✓	✓
Invasion Beyond Capsule Recode	-	-	✓	-	-	-	-	-	-
Gleason Patterns Clinical Recode	-	-	-	-	-	✓	-	-	-
Gleason Patterns Pathological Recode	-	-	-	-	-	✓	-	-	-
Gleason Score Clinical Recode	-	-	-	-	-	✓	-	-	-
Gleason Score Pathological Recode	-	-	-	-	-	✓	-	-	-
PSA Lab Value Recode	-	-	-	-	-	✓	-	-	-
Number of Cores Positive Recode	-	-	-	-	-	✓	-	-	-
Number of Cores Examined Recode	-	-	-	-	-	✓	-	-	-

(which we illustrated in Figure D2), and (3) applying a specific function or distribution to extend the curve [42, 43]. However, Rich et al. [44] noted that any form of KM curve extrapolation lacks justification, and any prediction after the study conclusion is unreliable. Take the prostate # 1 dataset as an instance, where the survival curve does not reach the probability of 0.5. If we use linear extrapolation – from the starting point of the curve (0,1) to the final time point, then continue the line to reach the probability of 0.5 or 0 – to continue the curve and find the median time, as demonstrated in Figure D2, then we can see that linear extrapolation exceeds 2100 months (175 years) of survival, and the median time is 1121 months (93 years). Given that the age average for the prostate #1 cancer dataset is 65 years, then a prediction of $65 + 93 = 158$ years is a wrong and unrealistic prediction.

Therefore, we follow the same suggestion as Rich et al [44], meaning that we drop the ISD cure vertically to zero post-study conclusion, bound the predictions of trained models and the best guess estimate for actual time to event by the length of the study (τ), as any prediction beyond the conclusion of the study is unreliable and lacks justification.

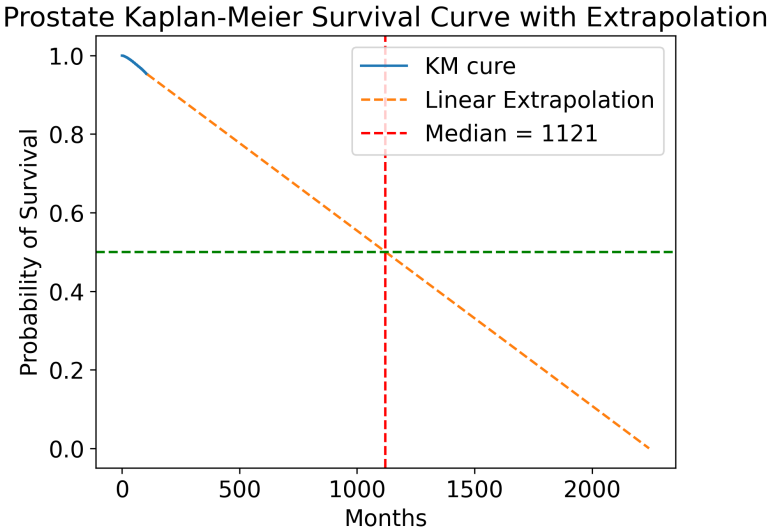

Fig. D2: KM curve linear extrapolation for prostate #1 dataset.

Supplementary E Evaluation Metrics in Details

In this section, we explain the formula of evaluation metrics and describe them in detail. Note that we used the SurvivalEVAL [45] package to implement this section.

1. **C-index:**

The C-index of a model, on a labeled survival dataset, is given by

$$\text{C-index}(S(.|.), \mathcal{D}) \quad = \quad \frac{\text{Number of concordant pairs}}{\text{Number of comparable pairs}}\,, \tag{E3}$$

where a pair of instances is considered concordant if the predicted and the actual outcome follow the same ranking. Among all possible combinations of two subjects from a sample size of N, a comparable pair means we know which one of the subjects experienced the event first. For example, as shown in Figure E3, patients A and B can be considered a comparable pair because it is clear that the event occurred first with patient A. In contrast, patients B and C do not form a comparable pair since patient B is censored prior to patient C's event, leaving ambiguity about whether patient B experienced the event before or after patient C. Hence, for patients B and C we do not know who experienced the event first, and remains uncertain and incomparable. Additionally, any two patients who are not censored are comparable, making patients A and C a comparable pair. Therefore in Figure E3, we have 2 comparable pairs: {A, B}, and {A, C}.

