HHS Public Access

Author manuscript

Comput Biol Med. Author manuscript; available in PMC 2026 April 07.

Published in final edited form as:

Comput Biol Med. 2026 April 15; 206: 111603. doi:10.1016/j.compbiomed.2026.111603.

Empirical evaluation of variability and multi-institutional generalizability of deep learning survival models: application to renal cancer CT scans

Brennan Flannery[a,*], **Thomas DeSilvio**[a], **Mohsen Hariri**[b], **Amir Reza Sadri**[a], **Nicholas Heller**[c], **Christopher Weight**[c], **Satish E. Viswanath**[a,d]

[a]Case Western Reserve University, Department of Biomedical Engineering, 10900 Euclid Ave, Cleveland, OH 44106, USA

[b]Case Western Reserve University, Department of Computer and Data Sciences, 10900 Euclid Ave, Cleveland, OH 44106, USA

[c]Cleveland Clinic, Department of Urology, 9500 Euclid Ave, Cleveland, OH 44106, USA

[d]Louis Stokes Veteran Affairs Medical Center, 10701 East Blvd, Cleveland, OH 44106, USA

Abstract

Deep learning survival (DLS) modeling has shown significant promise across multiple disease conditions for predicting patient outcomes via biomedical data, but remain relatively under-explored in conjunction with medical imaging. This poses the question of what methodological choices promote model generalization, especially when predicting continuous survival outcomes in external validation cohorts. In this work, we systematically examined the impact of data partitioning, data order, model initialization, and augmentation strategies on the generalization of CT-based DLS survival models in renal cancers. A multi-institutional CT cohort was assembled, leveraging public repositories such as TCGA-KIRC, TCGA-KICH, TCGA-KIRP, and KITS19; totaling 525 patients from across 9 institutions. A 3D ResNet-18 model was trained to predict overall patient survival with a Cox loss function and evaluated across three controlled experiments: (i) random versus covariate-balanced data partitioning, (ii) randomizing vs fixing data ordering and initial model weights, and (iii) varying intensity, spatial, and noise based augmentations; evaluated in terms of concordance index (c-index) and hazard ratio (HR). Intelligent covariate-balanced data partitioning demonstrated optimal performance on an external validation set (c-index 0.74, HR 5.08) compared to random partitioning (c-index 0.56, HR 0.29). Different

*Corresponding author. bxf169@case.edu (B. Flannery), satish.viswanath@emory.edu (S.E. Viswanath).

CRediT authorship contribution statement

Brennan Flannery: Writing – review & editing, Writing – original draft, Visualization, Validation, Software, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Thomas DeSilvio:** Visualization, Conceptualization. **Mohsen Hariri:** Conceptualization. **Amir Reza Sadri:** Data curation. **Nicholas Heller:** Conceptualization. **Christopher Weight:** Conceptualization. **Satish E. Viswanath:** Writing – review & editing, Supervision, Resources, Project administration, Methodology, Funding acquisition, Conceptualization.

Ethical statement

This study was considered exempt from IRB review as it used public repositories of de-identified medical images only. No human subjects were directly involved, and no new data collection was performed.

model initializations significantly impacted the average performance and variance across replicate trainings, while data ordering had little to no effect. Combining multiple augmentations produced the best performance on external validation (c-index +4.76%, HR +44.39%) compared to no augmentations, while most impactful individual augmentation type was additive gaussian noise (c-index +3.17%, HR +23.47%). Our results demonstrate that careful selection of strategies for data partitioning and ordering, weight initialization, and image augmentation are critical for developing robust and generalizable DL models for continuous survival prediction using medical imaging data.

Keywords

Deep learning; Survival; Variability; Radiology; CT

1. Introduction

Deep learning (DL) has recently shown significant promise across multiple diseases for predicting patient outcomes via biomedical data by iteratively optimizing and learning features associated with disease characteristics [1]. In the context of prognostic applications, the sub-class of deep learning-based survival (DLS) models have been designed to predict patient events such as mortality or clinical complications. The overarching objective of such models is to transform the clinical workflow by enabling the targeting of intensive therapies to patients identified to be at high risk [2] of such events, typically employing Cox-like loss functions [3] to learn risk scores associated with target outcomes. DLS models output a single continuous risk score per patient, where higher risk scores represent a higher likelihood of early adverse events on a continuous time scale. These models leverage data available for adverse events (e.g., death, recurrence, etc.) as well as time-to-event information, which are used as the training targets to optimize model risk scores for accurately ordering patient events. Similar to Cox proportional hazards models [4], these risk scores are evaluated in terms of (i) how accurately they predict a given event occurring within a specified time interval via the concordance index, and (ii) stratifying patients into separable risk groups based on Kaplan-Meier analysis. Thus far, DLS models have been primarily applied to tabular data, including clinical information [5–10], genomics [6,11,12], and proteomics [6].

