CURE-OOD: Benchmarking Out-of-Distribution Detection for Survival Prediction

Anonymous authors
Paper under double-blind review

Abstract

"How long can I live and remain free of cancer?" is often the first question a patient asks after receiving a cancer diagnosis and treatment. Accurate survival prediction helps alleviate psychological distress and supports risk stratification and personalized treatment planning. Recent survival prediction frameworks have shown strong performance using computed tomography (CT) images. However, variations in imaging acquisition introduce out-of-distribution (OOD) samples caused by covariate shifts that undermine model reliability. Despite this challenge, to our knowledge, no existing benchmark systematically studies OOD detection in cancer survival prediction. To address this gap, we introduce the Cancer sURvival bEnchmark for OOD Detection (CURE-OOD), the first benchmark for systematically evaluating OOD detection in survival prediction under controlled acquisition-induced distribution shifts. CURE-OOD defines scanner-parameter-based training, in-distribution (ID), and OOD test splits across four survival prediction tasks. Our experiments show that covariate shifts notably reduce survival prediction performance. It also shows that mainstream classification-oriented OOD detectors can fail in survival prediction. Finally, we include HazardDev as a simple survival-aware reference baseline for OOD detection. CURE-OOD enables systematic analysis of how distribution shifts affect both downstream survival performance and OOD detectability.

1 Introduction

Accurate survival prediction is critical in cancer prognosis, as it helps clinicians assess patient risk and make informed, personalized treatment decisions. In head and neck cancer (HNC), radiation therapy is typically guided by stage-dependent risk stratification protocols, where patients within the same stage receive identical dose prescriptions (Pan et al., 2016; Caudell et al., 2017). Yet patients with similar staging or tumor characteristics can experience markedly different outcomes, ranging from long-term recurrence-free survival to early disease progression or treatment-related toxicity (Ang et al., 2010; Kawecki & Krajewski, 2014). This heterogeneity motivates reliable, individualized survival prediction models that complement traditional staging systems and guide treatment planning (Beesley et al., 2019; Kang et al., 2015; Amin et al., 2017).

Recent advances in deep learning have enabled survival modeling from medical images. Typically, CNNs or Vision Transformers (ViTs) extract imaging features that are passed to survival prediction frameworks (Chen et al., 2024b; Saeed et al., 2024; Chen et al., 2024a). A representative formulation is multi-task logistic regression (MTLR), which reformulates survival analysis into ordered binary subtasks across discrete time intervals to jointly learn a survival distribution (Zhang & Yang, 2021; Yu et al., 2011; Kvamme et al., 2019; Nagpal et al., 2021). In real-world practice, CT scans are often collected over long periods and across hospitals or imaging centers. During this process, scanner hardware and acquisition protocols may be updated or replaced (Welch et al., 2023), while cross-center variation further introduces differences in scanner models and acquisition settings (Guan & Liu, 2021), as illustrated in Fig. 1. This variability changes image appearance and induces covariate shifts while the clinical prediction task remains fixed. In image classification, segmentation, and related medical imaging tasks, such shifts have been shown to reduce model reliability (Gutbrod et al., 2025; Liu et al., 2025; Baek et al., 2024; Zhao et al., 2025; 2026). However,

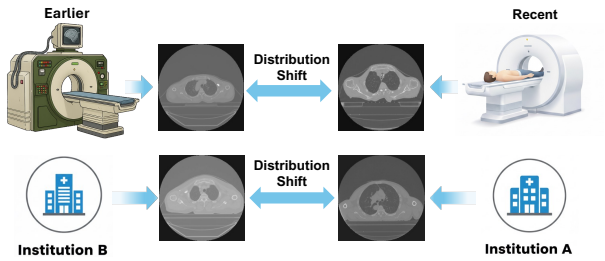


Figure 1: CT data distribution shifts caused by cross-institution variability and scanner changes over time.

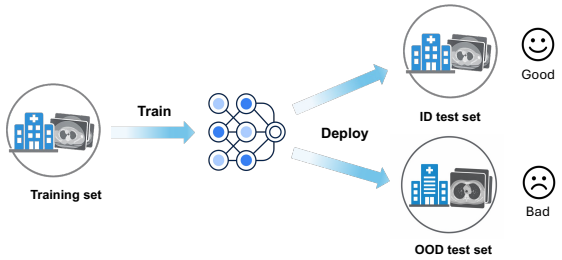


Figure 2: A model trained on ID data performs well on ID test data but degrades on OOD test data.

for cancer survival prediction, whether acquisition-induced covariate shifts cause significant downstream performance degradation has not been systematically studied.

Existing OOD benchmarks and detection methods mainly target classification, where detectors often rely on class-confidence or logit-based signals. Survival prediction differs from classification because its outputs are time-indexed risk or survival distributions rather than class confidence. Therefore, whether classification-oriented OOD detectors remain effective for survival prediction under acquisition-induced shifts has not been systematically studied.

To address these gaps, we introduce CURE-OOD, a benchmark for OOD detection in cancer survival prediction. Table 1 summarizes key differences between CURE-OOD and existing OOD benchmarks. Following standard OOD terminology, in-distribution (ID) data are drawn from the same distribution as the training set, whereas OOD data deviate from that distribution. We build CURE-OOD as a controlled evaluation protocol with non-overlapping training, in-distribution (ID) test, and OOD test splits. These splits are defined by four acquisition parameters, including pixel spacing, exposure time, slice thickness, and X-ray tube current. This protocol enables systematic and reproducible evaluation of two questions. First, whether acquisition-induced covariate shifts degrade survival prediction. Second, whether existing OOD detection methods can reliably identify shifted samples in survival modeling. On CURE-OOD, we find that these shifts reduce survival prediction performance and that several classification-oriented OOD detectors fail under survival formulations, sometimes even assigning more OOD-like scores to ID samples. We further include HazardDev as a simple survival-aware reference baseline that uses predicted hazard as the detection signal.

Our main contributions are summarized as follows:

- We present CURE-OOD, to our knowledge, the first benchmark for out-of-distribution detection in cancer survival prediction built on real-world clinical CT data, extending medical imaging OOD evaluation beyond classification and segmentation.
- We systematically verify the degradation of survival prediction performance under acquisition-induced covariate shifts and benchmark representative OOD detectors in this setting.
- We introduce HazardDev, a simple survival-aware OOD detection baseline based on predicted hazard, and analyze the limitations of existing classification-oriented detectors for survival prediction.

CURE-OOD facilitates more rigorous evaluation of OOD detection in survival prediction and supports the development of survival-aware methods under acquisition-induced shifts.

2 Related work

Survival prediction in medical imaging. Modern survival modeling has evolved from classical statistics to deep learning. Early work relied on the Cox proportional hazards model (Cox, 1972) and Aalen regression models (Aalen, 1989), which are powerful for statistical inference but limited for individualized prediction and rely on the proportional hazards assumption. DeepSurv (Katzman et al., 2018) and other "deep-Cox"

Table 1: Comparison with existing OOD benchmarks. While most existing benchmarks focus on natural image classification under semantic or synthetic shifts, CURE-OOD addresses acquisition-induced covariate shifts in clinical imaging for survival prediction.

Benchmark	**Med.**	**Primary Task**	**Shift Type**	**Source**
OpenOOD (Yang et al., 2022)	✗	Classification	Semantic	Mixed
OpenMIBOOD (Gutbrod et al., 2025)	✓	Classif. / Segmen.	Mixed	Real
ImageNet-ES (Baek et al., 2024)	✗	Classification	Covariate	Real
ImageNet-C/ CIFAR-C (Hendrycks & Dietterich, 2019)	✗	Classification	Covariate	Synthetic
ImageNet-A/-O/-R (Hendrycks et al., 2021)	✗	Classification	Semantic	Real
WILDS (Koh et al., 2021)	Part.	Classif. / Regress.	Covariate	Real
MIDOG (Aubreville et al., 2023)	✓	Detection	Covariate	Real
PhaKIR (Rueckert, 2024)	✓	Detect. / Classif.	Covariate	Real
OASIS-3 (LaMontagne et al., 2019)	✓	Classification	Covariate	Real
MedMNIST v2 (Yang et al., 2023)	✓	Classification	Mixed	Mixed
BREEDS (Santurkar et al., 2020)	✗	Classification	Subpopulation	Real
CURE-OOD	✓	Risk modeling	Covariate	Real

approaches (Martinussen & Scheike, 2006; Ching et al., 2018; Nagpal et al., 2021; Kvamme et al., 2019) extend these hazard-based formulations while using neural networks for representation learning. In contrast, MTLR (Yu et al., 2011) discretizes time and learns a sequence of dependent logistic classifiers, directly modeling the survival function and naturally handling censored data. This reformulation makes survival prediction operationally similar to classification across time intervals. In current imaging pipelines, MTLR is commonly combined with visual backbones such as CNNs or ViTs to extract features and predict patient-specific risk or hazard over time (Wiegrebe et al., 2024). It has also been adopted in recent multilabel survival prediction systems (Kim et al., 2021).

Covariate shift in medical datasets. Medical imaging datasets are often collected over many years, especially for survival prediction tasks requiring decade-long follow-up (Welch et al., 2024; Henschke et al., 2023; Littlejohns et al., 2020). During this span, scanner hardware, protocols and reconstruction algorithms are frequently upgraded or replaced to meet evolving clinical and technical standards (Fortin et al., 2018). In addition, multi-center studies introduce further variability across institutions and device vendors, as identical scanners may operate under different acquisition conditions (Mackin et al., 2015). These factors yield heterogeneous acquisition settings that alter image appearance and quantitative features, leading to covariate shift and degraded downstream performance (Zech et al., 2018). Unlike many computer vision applications, medical imaging demands much higher reliability and safety, where incorrect predictions can directly impact clinical decisions and patient outcomes. Therefore, rather than solely adapting models to unseen domains, it is crucial to detect OOD samples that arise from covariate shifts (Cao et al., 2020; Gutbrod et al., 2025), allowing clinicians to recognize potentially unreliable predictions before they affect patient care.