After computing the number of comparable pairs, given the model's prediction versus the ground truth, we compute the number of concordant pairs. So following our example, if we predict the following time to events: $A = 5$, $B = 13$, $C = 8$, then we have correctly ranked both of our comparable pairs, since time to event prediction for B is greater than A, and C is also greater than A. Thus, C-index is equal to 1.

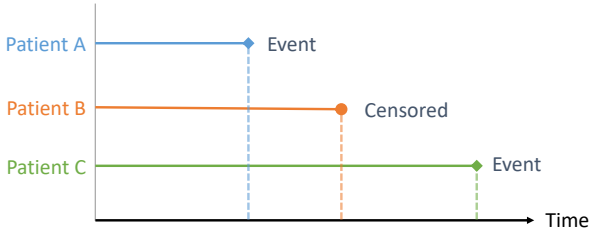

Fig. E3: Time to event/censorship for three patients.

2. **Brier Score:**
Brier Score (BS) is the squared difference between the predicted probability of survival at a specific time t and the true event value (0 or 1) [29]. It ranges between zero to one, and a value of zero means perfect prediction. For censored patients with unknown event values, BS uses the inverse probability censoring weight (IPCW) [46], which uniformly transfers each censored patient's weight to uncensored patients after that time.

BS is defined as:

$$BS(t, \mathcal{D}) \quad = \quad \frac{1}{N}\sum_{i=1}^{N} \frac{(0 - S_{\text{m}}(\,t \mid \boldsymbol{x}_i\,))^2 \cdot \mathbb{1}_{t_i \leq t, \delta_i = 1}}{G_i(t_i)} \;+\; \frac{(1 - S_{\text{m}}(\,t \mid \boldsymbol{x}_i\,))^2 \cdot \mathbb{1}_{t_i > t}}{G_i(t)}\,,$$

where $G_i(t)$ is the probability of not being censored until time t, which is commonly estimated by running the KM algorithm, but with the censor-bit (event flag) flipped.

3. **MAE:**
Mean absolute error (MAE) measures the average absolute difference between the predicted time ($\hat{t}_i$) and the actual (truth) time (t_i):

$$\text{MAE}(\{\hat{t}_i\},\ \{t_i\}) \quad = \quad \frac{1}{N}\sum_{i=1}^{N} |\hat{t}_i - t_i| \ . \tag{E4}$$

For the prediction time ($\hat{t}_i$), we use the median time of the ISD model ($\hat{t}_i = S^{-1}(\, P = 0.5 \mid \boldsymbol{x}_i\,)$). However, to compute this MAE, the actual time (t_i) is unknown for censored patients. Hence, we need to use another variation of MAE that can estimate the truth time for censored patients. In this study, we use the MAE-PO that employs pseudo-observation to estimate the actual time of survival for censored patients [28].

4. **MAE-PO:**
Qi et al. [28] proposed the MAE-PO that employs pseudo-observation to estimate the actual time of survival for censored patients and uses $\hat{\theta}$ as a predictor, which can be based on the mean value of the KM estimator, $\hat{\theta} = \mathbb{E}_t[S_{\text{KM}(\mathcal{D})}(\, t\,)]$, where $S_{\text{KM}(\mathcal{D})}(\, t\,)$ is the group-level survival probability, estimated using KM model on the dataset $\mathcal{D}$. The idea here is that we measure the contribution of patient i to the unbiased predictor $\hat{\theta}$. The best guess for MAE-PO can be defined as:

$$e_{\text{T-pseudo-obs}}(t_i, \mathcal{D}) \quad = \quad N \times \hat{\theta} \ - \ (N-1) \times \hat{\theta}^{-i} \ , \tag{E5}$$

where $\hat{\theta}^{-i}$ is $\mathbb{E}_t[S_{\text{KM}(\mathcal{D}^{-i})}(\, t\,)]$, the predictor applied to the $N-1$ data instances, after removing the patient i. This best guess can be unreliable for patients who get censored earlier in the study since we do not have much information about them. Therefore, as suggested by Haider et al. [5], we assign less confidence weight to the best guess of early censored patients. This confidence weight is calculated as:

$$\omega_i \quad = \quad 1 - S_{\text{KM}(\mathcal{D})}(t_i) \ . \tag{E6}$$

Note ω_i is zero in the beginning (at time zero), and increases after that. Lastly, MAE-PO is defined as:

$$\mathbb{E}_{i\sim\mathcal{D}}[\mathcal{R}_{\text{MAE-PO}}(\hat{t}_i, t_i, \delta_i)] = \tag{E7}$$
$$\frac{1}{\sum_{i=1}^{N}\omega_i}\sum_{i=1}^{N}\omega_i \left|[(1-\delta_i)\cdot e_{\text{T-pseudo-obs}}(t_i,\mathcal{D}) + \delta_i \cdot t_i] - \hat{t}_i\right| ,$$

where symbol $\mathcal{R}$ means a scoring rule, which is used to compute the MAE-PO error. Note that here the prediction time ($\hat{t}_i$) is the median of the ISD model.

5. Truncated MAE-PO:

As discussed in Section 3, we choose to bound the prediction time and best guess by the end of the study and use the truncated variation of MAE-PO. Hence, the best guess for truncated MAE-PO can be defined as:

$$e_{\text{T-pseudo-obs},\tau}(t_i, \mathcal{D}) \quad = \quad \min\{e_{\text{pseudo-obs}}(t_i, \mathcal{D}) \ \ , \tau\}, \tag{E8}$$

where $e_{\text{pseudo-obs}}(t_i, \mathcal{D})$ is defined using Euqation E5. Further, we use the same weighting as described in Equation E6. Therefore, the truncated MAE-PO is defined as follows:

$$\mathbb{E}_{i\sim\mathcal{D}}[\mathcal{R}_{\text{T-MAE-PO},\tau}(\hat{t}_i,\, t_i,\, \delta_i)] = \tag{E9}$$
$$\frac{1}{\sum_{i=1}^{N}\omega_i}\sum_{i=1}^{N}\omega_i\left|[(1-\delta_i)\cdot e_{\text{T-pseudo-obs},\tau}(t_i,\mathcal{D}) + \delta_i\cdot t_i] - \hat{t}_i\right|,$$

where the prediction time ($\hat{t}_i$) is the **truncated median time** ($\hat{t}_i = \hat{t}_{i,\text{T-median},\tau}$) of the ISD defined in Equation B2, and we use the **truncated best guess** ($e_{\text{T-pseudo-obs},\tau}(t_i, \mathcal{D})$) defined in Equation E8.

6. Truncated-Log MAE-PO:

The truncated-log (TL) adaptation of MAE-PO is:

$$\mathbb{E}_{i\sim\mathcal{D}}[\mathcal{R}_{\text{TL-MAE-PO},\tau}(\hat{t}_i,\, t_i,\, \delta_i)] = \tag{E10}$$
$$\frac{1}{\sum_{i=1}^{N}\omega_i}\sum_{i=1}^{N}\omega_i\left|[(1-\delta_i)\cdot \log(e_{\text{T-pseudo-obs},\tau}(t_i,\mathcal{D})) + \delta_i\cdot \log(t_i)] - \log(\hat{t}_i)\right|,$$

where again the prediction time ($\hat{t}_i$) is the **truncated median time** ($\hat{t}_i = \hat{t}_{i,\text{T-median},\tau}$) of the ISD defined in Equation B2, and we use the **truncated best guess** ($e_{\text{T-pseudo-obs},\tau}(t_i, \mathcal{D})$) defined in Equation E8. For Equation E10, if the predicted time to event or the ground truth is zero, we initially add a small value (ϵ) to prevent the logarithm function from yielding minus infinity. Moreover, we choose to use log base e.