By comparison, the utility of deep learning for continuous survival analysis in medical imaging, particularly radiology, remains underexplored. Studies that do investigate prediction of patient survival using medical imaging often reduce problem complexity by considering binary outcomes (e.g., long- or short-term survival [3,6,11,13–17]). This is a notable design shortcoming as continuous outcomes are critical to the performance of DLS models [5–9,11,12,16,18]. In addition to compressing outcomes from continuous to discrete, the binary classification approach fails to address complexities of censored outcomes, which occur when patient status cannot be confirmed after a certain time [19]. A majority of DLS studies in medical imaging also have not specifically used distinct train-test-holdout validation splits [6,8,9,18], instead favoring cross-validation alone [5,10,12,17], or omitting a holdout validation cohort entirely [5,7,10,12,16,17]. While this may be a

function of limited cohort sizes, cross-validation approaches have known limitations in terms of demonstrating the “true” generalizability and robustness of DL models [20,21].

We thus posit that there still remain significant gaps in knowledge with regard to time-to-event modeling via DL as well as confirming the generalizability of DLS models, especially in the context of medical imaging data. While a number of model training paradigms are commonly used in DLS models, there is a lack of clarity given the unique loss functions such models typically employ [22]. The impact of such parameters when considering data from multiple institutions is also relatively unexplored. To our knowledge, the following key factors that influence the performance and generalizability of DLS models have not been widely explored:

- *Data partitioning*: This process dictates how data is split into training, testing, and validation sets. Patients are often randomly assigned to these groups, which could fail to account for variations in acquisition parameters (known to exist in medical imaging cohorts [23]). The resulting sub-optimal distribution and unrepresentative subsets can negatively impact model performance and generalizability [24,25].
- *Data ordering*: The order in which data is presented to the model during training could potentially have an effect on model performance. Similar to partitioning approaches, most studies use random selection when choosing items for batches during neural network training [17,26], and there is limited understanding of how this parameter affects model performance, especially given appearance and batch variations which are known to exist in medical imaging cohorts [27].
- *Model initialization*: Typically performed randomly within predefined ranges [28], this integral part of model training is known to impact model performance significantly, especially in terms of convergence rate [29]. While pretraining and transfer learning approaches have been adopted to provide more optimal starting model weights for training, [30–33], it is not well known how model initialization specifically affects the performance of DLS models.
- *Image augmentations*: Random augmentations are frequently applied during training to enhance model robustness [34]. These augmentations work for medical imaging data by altering image appearance through additive noise or blurring, deformation, or intensity transforms. Critically, the specific effect of each type of augmentation on the validation performance of DLS models remains underexplored for imaging cohorts.

In this study, we present a detailed and comprehensive empirical investigation of deep learning survival models in a multi-institutional setting, in the context of the challenging clinical problem of prognosticating overall survival in renal cancer patients via CT scans. This clinical question is motivated by the fact that the median survival rate for renal cancers is 78%, and there is a wide disparity in 5-year outcomes independent of tumor grade and stage [35]; suggesting the need for more detailed risk stratification approaches. Our study utilizes large publicly available resources of CT imaging and survival outcomes from renal cancer patients [36–40], encompassing a representative heterogenous real-world

cohort encompassing 525 patients, 46 scanner models, and 9 institutions. The specific questions that we attempted to examine were (1) what is the impact of random versus metadata-informed data partitioning strategies on DLS models? (2) what is the interplay between model initialization and data ordering on the training performance of DLS models? and (3) which augmentation strategies can impact DLS model optimization, especially when considering the degree and combinations of augmentations applied? Our goal was to provide actionable insights and recommend data-driven best practices to ensure the robustness and generalizability of DL-based survival models, particularly for clinical applications via medical imaging data.