OOD detection. OOD detection aims to identify inputs that deviate from the training distribution and has been widely studied in computer vision. Most existing methods are designed for image classification datasets such as CIFAR and ImageNet, relying on confidence- or representation-based scores, including maximum softmax probability (MSP) (Hendrycks & Gimpel, 2016), ODIN (Liang et al., 2017), energy-based scoring (Liu et al., 2020) and activation shaping (ASH) (Djurisic et al., 2022). Some studies further extend OOD detection to segmentation tasks, focusing on pixel-level anomaly localization (Zhao et al., 2025; 2024; Vojíř et al., 2024; Shoeb et al., 2024). In the medical domain, Cao et al. (Cao et al., 2020) provided one of the earliest systematic evaluations of OOD detection across several medical imaging modalities. More recently, Gutbrod et al. introduced OpenMIBOOD (Gutbrod et al., 2025), a large-scale benchmark inspired by OpenOOD (Yang et al., 2022) that comprehensively assesses OOD detection under medical image classification settings. Despite considerable progress, OOD detection has been explored mainly in classification and segmentation tasks. To our knowledge, survival prediction, with its distinct time-to-event objectives and censoring-aware formulations, has not yet been systematically studied.

3 Background

3.1 Survival Analysis

Survival prediction aims to estimate the likelihood of a clinical event, such as cancer recurrence or death, occurring over time. Unlike standard classification tasks that predict discrete labels, survival models estimate a time-dependent probability distribution describing event risk across time intervals. In medical imaging, such models leverage features extracted from CT scans to capture patterns associated with disease progression. Among existing formulations, multi-task logistic regression (MTLR) (Yu et al., 2011) is particularly suitable for censored survival prediction because it represents the task as a sequence of dependent binary classification tasks across discretized time intervals.

Let $\{t_1, \ldots, t_m\}$ denote m time intervals and $\mathbf{x}$ the features extracted by the visual backbone. MTLR parameterizes the survival probability at each interval as $P_{\theta_i}(T \geq t_i \mid \mathbf{x}) = (1 + \exp(\boldsymbol{\theta}_i^\top \mathbf{x} + b_i))^{-1}$, and enforces temporal consistency through the joint likelihood

$$P_\Theta(Y = y \mid \mathbf{x}) = \frac{\exp\big(\sum_{i=1}^{m} y_i(\boldsymbol{\theta}_i^\top \mathbf{x} + b_i)\big)}{\sum_{k=0}^{m} \exp(f_\Theta(\mathbf{x}, k))}, \quad f_\Theta(\mathbf{x}, k) = \sum_{i=k+1}^{m} (\boldsymbol{\theta}_i^\top \mathbf{x} + b_i). \tag{1}$$

where the resulting logits $f_\Theta(\mathbf{x}, k)$ summarize interval-wise risk over time. Model parameters are learned by minimizing the negative log-likelihood of this distribution.

From the predicted distribution, the event probability is $p_\Theta(\mathbf{x}, t_i) = P_\Theta(T = t_i \mid \mathbf{x})$, the survival probability is $G_\Theta(\mathbf{x}, t_i) = P_\Theta(T \geq t_i \mid \mathbf{x})$, and the discrete hazard is

$$h(t_i \mid \mathbf{x}) = \frac{P_\Theta(T = t_i \mid \mathbf{x})}{P_\Theta(T \geq t_i \mid \mathbf{x})} = \frac{p_\Theta(\mathbf{x}, t_i)}{G_\Theta(\mathbf{x}, t_i)}. \tag{2}$$

3.2 Medical Image Acquisition for Survival Analysis

CT appearance is determined by both acquisition and reconstruction settings. Acquisition parameters directly affect image resolution, contrast, and noise characteristics. Variations in these settings, caused by scanner differences or protocol changes across sites and time, lead to systematic changes in image appearance, introducing covariate shifts. Therefore, we can leverage these parameters to define structured distribution shifts. We leverage this property to construct controlled ID and OOD splits, as detailed in Sec. 4.2.

4 CURE-OOD Benchmark

4.1 Data and Preparation

To construct CURE-OOD with controlled acquisition-induced shifts, we require a CT cohort that (i) provides scanner metadata for partitioning and (ii) supports standardized survival endpoints. We therefore construct CURE-OOD by curating and partitioning the RADCURE cohort (Welch et al., 2024), which provides head and neck cancer (HNC) radiotherapy planning CT scans collected between 2005 and 2017. Acquisition parameters are recorded in the DICOM metadata, including KVP, pixel spacing, exposure time, slice thickness, and X-ray tube current. Since all scans were acquired with the same KVP setting, we exclude KVP from the partitioning design. After filtering cases with incomplete CT volumes or missing primary GTVp masks, we retain 2,366 patients for benchmark construction.

Using this filtered cohort, we consider four survival prediction tasks: overall survival (OS), local failure-free survival (LFFS), regional failure-free survival (RFFS), and distant failure-free survival (DFFS). The corresponding numbers of events are 792, 300, 134, and 293, respectively, and we treat each task independently during training and evaluation.

Table 2: Dataset partitioning based on acquisition parameters for CURE-OOD.

Parameter	Train	ID	OOD	ID Range	OOD Range
Pixel spacing (mm)	1293	100	100	0.976–1.172	0.805–0.969
Exposure time (ms)	1288	100	100	1000	1831–2289
Slice thickness (mm)	1200	100	100	2	2.5
X-ray tube current	1541	100	100	300	305–490

4.2 Key Acquisition Parameters

We focus on four CT acquisition parameters that directly influence image appearance and feature distributions: pixel spacing, exposure time, slice thickness, and X-ray tube current. Together, they control effective resolution and noise, leading to realistic acquisition-induced shifts across sites and time (Fig. 3).

Pixel Spacing sets the spatial sampling resolution. Smaller spacing preserves finer anatomical details but can amplify noise, while larger spacing smooths textures and yields a wider field of view.

Exposure Time controls photon accumulation at the detector. Longer exposure generally improves signal-to-noise ratio and contrast, whereas shorter exposure increases noise and may obscure subtle structures.

Slice Thickness determines the degree of through-plane averaging. Thicker slices reduce random noise but blur edges and fine structures, while thinner slices preserve sharper boundaries at the cost of higher noise.

X-ray Tube Current reflects the X-ray intensity. Higher current typically produces cleaner, higher-contrast scans, whereas lower current leads to dimmer and noisier images.

These parameters are recorded in the RADCURE metadata and provide interpretable, reproducible axes for defining clinically realistic distribution shifts.

4.3 Data Partitioning Strategy

For each acquisition parameter, we first sort scans by the recorded value and group them according to distinct settings. We then construct an ID/OOD split according to the natural distribution of the target acquisition parameter, assigning the larger group to the ID domain to ensure sufficient training data and the smaller group to the OOD domain. For instance, for slice thickness, 2 mm scans (1,614) form the ID set and 2.5 mm scans (752) form the OOD set.

Fig. 4 visualizes the resulting separation for pixel spacing. Additional distributions for the other acquisition parameters are provided in Fig. 9 in the appendix. The ID/OOD assignment rules are summarized in Table 2. For each task, samples without the corresponding acquisition metadata are excluded before partitioning. We randomly sample 100 cases from each group to form the ID and OOD test sets. The same partitioning strategy is applied to all four acquisition parameters, and the training set size ranges from approximately 1,200 to 1,600 cases across different splits.

4.4 A Survival-Aware Reference Baseline

Because MTLR does not produce class confidence scores, classification-oriented OOD scores are not naturally aligned with its outputs. We therefore include **HazardDev** as a simple survival-aware reference baseline defined on a survival-native quantity. We use hazard as the reference signal because it is directly derived from the predicted survival distribution and provides an interpretable summary of how event risk evolves over time. HazardDev then assigns an OOD score by measuring how much a sample's predicted hazard curve deviates from the average hazard curve on the training distribution.

Training-phase statistics. Given a trained MTLR model and the discrete-time hazard definition in Sec. 3.1, we first compute the hazard function $h_n(t_i)$ for each training sample n at each discretized time bin t_i. We then calculate the average hazard over all N training samples to obtain the expected hazard profile

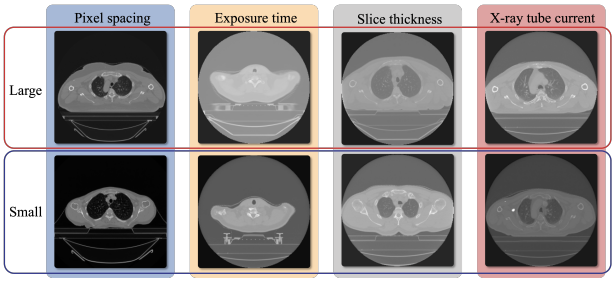


Figure 3: Examples showing how acquisition parameters affect CT image appearance. Larger pixel spacing covers a wider field of view with decreased resolution, longer exposure time and higher tube current reduce noise, and thicker slice thickness produces smoother but blurrier images.

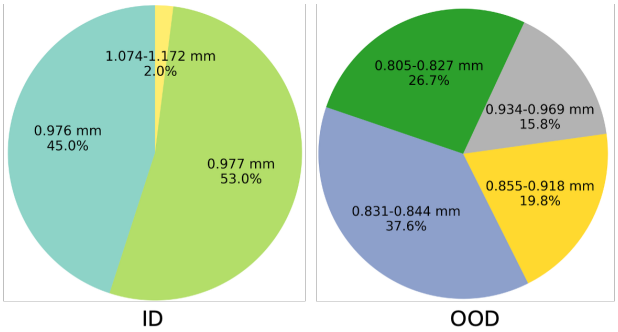


Figure 4: Overall distribution of pixel spacing values across the RADCURE dataset. The variability across parameter ranges enables a clear distinction between ID and OOD domains used in CURE-OOD.

of the training distribution:

$$\bar{h}(t_i) = \frac{1}{N} \sum_{n=1}^{N} h_n(t_i). \tag{3}$$

Testing-phase deviation. For a test sample with hazard prediction $h_{\text{test}}(t_i)$, we measure its deviation from the training distribution by computing the difference at each time bin and aggregating across all bins:

$$\text{HazardDev}(\mathbf{x}) = \sum_{i=1}^{m} h_{\text{test}}(t_i) - \bar{h}(t_i). \tag{4}$$

A larger HazardDev value indicates that the sample's hazard pattern differs substantially from the training population, suggesting a higher likelihood of being out-of-distribution.

5 Experiments

Evaluation metrics. We report metrics for both survival prediction and OOD detection. For survival prediction, we use the concordance index (C-index), where higher values indicate better agreement between predicted risk and observed time-to-event. For OOD detection, we use AUROC and AUPRC, measuring overall separability.