To further understand how we can interoperate error measured by TL-MAE-PO, recall that the TL-MAE-PO for AFT on the Prostate #1 dataset is 0.62 ± 0.001. Here, given that $\exp(0.62) = 1.86$, this is claiming that we expect each prediction to be within a multiplicative factor of 1.86 of the correct value. So, for instance, if we predict patient A will live 9.02 months, we are saying that we anticipate that patient A will live between $(9.02/1.86,\ 9.02 \times 1.86) = (4.82,\ 31.19)$ months. If another patient was predicted to live 9.02 days, then we would anticipate that person would live between $(4.82,\ 31.19)$ days. This is the nature of multiplicative bounds.

Supplementary F Model Implementation Details

In this section, we included details of the model implementation that was used in this paper.

- **Kaplan Meier (KM)** is a popular estimator that uses the information of a group of patients. The KM curve provides a stepwise estimate of the probability of event occurrence. We used `KaplanMeierFitter` class from `lifelines` library, and we used the median time of the training population as the time to event prediction for the test set.
- **Random Survival Forest (RSF):** is an extension of the Random Forest algorithm for time-to-event data, offering a non-parametric approach to model survival outcomes. RSF is an ensemble of survival trees, each learned on a bootstrapped version of the training dataset. We used `RandomSurvivalForest` class from `sksurv.ensemble` library for implementation, with 150 trees, min samples split of 25, and min samples leaf of 20.
- **Multi-Task Logistic Regression (MTLR)** is a machine learning approach designed for survival analysis and gives individualized curve prediction. The model is implemented using `MTLR` class from `torchmtlr` package, with a learning rate of 0.001, batch size of 512, and 500 epochs.
- **Deep MTLR** is another method that predicts individualized survival distribution and uses the MTLR models as its base and a deep learning model as its core. We implemented D-MTLR using `DeepMTLR` class from `torchmtlr` package, with the same configuration as MTLR. The architecture of the model is provided in the code base, and it includes layers of NN nodes, dropout of 0.4, and Exponential Linear Units (ELUs).
- **DeepHit** learns the individualized survival distribution using deep learning. The model is implemented using `DeepHitSingle` class from `pycox.models` package. We used Adam optimizer with early stopping.
- **Cox Proportional Hazard (Cox-PH)** is a semi-parametric method used in survival analysis to assess the impact of several risk factors on survival time. It provides hazard ratios, indicating the relative risk of event occurrence given a change in predictor variables. It is composed of a baseline hazard function at the population level (non-parametric) and a parametric partial hazard function. We implemented Cox-PH using `CoxPHSurvivalAnalysis` class from `sksurv.linear_model` library.
- **Accelerate Failure time (AFT)** is a parametric survival analysis technique that directly models the time to event and provides individualized prediction. We employed AFT with Weibull parametric assumption. For implementation, we employed the `WeibullAFTFitter` class from `lifelines` library. We used median time for the time-to-event prediction, and based on our experiments, it works better than using the average time.

Supplementary G Results in Details

Tables G2, G3, G4, G5, G6, G7, G8, G9, G10, and G11 show the evaluation of various models by each of the discussed metrics for the selected cancer types. For all

the tables, the reported C-index and BS are computed at the median time, except for table G7, in which we computed the C-index and BS at the 10-year time point since we wanted to compare our results with the reported results of Survival Quilts model [3]. In terms of TL-MAE-PO and T-MAE-PO, our results show that RSF followed by Deep-MTLR are the top-performing methods in all the datasets except for the prostate # 1 dataset where AFT is the best.

Table G2: Model comparison using all the discussed metrics for **brain** cancer.