The rest of the paper is organized as follows: we next describe the overall experimental design of this work (illustrated in Fig. 1), including the data, pre-processing, and model architecture, as well as the experimental methodology for evaluating each aspect of DLS model optimization. Next, our experimental results for multi-institutional benchmarking of DLS models within different optimization setups are presented and discussed, followed by the concluding remarks.

2. Experimental design

2.1. Data description, pre-processing, and kidney localization on CT scans

Pre-operative contrast-enhanced CT scans with death event indicators and followup times were obtained from publicly available repositories, summarized in Table 1. All CT scans first underwent resampling to a uniform resolution of 1 mm × 1 mm × 1 mm to account for voxel size differences between scanners and institutions. Kidney and tumor regions were localized on each CT scan based on an automated segmentation model [41]. Kidney and tumor annotations available in KITS23 [40] were used to train a UNet model for 100 epochs with an 80/20% train-test split, while using Dice loss and an Adam optimizer. Model performance was evaluated via Dice score and intersection-over-union (IoU) within the testing set from KITS23. This optimized U-net model was utilized to segment out kidney and tumor regions on all remaining CT scans prior survival analysis. Further details of U-net model performance and comparisons are presented in Supplementary Figs. 1 and 2, indicating Dice scores of 0.92 for kidney and 0.71 for tumor localization specifically as well as a 98.5% accuracy in identifying bounding boxes for kidney regions. The kidney with the largest tumor mass in each patient was then retained for all further analysis. Based on U-net model outputs, regions of interest (ROIs) were extracted by cropping a 100 × 100 × 100 pixel region around the automatically segmented kidney and tumor regions on each CT scan.

2.2. DLS model architecture

A 3D ResNet-18 architecture was utilized to construct survival models with a single neuron in the final layer which output a continuous risk score. ResNet-18 was chosen due to its broad adoption and parameter efficiency in medical imaging [42], enabling reproducible benchmarking in a data-limited setting. A 3D implementation was specifically used to account for the volumetric nature of CT data utilized here. Models were trained for 200 epochs using the Adam optimizer with early stopping, a Cox proportional hazards-based loss function, model initialization using He initialization, and a learning rate of 0.001. These

choices were driven by their wide usage in the medical imaging domain, thus ensuring any experimental insights can be immediately actionable. Training was conducted on an NVIDIA GeForce RTX 3080 GPU with a batch size of 32. To address the imbalanced nature of survival datasets, a weighted random sampler was utilized during training, ensuring that each batch always included event cases (i.e., non-censored cases), as required by the Cox loss function. For each of the following experiments, risk scores output by the DLS models were used to stratify datasets into high- and low-risk groups based on the median risk score. Model performance was quantified using:

- Concordance index (CI) measures how well a model can predict an order of events. Since an earlier death event is indicative of high risk, CI specifically measures the ability of a model to identify high risk patients. CI ranges from 0 to 1, where 1 indicates perfect ordering of death events, while 0 represents complete misordering.
- Hazard Ratio (HR) compares the frequency of death events in the model predicted high risk group to the model predicted low risk group over a designated time period. A higher HR indicates more and earlier death events in the high risk group compared to the low risk group.

Risk group separation was additionally visualized via Kaplan-Meier curves [43], which convey population based trends in survival. When comparing curves from two populations (e.g., high risk and low risk) a large separation between the curves indicates different survival outcomes between the groups while overlapping curves indicates similar survival outcomes.

2.3. Experiment 1 - evaluating the impact of data partitioning strategies

Two strategies were considered for data partitioning, to determine how best to segregate data into training, testing, and validation cohorts.

1. Random Partitioning: As is typically applied across a majority of deep learning studies [44,45], the entire cohort was randomly segregated such that 60% was allocated to the training set (D_R^{Tr}), 20% to the testing set (D_R^{Te}), and 20% to the validation set (D_R^{V}).
2. Intelligent Partitioning: An intelligent data-partitioning scheme was implemented based on the CohortFinder algorithm [46], which is an open-source data-driven partitioning tool designed to maximally balance batch effects in a cohort to yield more representative data partitions and thus more generalizable DL models. CohortFinder utilizes the output of an open source quality control tool [47], to extract image metadata (e.g., scanner manufacturer, resolution, convolution kernel, exposure time, and tube current), followed by unsupervised clustering to determine batch effect groupings. Six distinct clusters were identified using CohortFinder, and each cluster was proportionally assigned to training (60%, D_{CF}^{Tr}), testing (20%, D_{CF}^{Te}), and validation (20%, D_{CF}^{V}) sets.