Survival prediction training. To evaluate OOD detection under different distribution shifts, we train separate MTLR (Yu et al., 2011) survival models for each acquisition parameter–outcome pair. We consider four acquisition parameters and four outcomes (OS, LFFS, RFFS, DFFS), resulting in 16 models. Each model is trained on its corresponding ID training set and evaluated on both the matched ID test set and the paired OOD test set, reflecting deployment where acquisition conditions may change at inference time.

Experimental setup. We use a ViT-based 3D model with a UNETR (Hatamizadeh et al., 2022) encoder to extract volumetric features from CT, followed by two fully connected layers and an MTLR head that outputs survival distributions over discretized time. CT volumes are resampled to a common resolution and cropped around the primary tumor; the timeline is discretized into 8 intervals. We train with AdamW (weight decay 0.01, batch size 32) for up to 400 epochs with early stopping on validation loss; learning rates are 1×10^{-3} (OS/LFFS), 8×10^{-4} (RFFS), and 3×10^{-4} (DFFS). Augmentation includes random 3D rotation and shifts. All experiments use PyTorch on NVIDIA RTX A5000 GPUs.

Table 3: OOD detection performance across survival tasks measured by AUROC and AUPRC. Average results underscore HazardDev's robustness under acquisition shifts.

	Method	Pixel Spacing		Exposure Time		Slice Thick.		Tube Current		**Average**	
		AUROC	AUPRC	AUROC	AUPRC	AUROC	AUPRC	AUROC	AUPRC	AUROC	AUPRC
OS	ASH	0.3910	0.4505	0.3815	0.4354	0.4670	0.4783	0.4381	0.4795	0.4194	0.4609
	Dropout	0.5353	0.5040	0.4040	0.4458	0.4778	0.4847	0.4711	0.4851	0.4721	0.4799
	Energy	0.3938	0.4534	0.3807	0.4300	0.4594	0.4779	0.4201	0.4745	0.4135	0.4589
	GEN	0.5411	0.5101	0.4223	0.4441	0.4843	0.4909	0.5074	0.5105	0.4888	0.4889
	Dice	0.4200	0.4631	0.4001	0.4370	0.4615	0.4789	0.4329	0.4763	0.4286	0.4638
	MLS	0.4073	0.4586	0.3980	0.4366	0.4580	0.4780	0.4207	0.4728	0.4210	0.4615
	MSP	0.5472	0.5121	0.4279	0.4466	0.4995	0.4983	0.5120	0.5098	0.4967	0.4917
	ODIN	0.5169	0.4980	0.4070	0.4320	0.4784	0.4819	0.4912	0.5043	0.4734	0.4790
	SCALE	0.3994	0.4544	0.3842	0.4384	0.4635	0.4801	0.4294	0.4798	0.4191	0.4632
	HazardDev	**0.6043**	**0.5726**	**0.6237**	**0.6098**	**0.5416**	**0.5455**	**0.5797**	**0.5465**	**0.5873**	**0.5686**
LFFS	ASH	0.4056	0.4480	0.4284	0.4736	0.4458	0.4580	0.4686	0.5104	0.4371	0.4725
	Dropout	0.3931	0.4378	0.4433	0.4932	0.4566	0.4659	0.5014	0.5226	0.4486	0.4799
	Energy	0.3942	0.4400	0.3578	0.4028	0.4661	0.4658	0.4521	0.5044	0.4175	0.4532
	GEN	0.4551	0.4634	0.6333	0.5950	0.4490	0.4837	0.5483	**0.5506**	0.5214	0.5232
	Dice	0.3991	0.4424	0.4579	0.4951	0.4461	0.4626	0.4575	0.5081	0.4401	0.4771
	MLS	0.3953	0.4406	0.4319	0.4790	0.4607	0.4684	0.4642	0.5115	0.4380	0.4749
	MSP	0.4376	0.4550	0.6242	0.5891	0.4447	0.4688	0.5192	0.5348	0.5064	0.5119
	ODIN	0.4397	0.4548	0.6097	0.5708	0.4638	0.4736	0.4754	0.5100	0.4971	0.5023
	SCALE	0.3966	0.4410	0.4160	0.4655	0.4607	0.4679	0.4775	0.5159	0.4377	0.4726
	HazardDev	**0.6082**	**0.5951**	**0.6459**	**0.6607**	**0.5347**	**0.5864**	**0.5527**	0.5203	**0.5854**	**0.5906**
RFFS	ASH	0.4653	0.4834	0.4197	0.4672	0.4850	0.4836	0.5410	0.5405	0.4778	0.4937
	Dropout	0.4310	0.4474	0.3781	0.4161	0.5147	0.5397	0.5408	0.5296	0.4661	0.4832
	Energy	0.4128	0.4372	0.4019	0.4219	0.4879	0.5073	0.5088	0.5018	0.4528	0.4671
	GEN	0.5438	0.5143	0.5305	0.5208	0.5272	**0.5757**	0.5417	0.5660	0.5358	0.5442
	Dice	0.4317	0.4457	0.3930	0.4203	**0.5288**	0.5359	0.5313	0.5186	0.4712	0.4801
	MLS	0.4271	0.4446	0.3972	0.4215	**0.5288**	0.5395	0.5366	0.5242	0.4724	0.4824
	MSP	0.4998	0.4848	0.5517	0.5216	0.5084	0.5651	0.5378	**0.5671**	0.5244	0.5346
	ODIN	0.5066	0.4927	0.5085	0.5026	0.4950	0.5520	0.5194	0.5471	0.5074	0.5236
	SCALE	0.4330	0.4558	0.3978	0.4198	0.5074	0.5169	0.5398	0.5354	0.4695	0.4820
	HazardDev	**0.6060**	**0.6089**	**0.5984**	**0.6421**	0.5233	0.5112	**0.5500**	0.5376	**0.5694**	**0.5749**
DFFS	ASH	0.4608	0.4653	0.3638	0.4125	0.4811	0.5121	0.5170	0.5179	0.4557	0.4769
	Dropout	0.5387	0.5197	0.4728	0.4528	**0.5570**	0.5449	0.4918	0.5086	0.5151	0.5065
	Energy	0.4633	0.4713	0.3579	0.4067	0.4699	0.5086	0.4878	0.5066	0.4447	0.4733
	GEN	0.5420	0.5168	0.5543	0.4964	0.5486	0.5437	0.5374	0.5235	0.5456	0.5201
	Dice	0.5316	0.5112	0.4162	0.4282	0.5376	0.5377	0.4911	0.5074	0.4941	0.4961
	MLS	0.4725	0.4787	0.3861	0.4181	0.4656	0.5070	0.5046	0.5107	0.4572	0.4786
	MSP	**0.5587**	0.5277	0.5314	0.4800	0.5522	**0.5483**	0.5483	**0.5258**	0.5476	0.5204
	ODIN	0.5531	0.5380	0.4945	0.4605	0.5074	0.5173	**0.5514**	0.5171	0.5266	0.5082
	SCALE	0.4668	0.4738	0.3606	0.4108	0.4832	0.5114	0.5166	0.5141	0.4568	0.4775
	HazardDev	0.5386	**0.5554**	**0.6464**	**0.6630**	0.5338	0.5134	0.5201	0.5152	**0.5597**	**0.5617**

Table 4: Comparison of C-index between ID and OOD test sets across four survival prediction tasks. The consistent performance decrease (Dec.) on OOD data underscores the impact of acquisition-induced distribution shifts on survival prediction accuracy.

Task	**Pixel Spacing**			**Exposure Time**			**Slice Thickness**			**Tube Current**		
	ID	OOD	Dec.	ID	OOD	Dec.	ID	OOD	Dec.	ID	OOD	Dec.
OS	0.7163	0.6675	0.0488	0.7254	0.6857	0.0397	0.6583	0.6570	0.0013	0.7430	0.6819	0.0611
LFFS	0.7227	0.5969	0.1258	0.6431	0.5969	0.0462	0.7443	0.6937	0.0506	0.7411	0.6045	0.1366
RFFS	0.7974	0.7704	0.0270	0.8145	0.6125	0.2020	0.7405	0.5772	0.1633	0.6901	0.6376	0.0525
DFFS	0.7103	0.6784	0.0319	0.8912	0.5925	0.2987	0.7208	0.6751	0.0457	0.7646	0.6526	0.1120

6 Results on CURE-OOD

6.1 Performance Degradation under Data Shifts

To verify the impact of acquisition-induced shifts on downstream survival performance, we compare C-index on the ID and OOD test sets for each acquisition parameter. Table 4 reports C-index results across four tasks. When the ID-trained model is evaluated on OOD test data, survival accuracy generally decreases, showing that acquisition changes can degrade predictive performance. For OS, the degradation is modest for three shifts (0.0397–0.0611) but nearly disappears under slice thickness (0.0013), suggesting that the effect of acquisition shift is task- and parameter-dependent. In contrast, LFFS, RFFS, and DFFS show larger drops, often > 0.1, which may reflect the smaller number of events and the resulting class imbalance, making the models less robust under shifts. Overall, these results suggest that robustness under acquisition shift depends not only on the shift itself but also on endpoint-specific event characteristics. Together, the observed degradation patterns suggest that distribution shifts can have measurable downstream consequences for survival prediction.

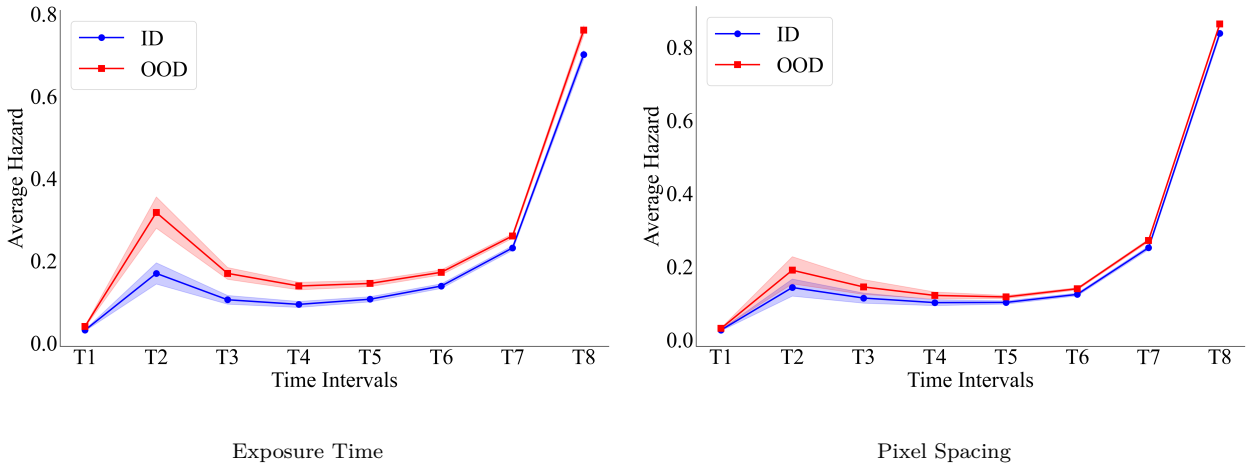


Figure 5: Mean hazard curves for ID and OOD test sets on OS under two acquisition shifts. Shaded regions indicate ± 1 std. OOD shows higher mean hazard than ID, suggesting hazard magnitude is an OOD signal.