Metric	RSF	MTLR	Deep-MTLR	DeepHit	Cox-PH	AFT	KM (baseline)
C-index	**74.91 ± 0.35**	72.46 ± 0.39	74.68 ± 0.39	73.45 ± 0.28	72.63 ± 0.41	72.64 ± 0.43	50.00 ± 0.00
BS(median)	**0.16 ± 0.00**	0.18 ± 0.00	**0.16 ± 0.00**	0.20 ± 0.00	0.17 ± 0.00	0.17 ± 0.00	0.22 ± 0.00
T-MAE-PO	**46.12 ± 0.53**	51.97 ± 0.82	47.99 ± 0.80	58.49 ± 0.26	48.99 ± 0.51	49.30 ± 0.57	156.56 ± 0.77
TL-MAE-PO	**1.42 ± 0.01**	1.52 ± 0.02	1.44 ± 0.01	1.70 ± 0.00	1.48 ± 0.01	1.49 ± 0.01	2.75 ± 0.00

The numbers represent a 10-fold cross-validation result with 95% CI. Bold numbers indicate superior performance compared to other models.

Table G3: Model comparison using all the discussed metrics for **breast** cancer.

Metric	RSF	MTLR	Deep-MTLR	DeepHit	Cox-PH	AFT	KM (baseline)
C-index	**82.66 ± 0.08**	77.23 ± 0.53	80.71 ± 0.15	80.76 ± 0.22	76.86 ± 0.09	77.03 ± 0.09	50.00 ± 0.00
BS(median)	**0.06 ± 0.00**	0.07 ± 0.00	**0.06 ± 0.00**	0.07 ± 0.00	0.07 ± 0.00	0.07 ± 0.00	0.08 ± 0.00
T-MAE-PO	**65.66 ± 0.13**	69.76 ± 3.17	65.74 ± 1.87	85.09 ± 0.63	72.46 ± 0.16	72.71 ± 0.15	99.15 ± 0.01
TL-MAE-PO	**0.93 ± 0.01**	0.97 ± 0.03	0.94 ± 0.02	1.14 ± 0.00	1.00 ± 0.00	1.00 ± 0.00	1.22 ± 0.00

The numbers represent a 10-fold cross-validation result with 95% CI. Bold numbers indicate superior performance compared to other models.

Table G4: Model comparison using all the discussed metrics for **kidney and renal pelvis** cancer.

Metric	RSF	MTLR	Deep-MTLR	DeepHit	Cox-PH	AFT	KM (baseline)
C-index	**86.26 ± 0.06**	79.65 ± 0.24	86.22 ± 0.11	84.59 ± 0.07	79.94 ± 0.14	80.03 ± 0.14	50.00 ± 0.00
BS(median)	**0.07 ± 0.00**	0.09 ± 0.00	**0.07 ± 0.00**	0.10 ± 0.00	0.09 ± 0.00	0.09 ± 0.00	0.12 ± 0.00
T-MAE-PO	51.94 ± 1.26	66.59 ± 2.13	**51.59 ± 1.53**	78.70 ± 2.36	68.56 ± 1.18	69.55 ± 1.2	117.74 ± 1.47
TL-MAE-PO	**1.20 ± 0.02**	1.43 ± 0.04	1.21 ± 0.03	1.71 ± 0.03	1.45 ± 0.02	1.46 ± 0.02	2.00 ± 0.02

The numbers represent a 10-fold cross-validation result with 95% CI. Bold numbers indicate superior performance compared to other models.

Table G5: Model comparison using all the discussed metrics for **liver** cancer.