Fig. 2 visualizes UMAP embeddings of image metadata to illustrate the presence of institution-specific batch effects in the cohort (where individual institutions cluster

separately from each others), demonstrating that partitions obtained via D_{CF} comprise representative samples from each institution-specific cluster. To evaluate the impact of data partitioning on DLS approaches, a single independent model was trained, tested, and validated for each partitioning scheme, e.g., M_{CF} was trained using D_{CF}^{Tr}, optimized on D_{CF}^{Te} (after each training epoch), with the best performing model validated on D_{CF}^{V}. A similar approach was implemented to train, test, and validate M_R. M_{CF} and M_R were then compared in terms of their performance in prognosticating survival, as described in Section 2.6. To isolate the effects of data partitioning, models were trained with a fixed initialization seed, a fixed data order, and no image augmentations.

2.4. Experiment 2 - evaluating the impact of model initialization and data ordering strategies

Two possibilities were considered for each of (i) initializing the model weights, and (ii) ordering of data being presented to the model during the training process. *Randomization* involved randomly initializing the weights and data ordering at each epoch, while *fixed* involved pre-determining a random seed at the start of training which was used to initialize model weights and data ordering prior to each epoch. Four combinations were considered based on these options:

1. Fixed data ordering and fixed weight initialization.
2. Fixed data ordering and random weight initialization.
3. Random data ordering and fixed weight initialization.
4. Random data ordering and random weight initialization.

For each of the 4 configurations, 50 independent DLS models were trained using D_{CF}^{Tr}, tested on D_{CF}^{Te} after each epoch, and validated on D_{CF}^{V}; for a total of 200 models. Performance was averaged across all 50 models for each configuration, and the distributions of model performances were compared based on c-index and hazard ratio with statistical analysis as described in Section 2.6. To isolate the effects of data ordering and initialization, models were training using intelligent partitioning only as well as no image augmentations.

2.5. Experiment 3 - evaluating the effect of data augmentation

To identify how different augmentation strategies impact model generalizability and stability, three types of augmentations were tested:

1. Intensity Transforms: Random histogram shifting of images to simulate contrast differences. Three levels of shifting were tested by varying the number of control points (3, 7, and 10). These control points define anchor intensities where random shifts are applied, such that a smaller numbers of control points allow for larger intensity shifts.
2. Spatial Transforms: Affine transformations of the images to simulate variations in patient positioning or organ size. Affine transformations required altering three parameters simultaneously, including the rotate range (15°, 30°, 45°), scale range (0.05, 0.15, 0.30), and shear range (0.05, 0.15, 0.30).

3. Artifact-Inducing Transforms: Gaussian noise was added to the scans to mimic acquisition artifacts. Three levels of noise were tested by varying the standard deviation of the additive noise (10, 25, and 50).

Representative images illustrating the effect of each augmentation are depicted in Fig. 3. For each augmentation type, models were trained with each of the three levels of transforms individually; all implemented using the Monai package [48]. Additionally, a combination of all three augmentations was tested to evaluate their cumulative effect on model performance. For each of the 10 configurations considered, 50 independent DLS models were trained using D_{CF}^{Tr}, tested on D_{CF}^{Te} after each epoch, and validated on D_{CF}^{V}; for a total of 500 models. Performance was averaged across all 50 models for each configuration, and the distributions of model performances were compared based on c-index and hazard ratio with statistical analysis as described in Section 2.6. To isolate the effect of image augmentations, all models are trained with intelligent partitioning, fixed initialization, and fixed data order.

2.6. Statistical analysis

Survival group comparisons were performed using the Wilcoxon log-rank test to determine statistical significance in separating risk groups. When comparing multiple configurations of DLS models, one-way ANOVA with post-hoc Tukey tests was used to compare performance with p-values adjusted for multiple comparisons using the Holm-Bonferroni correction. Variance differences between model configurations were assessed using pairwise F-tests, with corrected p-values.