6.2 Why Classification-Style OOD Scores Fail

Classification-style OOD detectors are typically designed for class-confidence scores, whereas survival prediction produces structured, time-dependent outputs. In this setting, these detectors can still be directly applied to interval-wise logits $f_\Theta(\mathbf{x}, k)$, enabling a fair comparison between conventional post-hoc OOD scores and survival-aware alternatives on the same benchmark. In practice, Table 3 shows that several mainstream logit-based methods, including Energy, ASH, and SCALE, do not provide reliable OOD signals in survival prediction; they often show weak discrimination and, in some settings, even reversed ordering, where ID samples receive more OOD-like scores than OOD samples.

The reason is that survival logits do not encode class confidence in the same way as classification logits. Instead, as described in Sec. 3.1, they parameterize a time-ordered survival distribution and therefore reflect a structured, time-dependent risk profile. At early intervals, most patients have not yet experienced an event, so the training objective places greater emphasis on survival through those intervals, which in turn encourages the corresponding logits to move toward smaller, more negative values. Many classification-style OOD scores implicitly assume that larger logit magnitude or score values correspond to stronger in-distribution confidence. In survival prediction, however, larger logits are not necessarily more confident predictions, so the resulting score ordering need not follow the confidence assumptions used by classification-style OOD detectors.

Under acquisition-induced shift, OOD samples fall into less familiar feature regions, making the model more likely to assign higher event risk than it does for well-aligned ID samples. Their logits therefore become larger and drift toward zero. Consequently, ID samples often produce more negative logits than OOD samples, which is the opposite of the usual classification pattern where ID examples yield stronger, larger-magnitude evidence for the predicted class. When classification-style scores are applied on top of these logits, the ordering can invert: scores that are supposed to be larger for OOD inputs may instead become larger for ID inputs, causing systematic ID-as-OOD errors.

This effect is shown in Fig. 6, where the average OOD logits are consistently higher and closer to zero than the corresponding ID logits under a representative acquisition shift. This is fundamentally different from standard classification, where ID samples typically induce stronger class evidence and more extreme logits than OOD samples. The logits distributions in Fig. 8 and the gradient analysis in Sec. A further support the same mechanism. Together, these results explain why the evaluated classification-oriented OOD scores are not reliably transferable to survival prediction, even when the same methods are effective in standard classification benchmarks.

Table 5: OOD detection results on the OS task with a ResNet backbone. We report AUROC and AUPRC across four acquisition shifts.

	Method	Pixel Spacing		Exposure Time		Slice Thick.		Tube Current		**Average**	
		AUROC	AUPRC	AUROC	AUPRC	AUROC	AUPRC	AUROC	AUPRC	AUROC	AUPRC
OS	ASH	0.4740	0.5079	0.3740	0.4167	0.4049	0.4357	0.4724	0.5100	0.4313	0.4676
	Dropout	0.5101	**0.5245**	0.4111	0.4335	0.4592	0.4554	0.5457	0.5201	0.4815	0.4834
	GEN	0.4965	0.5204	0.4708	0.4550	0.4828	0.4692	0.5405	0.5506	0.4976	0.4988
	Energy	0.4782	0.4934	0.3440	0.4175	0.3865	0.4386	0.4443	0.4454	0.4133	0.4487
	Dice	0.4673	0.5056	0.3853	0.4217	0.4139	0.4390	0.5565	0.5425	0.4557	0.4772
	MLS	0.4782	0.5096	0.3439	0.4076	0.3861	0.4303	0.4447	0.5042	0.4132	0.4629
	MSP	0.4862	0.5152	0.4985	0.4689	0.4956	0.4773	0.5392	0.5445	0.5049	0.5015
	ODIN	0.4769	0.5110	0.4851	0.4616	0.4897	0.4715	0.5382	0.5214	0.4975	0.4914
	SCALE	0.4771	0.5088	0.3604	0.4125	0.3797	0.4279	0.4754	0.5137	0.4232	0.4657
	HazardDev	**0.5218**	0.5118	**0.6559**	**0.6446**	**0.6135**	**0.6140**	**0.5657**	**0.5510**	**0.5892**	**0.5804**

6.3 Hazard-Based OOD Detection

In contrast, HazardDev remains aligned with the output semantics of survival prediction by operating on the predicted hazard curve instead of raw logits. As shown in Table 3, it provides the strongest average survival-aware reference among the compared methods across survival outcomes, with average AUROC improvements of 0.109, 0.062, 0.029, and 0.004 for OS, LFFS, RFFS, and DFFS, respectively, and average AUPR improvements of 0.095, 0.055, 0.022, and 0.004 in the same order. These results suggest that hazard-based deviation is a useful survival-aligned OOD signal under survival formulations.

To further probe this behavior, we visualize the mean hazard curves on ID vs. OOD test sets for OS under pixel spacing and exposure time shifts (Fig. 5). Across all discretized intervals, OOD samples exhibit systematically higher mean hazard than ID samples, supporting the use of hazard-based deviation as a task-consistent OOD signal. More broadly, this suggests that future OOD methods for survival prediction should be defined on survival-native predictive quantities rather than directly transferred classification confidence scores.

6.4 Architecture Sensitivity Analysis

To explore HazardDev's performance with a different feature extractor, we use a ResNet model and repeat training and OOD detection on the OS task under the same four acquisition shifts. Table 5 shows that the overall ranking changes with the feature extractor, which is expected because different backbones induce different feature geometries for post-hoc OOD scoring. However, the main qualitative trend remains: hazard-based scores stay competitive under the backbone change and continue to provide strong detection performance under acquisition shifts.

More specifically, the HazardDev score remains particularly effective on exposure time and slice thickness. Among the methods retained from Table 3, it attains the highest AUROC and AUPRC on exposure time, the highest AUROC and AUPRC on slice thickness, and the strongest average AUROC/AUPRC at 0.5892/0.5804. These results indicate that the utility of hazard-based deviation is not tied to a single backbone design and that this survival-aware signal remains informative under a different feature extractor. In turn, this suggests that the main benchmark findings extend across different feature extractors and capture a broader property of survival OOD detection.

6.5 Limits of OOD Detection for Robust Survival Prediction

We found that OOD separability does not necessarily reflect how much a shift harms the downstream survival task. To explore this relationship, we compare the C-index drop between ID and OOD test sets against OOD detection performance across all 16 task–shift combinations for three representative methods including HazardDev, GEN and MSP in Fig. 7. If stronger OOD signals always corresponded to larger deployment risk, points with larger Δ C-index would also show consistently higher OOD detection performance. Instead, the relationship is weak and unstable.

Across methods three methods, we found that larger C-index drops do not reliably correspond to better OOD detection. Even for HazardDev, the Pearson and Spearman correlations are only 0.258 and 0.050, respectively. This further shows that larger C-index drops do not reliably correspond to higher OOD detection

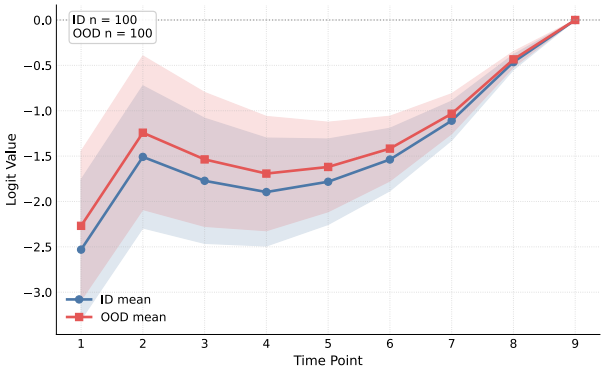


Figure 6: Mean logits on the OS task under the exposure time shift. Unlike standard classification logits, ID samples produce systematically smaller, more negative logits across time intervals, while OOD samples drift toward zero. This reversed ordering breaks the confidence interpretation assumed by classification-style OOD scores.

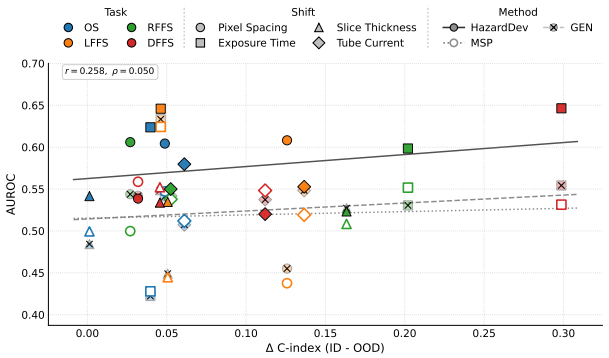


Figure 7: Relationship between downstream survival degradation and OOD detection quality across task–shift pairs. Each point is one method on one task–shift combination. The reported r and ρ are Pearson and Spearman correlations for HazardDev, indicating weak linear and rank associations between OOD detectability and downstream performance loss.

performance. Some shifts are easy to distinguish from ID data yet only modestly reduce survival accuracy, whereas others substantially degrade prediction while remaining only moderately separable. Taken together, these findings suggest that current methods remain limited in supporting robust survival prediction under acquisition-induced shifts.

7 Conclusion and Discussion

We introduced CURE-OOD, to our knowledge, the first benchmark for OOD detection in cancer survival prediction built on real-world clinical CT data, through a controlled evaluation protocol that uses acquisition parameter variation to instantiate realistic covariate shifts. Through experiments spanning multiple acquisition shifts and representative OOD detectors, we demonstrate that acquisition shifts substantially degrade survival prediction, and that OOD methods developed for classification can fail in survival prediction due to distinct logit dynamics. As a simple survival-aware reference baseline, HazardDev further suggests that hazard-based signals can provide a useful direction for OOD detection in survival prediction. Future work may build on CURE-OOD to develop more task-aligned OOD detection and robustness evaluation under distribution shifts.