Metric	RSF	MTLR	Deep-MTLR	DeepHit	Cox-PH	AFT	KM (baseline)
C-index	74.50 ± 0.21	70.30 ± 0.26	**74.56 ± 0.18**	73.41 ± 0.22	70.35 ± 0.25	70.30 ± 0.25	50.00 ± 0.00
BS(median)	**0.17 ± 0.00**	0.19 ± 0.00	**0.17 ± 0.00**	0.22 ± 0.00	0.19 ± 0.00	0.19 ± 0.00	0.23 ± 0.00
T-MAE-PO	**39.98 ± 0.95**	42.99 ± 3.36	40.86 ± 3.14	45.99 ± 4.50	44.10 ± 1.49	46.37 ± 1.34	174.60 ± 5.54
TL-MAE-PO	**1.93 ± 0.05**	2.07 ± 0.03	1.95 ± 0.04	2.24 ± 0.01	2.07 ± 0.05	2.11 ± 0.05	3.62 ± 0.11

The numbers represent a 10-fold cross-validation result with 95% CI. Bold numbers indicate superior performance compared to other models.

Table G6: Model comparison using all the discussed metrics for **lung and bronchus** cancer.

Metric	RSF	MTLR	Deep-MTLR	DeepHit	Cox-PH	AFT	KM (baseline)
C-index	**73.02 ± 0.07**	68.88 ± 0.23	72.20 ± 0.11	71.72 ± 0.12	68.96 ± 0.09	69.06 ± 0.09	50.00 ± 0.00
BS(median)	**0.18 ± 0.00**	0.20 ± 0.00	**0.18 ± 0.00**	0.21 ± 0.00	0.20 ± 0.00	0.20 ± 0.00	0.23 ± 0.00
T-MAE-PO	**42.81 ± 0.50**	47.93 ± 1.55	43.62 ± 1.44	51.08 ± 1.71	47.59 ± 0.78	47.43 ± 0.99	164.66 ± 1.49
TL-MAE-PO	**1.44 ± 0.00**	1.57 ± 0.02	1.47 ± 0.01	1.66 ± 0.01	1.55 ± 0.01	1.55 ± 0.01	2.81 ± 0.01

The numbers represent a 10-fold cross-validation result with 95% CI. Bold numbers indicate superior performance compared to other models.

Table G7: Model comparison using all the discussed metrics for **prostate #1** cancer.

Metric	RSF	MTLR	Deep-MTLR	DeepHit	Cox-PH	AFT	Survival Quilts	KM (baseline)
C-index	85.59 ± 0.71	83.11 ± 1.27	**86.05 ± 0.71**	75.57 ± 2.86	85.14 ± 0.74	85.31 ± 0.75	82.90 ± 0.09	50.00 ± 0.00
BS (10-years)	0.03 ± 0.00	0.03 ± 0.00	0.03 ± 0.00	0.03 ± 0.00	0.03 ± 0.00	0.03 ± 0.00	0.03 ± 0.00	0.04 ± 0.00
T-MAE-PO	46.52 ± 0.30	44.65 ± 2.54	45.74 ± 2.20	47.39 ± 0.89	45.05 ± 0.25	**44.69 ± 0.32**	-	49.37 ± 0.22
TL-MAE-PO	0.64 ± 0.00	0.62 ± 0.02	0.63 ± 0.02	0.65 ± 0.01	0.63 ± 0.00	**0.62 ± 0.00**	-	0.67 ± 0.00

The numbers represent a 10-fold cross-validation result with 95% CI. Bold numbers indicate superior performance compared to other models.

Table G8: Model comparison using all the discussed metrics for **prostate #2** cancer.

Metric	RSF	MTLR	Deep-MTLR	DeepHit	Cox-PH	AFT	KM (baseline)
C-index	**88.31 ± 0.10**	75.39 ± 5.93	84.5 ± 0.40	85.94 ± 1.10	83.60 ± 0.08	84.15 ± 0.09	50.00 ± 0.00
BS(median)	**0.04 ± 0.00**	0.10 ± 0.06	0.05 ± 0.00	0.05 ± 0.00	0.05 ± 0.00	0.05 ± 0.00	0.06 ± 0.00
T-MAE-PO	**56.01 ± 1.17**	81.89 ± 6.31	65.08 ± 3.52	80.36 ± 3.52	62.76 ± 1.26	64.15 ± 1.30	94.68 ± 1.49
TL-MAE-PO	**0.81 ± 0.01**	1.11 ± 0.17	0.89 ± 0.04	1.06 ± 0.02	0.88 ± 0.01	0.89 ± 0.00	1.14 ± 0.01

The numbers represent a 10-fold cross-validation result with 95% CI. Bold numbers indicate superior performance compared to other models.