3. Results

3.1. Experiment 1 - evaluating the impact of data partitioning strategies

Fig. 4 visualizes Kaplan-Meier curves obtained via random (M_R) vs CohortFinder (M_{CF})-based training, testing, and validation sets. In training, D_R^{Tr} and D_{CF}^{Tr} achieved similar CIs of 0.98 and 0.99, respectively. Both models also showed strong separation of high- and low-risk groups on their respective testing sets, with M_R achieving a CI of 0.81 and HR of 12.07 and M_{CF} achieving a marginally lower c-index of 0.65 and HR of 4.58. In holdout validation, M_R achieved a markedly lower c-index of 0.564 and a HR of 0.287 compared to M_{CF}, which demonstrated a c-index of 0.74 and HR of 5.08. Kaplan-Meier survival curves comparing predicted low and high risk patients in each set further demonstrate that M_R achieves statistically significant separation on training and testing, but not in holdout validation (Fig. 4A). By contrast, M_{CF} achieves statistically significant separation in all three splits (Fig. 4B). Additionally, M_{CF} resulted in improved performance beyond a baseline clinical-only model using TNM-staging (c-index of 0.62, Supplementary Table 1). Intelligent partitioning was also found to result in a proportional distribution of co-variates such as scanner manufacturer and institution between training, testing, and validation sets (see Supplementary Fig. 5).

3.2. Experiment 2 - evaluating the impact of model initialization and data ordering

Table 2 summarizes averaged model performances for the four different configurations considered, within training, testing, and validation sets (as determined via CohortFinder

partitioning, based on the results of Experiment 1). All configurations of data ordering and initialization were found to yield similar performance within the training set in terms of c-index as well as hazard ratios. By examining box plots and associated statistical testing results in Fig. 5, models using fixed initializations can be seen to yield significantly higher c-indices and HRs compared to random initializations on both testing and validation sets, as determined by ANOVA and post-hoc Tukey tests. Pairwise F-tests further confirmed that models trained with fixed initialization and fixed data ordering demonstrated significantly lower variation in validation c-index and HR compared to models trained with random initialization.

3.3. Experiment 3 - evaluating the effect of data augmentations

Table 3 summarizes the quantitative results in terms of c-index and HR for different configurations of image augmentations considered, with corresponding statistical analyses presented in Fig. 6 and in the Supplementary Materials. In training, increasing the intensity of histogram shifting and affine transforms resulted in a marginally worse but significant reduction in both c-index (Fig. 6A–C) as well as hazard ratio (Supplementary Materials, Fig. 2A–C). However, while random noise augmentations did not impact c-index in the training set (0.99 ± 0.01 c-index for low, medium and high intensity), there was a significant reduction in HR for high intensity noise (96.83 ± 8.57 vs 99.12 ± 9.29). This can be observed when comparing the group means as well as group variances, where more intense augmentations also result in significantly more variable performance. By contrast, increasing the intensity of histogram shifting and gaussian noise did not impact performance within the testing set (Fig. 6D,F and Fig. 2D,F in Supplementary Materials). Only increasing the intensity of affine transformations led to significantly improved c-index (Fig. 6E) and HR (Fig. 2E in Supplementary Materials). No significant differences could be observed in the variability of performance in the testing set across different augmentation configurations. The largest impacts of image augmentations were observed in the validation set. High-intensity histogram shifting resulted in significantly higher c-index (Fig. 6G) and HR (Supplementary Materials, Fig. 2G). By contrast, increasing intensity of noise augmentations resulted in significantly worse c-index (Fig. 6I) and HR (Fig. 2I in Supplementary Materials). No consistent trends were observed when considering affine augmentations, in addition to no significant differences in performance variability across different augmentation configurations within the validation set.

Fig. 7 summarizes performance trends when considering all image augmentations in combination, which yielded the best overall performance in terms of CI and HR in testing (Fig. 7B,E) and validation (Fig. 7C,F), though this was not always significantly different compared to individual augmentation types. Notably, augmenting random noise alone resulted in significantly less variation in HR (Fig. 7D) and CI Fig. 7(A), (B) across training, testing, and validation. Supplementary Fig. 4 additionally summarizes performance trends in terms of CI for every possible permutation of individual augmentations. Any combination of two or more augmenations was found to surpass a baseline model (no augmentations).