References

Odd O Aalen. A linear regression model for the analysis of life times. *Statistics in medicine*, 8(8):907–925, 1989.

Mahul B Amin, Stephen B Edge, Frederick L Greene, David R Byrd, Robert K Brookland, Mary Kay Washington, Jeffrey E Gershenwald, Carolyn C Compton, Kenneth R Hess, Daniel C Sullivan, et al. *AJCC cancer staging manual*, volume 1024. Springer, 2017.

K. Kian Ang, Jonathan Harris, Richard Wheeler, Randal Weber, David I. Rosenthal, Phuc Felix Nguyen-Tân, William H. Westra, Christine H. Chung, Richard C. Jordan, Charles Lu, Harold Kim, Rita Axelrod, C. Craig Silverman, Kevin P. Redmond, and Maura L. Gillison. Human papillomavirus and survival of patients with oropharyngeal cancer. *New England Journal of Medicine*, 363(1):24 – 35, 2010. doi: 10.1056/NEJMoa0912217. URL https://www.scopus.com/inward/record.uri?eid=2-s2.0-77954190940&doi=10.1056%2fNEJMoa0912217&partnerID=40&md5=53b452be4e88b8324284858bd1acf5d0. Cited by: 5671; All Open Access, Green Open Access.

Marc Aubreville, Nikolas Stathonikos, Christof A. Bertram, Robert Klopfleisch, Natalie ter Hoeve, Francesco Ciompi, Frauke Wilm, Christian Marzahl, Taryn A. Donovan, Andreas Maier, Jack Breen, Nishant Ravikumar, Youjin Chung, Jinah Park, Ramin Nateghi, Fattaneh Pourakpour, Rutger H.J. Fick, Saima Ben Hadj, Mostafa Jahanifar, Adam Shephard, Jakob Dexl, Thomas Wittenberg, Satoshi Kondo, Maxime W. Lafarge, Viktor H. Koelzer,

Jingtang Liang, Yubo Wang, Xi Long, Jingxin Liu, Salar Razavi, April Khademi, Sen Yang, Xiyue Wang, Ramona Erber, Andrea Klang, Karoline Lipnik, Pompei Bolfa, Michael J. Dark, Gabriel Wasinger, Mitko Veta, and Katharina Breininger. Mitosis domain generalization in histopathology images — the midog challenge. *Medical Image Analysis*, 84:102699, 2023. ISSN 1361-8415. doi: https://doi.org/10.1016/j.media.2022.102699. URL https://www.sciencedirect.com/science/article/pii/S1361841522003279.

Eunsu Baek, Keondo Park, Jiyoon Kim, and Hyung-Sin Kim. Unexplored faces of robustness and out-of-distribution: Covariate shifts in environment and sensor domains. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 22294–22303, 2024.

Lauren J Beesley, Peter G Hawkins, Lahin M Amlani, Emily L Bellile, Keith A Casper, Steven B Chinn, Avraham Eisbruch, Michelle L Mierzwa, Matthew E Spector, Gregory T Wolf, et al. Individualized survival prediction for patients with oropharyngeal cancer in the human papillomavirus era. *Cancer*, 125(1):68–78, 2019.

Tianshi Cao, Chin-Wei Huang, David Yu-Tung Hui, and Joseph Paul Cohen. A benchmark of medical out of distribution detection. *arXiv preprint arXiv:2007.04250*, 2020.

Jimmy J Caudell, Javier F Torres-Roca, Robert J Gillies, Heiko Enderling, Sungjune Kim, Anupam Rishi, Eduardo G Moros, and Louis B Harrison. The future of personalised radiotherapy for head and neck cancer. *The Lancet Oncology*, 18(5):e266–e273, 2017. ISSN 1470-2045. doi: https://doi.org/10.1016/S1470-2045(17)30252-8. URL https://www.sciencedirect.com/science/article/pii/S1470204517302528.

Meixu Chen, Kai Wang, and Jing Wang. Advancing head and neck cancer survival prediction via multi-label learning and deep model interpretation. *ArXiv*, pp. arXiv–2405, 2024a.

Meixu Chen, Kai Wang, and Jing Wang. Vision transformer-based multilabel survival prediction for oropharynx cancer after radiation therapy. *International Journal of Radiation Oncology* Biology* Physics*, 118(4):1123–1134, 2024b.

Travers Ching, Xun Zhu, and Lana X Garmire. Cox-nnet: an artificial neural network method for prognosis prediction of high-throughput omics data. *PLoS computational biology*, 14(4):e1006076, 2018.

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

Andrija Djurisic, Nebojsa Bozanic, Arjun Ashok, and Rosanne Liu. Extremely simple activation shaping for out-of-distribution detection. *arXiv preprint arXiv:2209.09858*, 2022.

Jean-Philippe Fortin, Nicholas Cullen, Yvette I Sheline, Warren D Taylor, Irem Aselcioglu, Philip A Cook, Phil Adams, Crystal Cooper, Maurizio Fava, Patrick J McGrath, et al. Harmonization of cortical thickness measurements across scanners and sites. *Neuroimage*, 167:104–120, 2018.

Hao Guan and Mingxia Liu. Domain adaptation for medical image analysis: a survey. *IEEE Transactions on Biomedical Engineering*, 69(3):1173–1185, 2021.

Max Gutbrod, David Rauber, Danilo Weber Nunes, and Christoph Palm. Openmibood: Open medical imaging benchmarks for out-of-distribution detection. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 25874–25886, 2025.

Ali Hatamizadeh, Yucheng Tang, Vishwesh Nath, Dong Yang, Andriy Myronenko, Bennett Landman, Holger R Roth, and Daguang Xu. Unetr: Transformers for 3d medical image segmentation. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pp. 574–584, 2022.

Dan Hendrycks and Thomas Dietterich. Benchmarking neural network robustness to common corruptions and perturbations. *arXiv preprint arXiv:1903.12261*, 2019.

Dan Hendrycks and Kevin Gimpel. A baseline for detecting misclassified and out-of-distribution examples in neural networks. *arXiv preprint arXiv:1610.02136*, 2016.

Dan Hendrycks, Kevin Zhao, Steven Basart, Jacob Steinhardt, and Dawn Song. Natural adversarial examples. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 15262–15271, 2021.

Claudia I Henschke, Rowena Yip, Dorith Shaham, Steven Markowitz, José Cervera Deval, Javier J Zulueta, Luis M Seijo, Cheryl Aylesworth, Karl Klingler, Shahriyour Andaz, et al. A 20-year follow-up of the international early lung cancer action program (i-elcap). *Radiology*, 309(2):e231988, 2023.

John Kang, Russell Schwartz, John Flickinger, and Sushil Beriwal. Machine learning approaches for predicting radiation therapy outcomes: A clinician's perspective. *International Journal of Radiation Oncology*Biology*Physics*, 93(5):1127–1135, 2015. ISSN 0360-3016. doi: https://doi.org/10.1016/j.ijrobp.2015.07.2286. URL https://www.sciencedirect.com/science/article/pii/S0360301615030783.

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):24, 2018.

Andrzej Kawecki and Romuald Krajewski. Follow-up in patients treated for head and neck cancer. *Memo-Magazine of European Medical Oncology*, 7(2):87–91, 2014.

Sejin Kim, Michal Kazmierski, and Benjamin Haibe-Kains. Deep-cr mtlr: A multi-modal approach for cancer survival prediction with competing risks. In *Survival Prediction-Algorithms, Challenges and Applications*, pp. 223–231. PMLR, 2021.

Pang Wei Koh, Shiori Sagawa, Henrik Marklund, Sang Michael Xie, Marvin Zhang, Akshay Balsubramani, Weihua Hu, Michihiro Yasunaga, Richard Lanas Phillips, Irena Gao, Tony Lee, Etienne David, Ian Stavness, Wei Guo, Berton Earnshaw, Imran Haque, Sara M Beery, Jure Leskovec, Anshul Kundaje, Emma Pierson, Sergey Levine, Chelsea Finn, and Percy Liang. Wilds: A benchmark of in-the-wild distribution shifts. In Marina Meila and Tong Zhang (eds.), *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 5637–5664. PMLR, 18–24 Jul 2021. URL https://proceedings.mlr.press/v139/koh21a.html.

Håvard Kvamme, Ørnulf Borgan, and Ida Scheel. Time-to-event prediction with neural networks and cox regression. *Journal of machine learning research*, 20(129):1–30, 2019.

Pamela J LaMontagne, Tammie LS Benzinger, John C Morris, Sarah Keefe, Russ Hornbeck, Chengjie Xiong, Elizabeth Grant, Jason Hassenstab, Krista Moulder, Andrei G Vlassenko, et al. Oasis-3: longitudinal neuroimaging, clinical, and cognitive dataset for normal aging and alzheimer disease. *medrxiv*, pp. 2019–12, 2019.

Shiyu Liang, Yixuan Li, and Rayadurgam Srikant. Enhancing the reliability of out-of-distribution image detection in neural networks. *arXiv preprint arXiv:1706.02690*, 2017.

Thomas J Littlejohns, Jo Holliday, Lorna M Gibson, Steve Garratt, Niels Oesingmann, Fidel Alfaro-Almagro, Jimmy D Bell, Chris Boultwood, Rory Collins, Megan C Conroy, et al. The uk biobank imaging enhancement of 100,000 participants: rationale, data collection, management and future directions. *Nature communications*, 11(1): 2624, 2020.

Weitang Liu, Xiaoyun Wang, John Owens, and Yixuan Li. Energy-based out-of-distribution detection. *Advances in neural information processing systems*, 33:21464–21475, 2020.

Yuhang Liu, Wenjie Zhao, and Yunhui Guo. H2st: Hierarchical two-sample tests for continual out-of-distribution detection. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 15413–15423, 2025.

Dennis Mackin, Xenia Fave, Lifei Zhang, David Fried, Jinzhong Yang, Brian Taylor, Edgardo Rodriguez-Rivera, Cristina Dodge, Aaron Kyle Jones, et al. Measuring computed tomography scanner variability of radiomics features. *Investigative radiology*, 50(11):757–765, 2015.

Torben Martinussen and Thomas H Scheike. *Dynamic regression models for survival data*. Springer, 2006.