Table G9: Model comparison using all the discussed metrics for **stomach** cancer.

Metric	RSF	MTLR	Deep-MTLR	DeepHit	Cox-PH	AFT	KM (baseline)
C-index	**77.00 ± 0.14**	72.76 ± 0.33	76.68 ± 0.13	75.25 ± 0.15	71.80 ± 0.21	71.87 ± 0.20	50.00 ± 0.00
BS(median)	**0.16 ± 0.00**	0.18 ± 0.00	**0.16 ± 0.00**	0.20 ± 0.00	0.18 ± 0.00	0.18 ± 0.00	0.22 ± 0.00
T-MAE-PO	**40.21 ± 0.52**	50.41 ± 0.93	49.26 ± 2.22	54.76 ± 2.40	51.07 ± 0.69	52.13 ± 1.31	159.69 ± 2.36
TL-MAE-PO	**1.53 ± 0.00**	1.72 ± 0.02	1.61 ± 0.01	1.85 ± 0.02	1.74 ± 0.01	1.78 ± 0.01	3.03 ± 0.02

The numbers represent a 10-fold cross-validation result with 95% CI. Bold numbers indicate superior performance compared to other models.

Table G10: Model comparison using all the discussed metrics for **thyroid** cancer.

Metric	RSF	MTLR	Deep-MTLR	DeepHit	Cox-PH	AFT	KM (baseline)
C-index	93.22 ± 0.22	89.38 ± 0.34	**93.30 ± 0.34**	92.22 ± 0.31	90.27 ± 0.20	90.20 ± 0.20	50.00 ± 0.00
BS(median)	**0.01 ± 0.00**	0.02 ± 0.00	**0.01 ± 0.00**	0.02 ± 0.00	0.02 ± 0.00	0.02 ± 0.00	0.02 ± 0.00
T-MAE-PO	**56.42 ± 0.34**	65.71 ± 3.70	57.35 ± 3.31	78.16 ± 3.16	72.01 ± 0.72	74.94 ± 0.60	98.00 ± 0.13
TL-MAE-PO	**1.29 ± 0.01**	1.43 ± 0.06	1.33 ± 0.06	1.66 ± 0.03	1.52 ± 0.01	1.55 ± 0.01	1.81 ± 0.00

The numbers represent a 10-fold cross-validation result with 95% CI. Bold numbers indicate superior performance compared to other models.

Table G11: Model comparison using all the discussed metrics for **urinary bladder** cancer.

Metric	RSF	MTLR	Deep-MTLR	DeepHit	Cox-PH	AFT	KM (baseline)
C-index	**81.73 ± 0.18**	74.41 ± 0.68	81.34 ± 0.27	80.94 ± 0.20	74.64 ± 0.19	74.73 ± 0.18	50.00 ± 0.00
BS(median)	**0.08 ± 0.00**	0.12 ± 0.01	0.08 ± 0.00	0.10 ± 0.00	0.10 ± 0.00	0.10 ± 0.00	0.12 ± 0.00
T-MAE-PO	64.55 ± 1.69	80.42 ± 10.29	**62.40 ± 2.95**	97.11 ± 2.92	79.88 ± 1.03	80.21 ± 1.05	117.51 ± 2.83
TL-MAE-PO	1.23 ± 0.01	1.44 ± 0.11	**1.22 ± 0.04**	1.68 ± 0.02	1.46 ± 0.01	1.46 ± 0.01	1.82 ± 0.02

The numbers represent a 10-fold cross-validation result with 95% CI. Bold numbers indicate superior performance compared to other models.