Chirag Nagpal, Steve Yadlowsky, Negar Rostamzadeh, and Katherine Heller. Deep cox mixtures for survival regression. In *Machine Learning for Healthcare Conference*, pp. 674–708. PMLR, 2021.

Hubert Y. Pan, Bruce G. Haffty, Benjamin P. Falit, Thomas A. Buchholz, Lynn D. Wilson, Stephen M. Hahn, and Benjamin D. Smith. Supply and demand for radiation oncology in the united states: Updated projections for 2015 to 2025. *International Journal of Radiation Oncology*Biology*Physics*, 96(3):493–500, 2016. ISSN 0360-3016. doi: https://doi.org/10.1016/j.ijrobp.2016.02.064. URL https://www.sciencedirect.com/science/article/pii/S0360301616002339.

Tobias Rueckert. Phakir challenge at miccai 2024. https://phakir.re-mic.de/, 2024. Accessed: 2025-11-08.

Numan Saeed, Muhammad Ridzuan, Fadillah Adamsyah Maani, Hussain Alasmawi, Karthik Nandakumar, and Mohammad Yaqub. Survrnc: Learning ordered representations for survival prediction using rank-n-contrast. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 659–669. Springer, 2024.

Shibani Santurkar, Dimitris Tsipras, and Aleksander Madry. Breeds: Benchmarks for subpopulation shift. *arXiv preprint arXiv:2008.04859*, 2020.

Youssef Shoeb, Robin Chan, Gesina Schwalbe, Azarm Nowzad, Fatma Güney, and Hanno Gottschalk. Have we ever encountered this before? retrieving out-of-distribution road obstacles from driving scenes. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 7396–7406, 2024.

Tomáš Vojíř, Jan Šochman, and Jiří Matas. Pixood: Pixel-level out-of-distribution detection. In *European Conference on Computer Vision*, pp. 93–109. Springer, 2024.

Mattea L Welch, Sejin Kim, Andrew J Hope, Shao Hui Huang, Zhibin Lu, Joseph Marsilla, Michal Kazmierski, Katrina Rey-McIntyre, Tirth Patel, Brian O'Sullivan, et al. Radcure: An open-source head and neck cancer ct dataset for clinical radiation therapy insights. *Medical Physics*, 51(4):3101–3109, 2024.

ML Welch, S Kim, A Hope, SH Huang, Z Lu, J Marsilla, M Kazmierski, K Rey-McIntyre, T Patel, B O'Sullivan, et al. Computed tomography images from large head and neck cohort (radcure). *The Cancer Imaging Archive (TCIA)*, 2023.

Simon Wiegrebe, Philipp Kopper, Raphael Sonabend, Bernd Bischl, and Andreas Bender. Deep learning for survival analysis: a review. *Artificial Intelligence Review*, 57(3):65, 2024.

Jiancheng Yang, Rui Shi, Donglai Wei, Zequan Liu, Lin Zhao, Bilian Ke, Hanspeter Pfister, and Bingbing Ni. Medmnist v2-a large-scale lightweight benchmark for 2d and 3d biomedical image classification. *Scientific Data*, 10(1):41, 2023.

Jingkang Yang, Pengyun Wang, Dejian Zou, Zitang Zhou, Kunyuan Ding, Wenxuan Peng, Haoqi Wang, Guangyao Chen, Bo Li, Yiyou Sun, et al. Openood: Benchmarking generalized out-of-distribution detection. *Advances in Neural Information Processing Systems*, 35:32598–32611, 2022.

Chun-Nam Yu, Russell Greiner, Hsiu-Chin Lin, and Vickie Baracos. Learning patient-specific cancer survival distributions as a sequence of dependent regressors. *Advances in neural information processing systems*, 24, 2011.

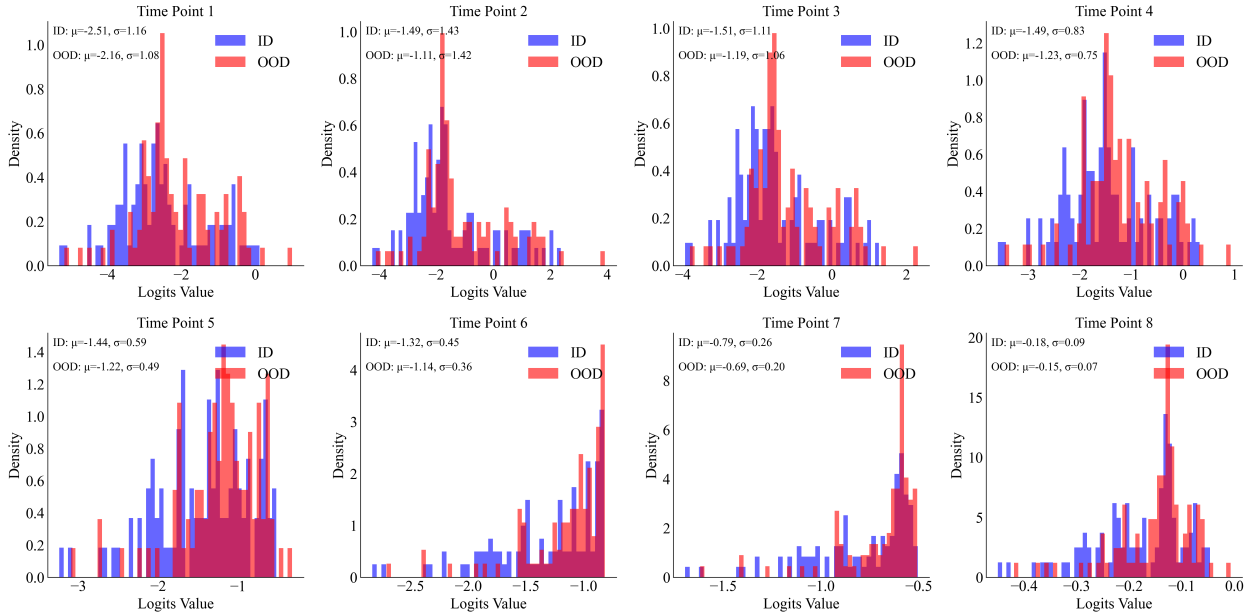
John R Zech, Marcus A Badgeley, Manway Liu, Anthony B Costa, Joseph J Titano, and Eric Karl Oermann. Variable generalization performance of a deep learning model to detect pneumonia in chest radiographs: a cross-sectional study. *PLoS medicine*, 15(11):e1002683, 2018.

Yu Zhang and Qiang Yang. A survey on multi-task learning. *IEEE transactions on knowledge and data engineering*, 34(12):5586–5609, 2021.

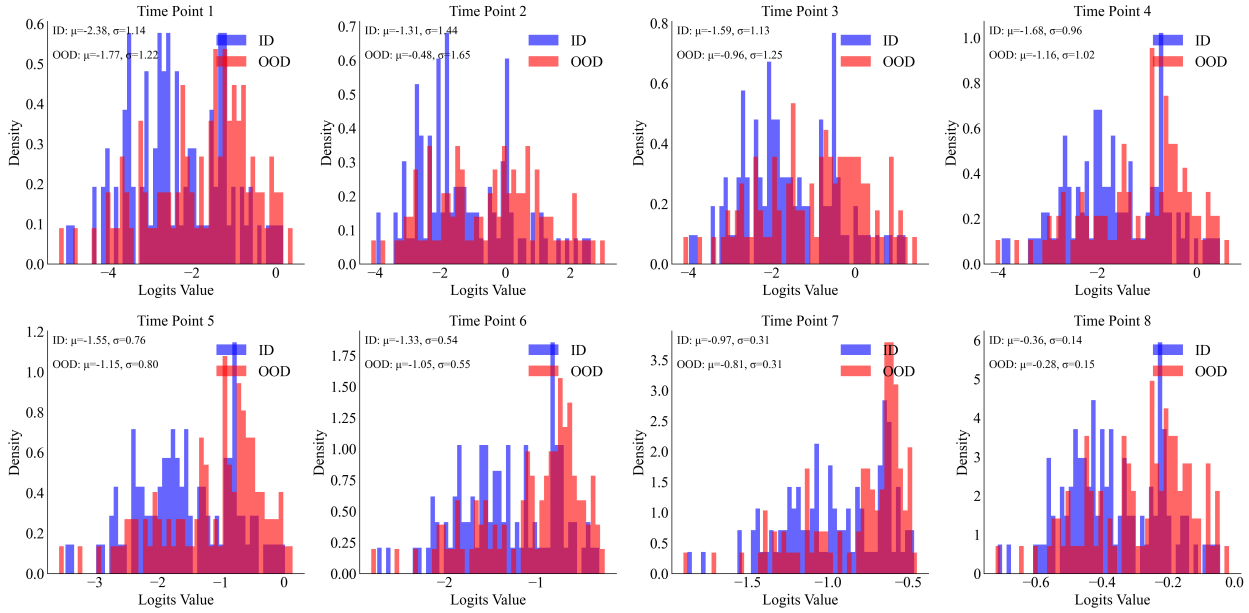
Wenjie Zhao, Jia Li, Xin Dong, Yu Xiang, and Yunhui Guo. Segment every out-of-distribution object. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3910–3920, 2024.

Wenjie Zhao, Jia Li, and Yunhui Guo. From pixel to mask: A survey of out-of-distribution segmentation. *arXiv preprint arXiv:2508.10309*, 2025.

Wenjie Zhao, Jia Li, Xin Dong, Yapeng Tian, Yu Xiang, and Yunhui Guo. Mitigating the id-ood tradeoff in open-set test-time adaptation. *arXiv preprint arXiv:2604.01589*, 2026.

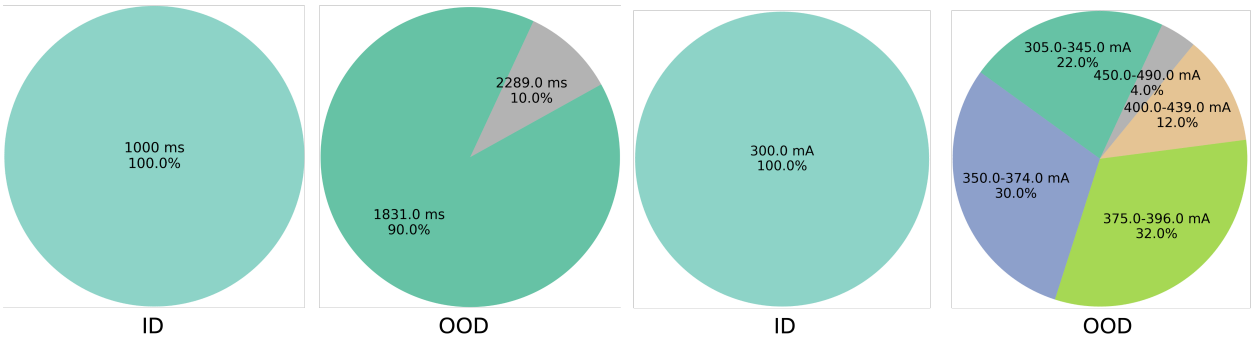


(a) Pixel Spacing



(b) Exposure Time

Figure 8: Visualization of MTLR logits distributions on the OS task under two acquisition shifts: (a) pixel spacing and (b) exposure time. For both shifts, ID samples tend to produce logits concentrated at lower values, whereas OOD samples show a clear tendency toward larger logits. This shift in the logits distribution indicates that acquisition differences systematically affect the model's output confidence, providing a useful signal for OOD detection.



(a) Exposure time distribution across ID and OOD.

(b) X-ray tube current distribution across ID and OOD.

Figure 9: Distributions of key acquisition parameters used to construct ID and OOD domains in the CURE-OOD benchmark. Each parameter exhibits distinct value ranges.

Appendix

A Gradient Dynamics of MTLR and Logit Behavior

To understand why logits-based OOD detection methods such as Energy and SCALE behave inversely on survival modeling tasks, we analyze the formulation of the MTLR model.

MTLR divides the continuous survival time into m discrete intervals $\{t_1, t_2, \ldots, t_m\}$ and models the conditional probability of survival at each interval. For a patient with feature $\mathbf{x}$ and survival status sequence $y = (y_1, y_2, \ldots, y_m)$, where $y_i = 0$ indicates survival up to t_i, the likelihood is defined as:

$$P_\Theta(Y = y \mid \mathbf{x}) = \frac{\exp\left(\sum_{i=1}^{m} y_i(\theta_i^\top \mathbf{x} + b_i)\right)}{\sum_{k=0}^{m} \exp(f_\Theta(\mathbf{x}, k))}, \tag{5}$$

$$\text{where } f_\Theta(\mathbf{x}, k) = \sum_{i=k+1}^{m} (\theta_i^\top \mathbf{x} + b_i).$$

Here, $f_\Theta(\mathbf{x}, k)$ represents the unnormalized logit (score) associated with the event occurring at time interval k. The denominator normalizes these scores over all possible event times, forming a valid probability distribution across discrete survival intervals.

The model is optimized by minimizing the *regularized negative log-likelihood*, which corresponds to maximizing the log-likelihood of the observed survival sequences with additional ℓ_2 and smoothness regularization terms (Yu et al., 2011). Formally, the optimization objective can be written as:

$$\min_\Theta \frac{C_1}{2} \sum_{j=1}^{m} \|\boldsymbol{\theta}_j\|_2^2 + \frac{C_2}{2} \sum_{j=1}^{m-1} \|\boldsymbol{\theta}_{j+1} - \boldsymbol{\theta}_j\|_2^2 - \mathcal{L}_{\text{MTLR}}, \tag{6}$$

where $\mathcal{L}_{\text{MTLR}}$ denotes the log-likelihood term:

$$\begin{aligned} \mathcal{L}_{\text{MTLR}} = \sum_{i=1}^{N} \Big[& \sum_{j=1}^{m} y_j(s_i)(\theta_j^\top \mathbf{x}_i + b_j) \\ & - \log \sum_{k=0}^{m} \exp f_\Theta(\mathbf{x}_i, k) \Big]. \end{aligned} \tag{7}$$

The first regularizer $\sum_j \|\boldsymbol{\theta}_j\|_2^2$ constrains the parameter magnitude to prevent overfitting, while the second term $\sum_j \|\boldsymbol{\theta}_{j+1} - \boldsymbol{\theta}_j\|_2^2$ enforces smoothness across consecutive time intervals.

During optimization, the first component of $\mathcal{L}_{\text{MTLR}}$ increases the scores for observed event intervals, whereas the second term normalizes the probabilities across all possible time sequences. This formulation naturally

drives the logits to form a temporal ordering. At early time intervals, where most samples remain alive, the logits become smaller (more negative). At later intervals, where events are more likely to occur, the logits gradually increase (become less negative). This trend is consistently observed across datasets, as shown in Fig. 8, where the mean logits shift upward from early to late time bins. We analyze the gradient dynamics below to theoretically explain why this monotonic pattern emerges during training.

A.1 Gradient Analysis and the Tendency Toward Monotonic Logits

We derive the gradient of the core likelihood term and show why optimization encourages smaller (more negative) logits at early intervals and larger (less negative) logits at later intervals.

Setup. For a sample $(\mathbf{x}_i, s_i)$, define the normalizer

$$S_i(\Theta) := \sum_{k=0}^{m} \exp f_\Theta(\mathbf{x}_i, k), \qquad \pi_{ik} := \frac{\exp f_\Theta(\mathbf{x}_i, k)}{S_i(\Theta)}.$$

Recall that $f_\Theta(\mathbf{x}, k) = \sum_{r=k+1}^{m} (\boldsymbol{\theta}_r^\top \mathbf{x} + b_r)$, so for a fixed $j \in \{1, \dots, m\}$,

$$\frac{\partial f_\Theta(\mathbf{x}_i, k)}{\partial \boldsymbol{\theta}_j} = \begin{cases} \mathbf{x}_i, & k < j, \\ \mathbf{0}, & k \geq j. \end{cases} \tag{A.1}$$

Gradient of the per-sample objective. For the core likelihood term

$$\ell_i(\Theta) = \sum_{r=1}^{m} y_r(s_i)\,(\boldsymbol{\theta}_r^\top \mathbf{x}_i + b_r) - \log S_i(\Theta),$$

the gradient with respect to $\boldsymbol{\theta}_j$ is

$$\frac{\partial \ell_i(\Theta)}{\partial \boldsymbol{\theta}_j} = y_j(s_i)\,\mathbf{x}_i - \frac{\partial}{\partial \boldsymbol{\theta}_j} \log S_i(\Theta). \tag{A.2}$$

The first term directly increases the score for interval j when the sample survives past t_j. The second term involves the derivative of the log-normalizer, which we compute next.

Derivative of the log-normalizer. Using equation A.1, we have

$$\begin{aligned} \frac{\partial}{\partial \boldsymbol{\theta}_j} \log S_i(\Theta) &= \frac{1}{S_i(\Theta)} \sum_{k=0}^{m} \exp f_\Theta(\mathbf{x}_i, k)\, \frac{\partial f_\Theta(\mathbf{x}_i, k)}{\partial \boldsymbol{\theta}_j} \\ &= \frac{1}{S_i(\Theta)} \sum_{k=0}^{j-1} \exp f_\Theta(\mathbf{x}_i, k)\, \mathbf{x}_i = \sum_{k=0}^{j-1} \pi_{ik}\, \mathbf{x}_i. \end{aligned} \tag{A.3}$$

Substituting equation A.3 into equation A.2 yields

$$\frac{\partial \ell_i(\Theta)}{\partial \boldsymbol{\theta}_j} = \left(y_j(s_i) - \sum_{k=0}^{j-1} \pi_{ik} \right) \mathbf{x}_i. \tag{A.4}$$

Summing over all samples gives the overall gradient

$$\frac{\partial \mathcal{L}_{\text{MTLR}}}{\partial \boldsymbol{\theta}_j} = \sum_{i=1}^{N} \left(y_j(s_i) - \sum_{k=0}^{j-1} \pi_{ik} \right) \mathbf{x}_i, \tag{A.5}$$

to which one may add the ℓ_2 and smoothness regularizers from (Yu et al., 2011) if included.

Interpretation. The term $y_j(s_i)$ indicates whether sample i has *survived past* t_j (with the 0/1 encoding in the main text). The cumulative softmax mass $\sum_{k=0}^{j-1} \pi_{ik}$ is the model's probability assigned to the event occurring *before* t_j. Hence the gradient equation A.5 pushes the parameters so that:

- if many samples survive past t_j (empirically common for early j), then $y_j(s_i)$ tends to dominate, and the update decreases $\sum_{k<j} \pi_{ik}$ by *lowering* the logits $f_\Theta(\mathbf{x}_i, k)$ for $k < j$ (via equation A.1), making early-interval logits more negative;

- conversely, for later j where events are more likely, the model is encouraged to allocate more mass to later intervals, effectively making those logits less negative.

Aggregated over the dataset, this drives a characteristic pattern in which early-interval logits become smaller (more negative) while later-interval logits become larger (less negative), so that the learned logits *tend to* exhibit a monotonic progression across time, consistent with increasing event risk. We provide a detailed derivation of this mechanism in Section A.2, showing how each optimization step explicitly suppresses the cumulative probability of early intervals.

A.2 From the Gradient to Early-Interval Probability Suppression

Building upon the gradient expression above, we further examine how each optimization step explicitly suppresses the early cumulative probability mass $\sum_{k<j} \pi_{ik}$, leading to a systematic decrease in early logits.

We show why maximizing the likelihood drives down the cumulative early probability $\sum_{k<j} \pi_{ik}$ and thus pushes early logits $f_\Theta(\mathbf{x}_i, k)$, $k < j$, to more negative values.

SGD update on $\boldsymbol{\theta}_j$. Consider one SGD step on sample $(\mathbf{x}_i, s_i)$ with stepsize $\eta > 0$:

$$\Delta\boldsymbol{\theta}_j = \eta \frac{\partial \ell_i}{\partial \boldsymbol{\theta}_j} = \eta\Big(y_j(s_i) - \sum_{k=0}^{j-1} \pi_{ik}\Big)\mathbf{x}_i, \tag{A.5}$$

where $\pi_{ik} = \exp f_\Theta(\mathbf{x}_i, k) / \sum_r \exp f_\Theta(\mathbf{x}_i, r)$.

By equation A.1, a small parameter update $\Delta\boldsymbol{\theta}_j$ induces a corresponding change in the logits $f_\Theta(\mathbf{x}_i, k)$ through the chain rule:

$$\begin{aligned} \Delta f_\Theta(\mathbf{x}_i, k) &\approx \Big\langle \nabla_{\boldsymbol{\theta}_j} f_\Theta(\mathbf{x}_i, k), \Delta\boldsymbol{\theta}_j \Big\rangle = \Big(\tfrac{\partial f_\Theta(\mathbf{x}_i,k)}{\partial \boldsymbol{\theta}_j}\Big)^\top \Delta\boldsymbol{\theta}_j \\ &= \mathbf{x}_i^\top \Delta\boldsymbol{\theta}_j, \qquad k < j, \end{aligned} \tag{A.6a}$$

since $\partial f_\Theta(\mathbf{x}_i, k)/\partial\boldsymbol{\theta}_j = \mathbf{x}_i$ only holds for $k < j$. Substituting the update rule from equation A.5 gives

$$\begin{aligned} \Delta f_\Theta(\mathbf{x}_i, k) &= \mathbf{x}_i^\top \Big[\eta\Big(y_j(s_i) - \sum_{r=0}^{j-1} \pi_{ir}\Big)\mathbf{x}_i\Big] \\ &= \eta\Big(y_j(s_i) - \sum_{r=0}^{j-1} \pi_{ir}\Big)\|\mathbf{x}_i\|_2^2, \qquad k < j, \end{aligned} \tag{A.6}$$

and leaves $f_\Theta(\mathbf{x}_i, k)$ unchanged for $k \geq j$ since $\partial f_\Theta/\partial\boldsymbol{\theta}_j = \mathbf{0}$ in that case.

Sign of the update at early j. With the $0 \to 1$ encoding in MTLR, $y_j(s_i) = 1$ iff the sample has *survived past* t_j. Empirically, at *early* j most samples survive past t_j, so for many i we have $y_j(s_i) \approx 1$ while the model still allocates nonzero mass to earlier events, $\sum_{r<j} \pi_{ir} > 0$. Hence the coefficient in equation A.6 is typically a positive number slightly less than 1,

$$y_j(s_i) - \sum_{r<j} \pi_{ir} \;\approx\; 1 - \epsilon, \quad \epsilon > 0.$$

Although this term itself is positive, its effect on earlier logits $f_\Theta(\mathbf{x}_i, k)$ with $k < j$ is *negative*. This arises because each logit $f_\Theta(\mathbf{x}_i, k)$ accumulates all later parameters $\{\boldsymbol{\theta}_r : r > k\}$ in $f_\Theta(\mathbf{x}_i, k) = \sum_{r=k+1}^{m} (\boldsymbol{\theta}_r^\top \mathbf{x}_i + b_r)$. When $\boldsymbol{\theta}_j$ increases, it raises the logits of later intervals more strongly than those of earlier ones, thereby reducing the relative magnitude of early logits under the softmax normalization. Consequently,

$$\Delta f_\Theta(\mathbf{x}_i, k) \leq 0, \qquad k < j, \tag{A.7}$$

meaning that SGD decreases all early logits, pushing them toward more negative values.

Effect on the early cumulative probability. The softmax probabilities are monotone in their logits. Writing $\boldsymbol{f}_i = (f_\Theta(\mathbf{x}_i, 0), \ldots, f_\Theta(\mathbf{x}_i, m))$ and $\boldsymbol{\pi}_i = \mathrm{softmax}(\boldsymbol{f}_i)$, the Jacobian is

$$\frac{\partial \pi_{iq}}{\partial f_\Theta(\mathbf{x}_i, k)} = \pi_{iq}\big(\mathbb{1}\{q = k\} - \pi_{ik}\big). \tag{A.8}$$

If all early logits $\{f_\Theta(\mathbf{x}_i, k) : k < j\}$ decrease by $\{\Delta f_\Theta(\mathbf{x}_i, k) \leq 0\}$ while the others stay fixed, then the total early mass strictly decreases:

$$\Delta\Big(\sum_{k<j} \pi_{ik}\Big) = \sum_{k<j} \sum_{q=0}^{m} \frac{\partial \pi_{iq}}{\partial f_\Theta(\mathbf{x}_i, k)} \, \Delta f_\Theta(\mathbf{x}_i, k) \;<\; 0. \tag{A.9}$$

Hence the update equation A.5 *suppresses* the cumulative early probability.

Population effect. Aggregating such SGD steps over the dataset, early intervals (where survival is prevalent) receive repeated updates of the form equation A.7, which consistently lower their logits and probability mass; later intervals receive comparatively fewer such decreases. Consequently, the learned logits tend to become smaller (more negative) at early k and larger (less negative) at late k, yielding the empirical monotonic trend reported in the main text.

A.3 Why ID Logits Are More Negative

For ID samples, the model has been optimized to align the logits with the empirical hazard trend of the training data. At early time points, the survival probability is high and the event likelihood is low, leading the logits $f_\Theta(\mathbf{x}, k)$ to take more negative values. At later time points, as the risk of event increases, the logits gradually become less negative. This pattern is a direct consequence of likelihood maximization and the monotonicity constraint imposed by the MTLR framework.

A.4 Why OOD Logits Become Larger

When encountering OOD samples, the feature representations $\mathbf{x}$ deviate from the training distribution. In these unseen regions, the learned parameters $\{\theta_i, b_i\}$ produce less calibrated responses, and the model tends to generate logits that are closer to zero. Formally, since $f_\Theta(\mathbf{x}, k)$ is a linear projection of $\mathbf{x}$,

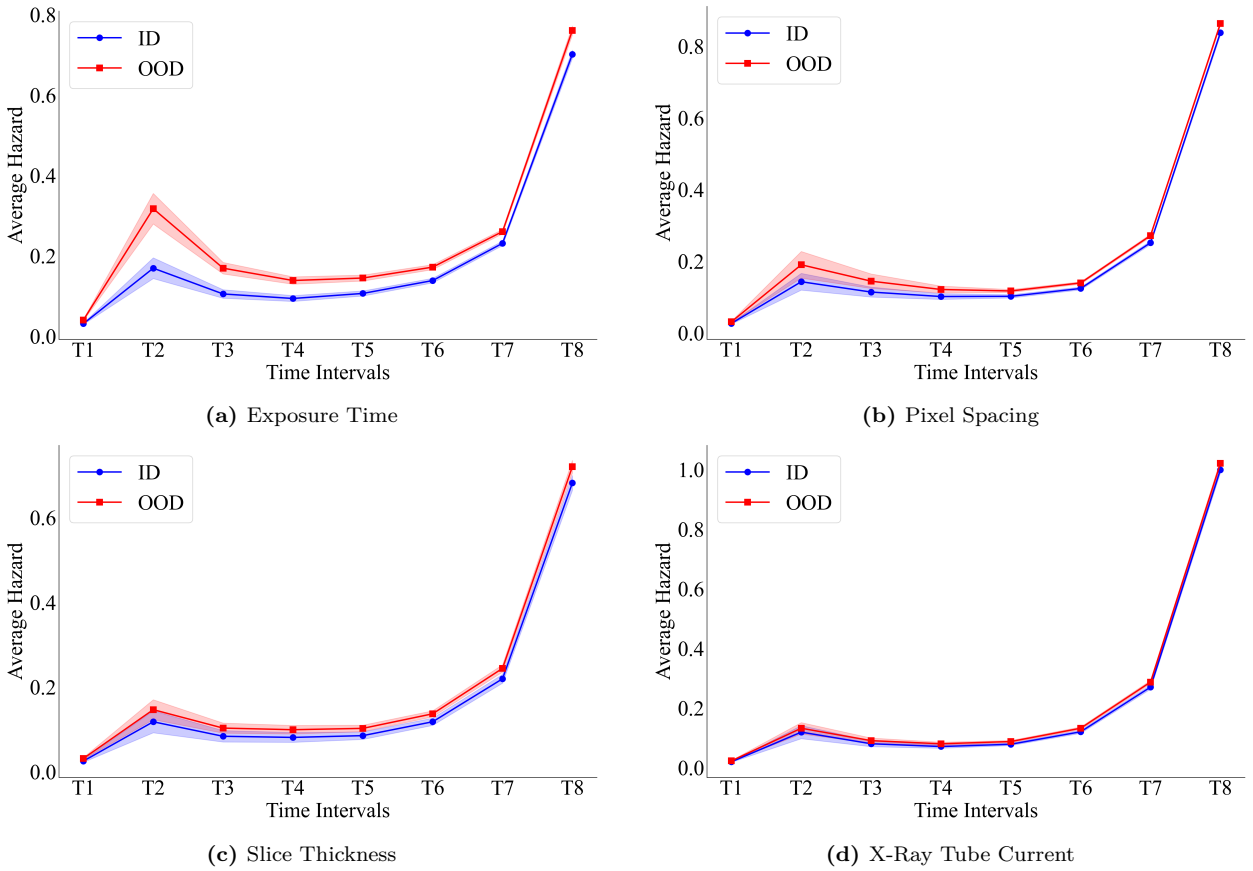
$$f_\Theta(\mathbf{x}, k) = \sum_{i=k+1}^{m} (\theta_i^\top \mathbf{x} + b_i),$$

a shift in $\mathbf{x}$ toward regions unobserved during training leads to a reduction in the projection magnitude $|\theta_i^\top \mathbf{x}|$, effectively making logits less negative. This reflects increased predictive uncertainty and weaker separation between survival intervals.

A.5 Consequence for Logit-Based OOD Detection

Most classification-based OOD detection methods, such as Energy or SCALE, assume that OOD samples produce logits with smaller magnitudes (lower confidence) than ID samples. However, in MTLR, the opposite occurs: ID samples yield strongly negative logits due to their alignment with the high-survival regions of the training data, while OOD samples result in less negative logits. Consequently, the conventional logit-based energy scores are inverted, causing ID samples to be misidentified as OOD and vice versa.

This inversion originates from the interplay between the monotonic survival constraint and distributional shifts in $\mathbf{x}$. Empirically, we observe this consistent polarity reversal across most shifts in the CURE-OOD benchmark.



(a) Exposure Time

(b) Pixel Spacing

(c) Slice Thickness

(d) X-Ray Tube Current

Figure 10: Mean hazard curves of ID and OOD test sets on the OS task across four acquisition shifts: (a) exposure time, (b) pixel spacing, (c) slice thickness, and (d) X-ray tube current. The shaded areas denote one standard deviation, showing the variability of predicted hazard values across samples. Across all shifts, ID samples consistently exhibit higher mean hazard values than OOD samples, indicating that hazard magnitude can serve as an effective signal for OOD detection.