Fig. 8 visualizes performance differences of different configurations from Experiments 2 and 3 as a shaded red-green cell map, where minimal performance improvements relative

to a baseline DLS model are shaded yellow. Intuitively, this suggests that the best model performance resulted from a combination of using fixed model initializations, fixed data order, and all augmentation types. The effect of different configurations (particularly augmentations) was more pronounced on the testing set than on the validation set, in terms of both c-index and hazard ratio. Additionally, configurations that resulted in reduced performance on the training set exhibited improved performance on the testing and validation sets.

4. Discussion

Advances in deep learning have led to wider interest in the development of DL-based survival (DLS) models for improved prognostication of patient survival and risk stratification [2], across a number of disease conditions. However, development of DLS models using medical imaging data to predict continuous outcomes remains relatively understudied, with specific gaps in knowledge regarding the choice of strategy for data partitioning, data ordering, model initialization, as well as image augmentations. The objective of our study was to conduct a comprehensive empirical investigation of DLS medical imaging models in a multi-institutional setting, enabling actionable insights and data-driven best practices to ensure their robustness and generalizability.

The specific clinical problem we considered was prognosticating overall survival in renal cancers via CT scans. While there have been some recent studies in this regard using multimodal or pathology data [26,42,49], there is only one other study which has specifically considered DL-based prognostication of kidney cancer survival via CT scans [50]. This approach utilized deep features (learned to predict grade) to optimize a simplified discrete time-survival model within the KITS21 [51] repository. While the performance reported in this study was similar to our results, we were additionally able to comprehensively examine the performance of DLS models on a multi-institutional cohort of data from additional public repositories [36–38,40], in addition to carefully interrogating the impact of different training configurations on DLS model performance.

The major takeaways from our study, with regard to specific training strategy choices, were as follows:

- When evaluating the impact of data partitioning strategies, intelligent data-partitioning via the CohortFinder algorithm [46] significantly enhanced model validation performance compared to randomized partitioning. This performance improvement is consistent with prior studies of classification tasks using this method [46], but our current study is the first indication of its efficacy in the context of survival prognostication. Based on the marked differences and institution-specific clusters observed in our dataset, our results underscore the importance of appropriately accounting for multi-institutional data variation when developing DLS models.
- Data ordering had little to no impact on model performance. Changing from fixed data order to randomized data order resulted in a 0% and 1.5% difference in testing and validation c-index, respectively. Hazard ratio demonstrated similar

trends, with a 2.5% and 2.6% difference on testing and validation, respectively. Variation in both metrics exhibited equally small differences. To our knowledge, there have been no other studies specifically evaluating the effect of data order on model performance in medical imaging or other machine learning contexts.

- Models trained with fixed initializations exhibited 8.5% higher c-indices and 107% higher hazard ratios compared to those with random initializations. They also demonstrated 25% less variation in c-index compared to those with random initializations. Similar observations regarding the importance of model initialization have recently been reported in the context of pretraining approaches in medical imaging [5,30] and multimodal/multi-omic deep learning survival methods [6,52].
- Different augmentation types were associated with markedly different trends in DLS model performance. Higher levels of random noise intensity produced a 3.2% decrease in c-index and 66% decrease in hazard ratio on validation, likely because introducing excessive Gaussian noise may be an unrealistic augmentation in CT scans where Poisson noise is more common [53]. By contrast, increasing the intensity of affine transformations improved validation c-index by 6.8% and hazard ratio by 20%. This may correspond to slight variations in patient positioning or natural differences in organ size and orientation, thus allowing the DLS model to generalize more accurately. Increasing the intensity of histogram shifting improved validation c-index by 4.8% and hazard ratio by 47%, likely due to enabling DLS models to better accommodate a diversity of image appearances. In a prior study of augmentation approaches for pre-training using natural images [54], addition of noise and a subset of affine transformations (rotation) were also found to be among the most impactful augmentations. However, to our knowledge, no previous studies have systematically examined the effect of specific augmentations on medical imaging-based DL model performance, let alone DLS models specifically.
- Applying all augmentations together generally yielded the best performance (+13.56% on testing, +4.76% on validation), which is consistent with current best practices in deep learning [34]. Among the individual augmentations, affine transformations had the greatest impact on testing set c-index (+15.25%), while random noise exerted the most influence on validation c-index (+3.17%). Notably, none of the augmentation types individually improved performance compared to using no augmentations (under identical initialization and data ordering configurations). This finding suggests that individual augmentations are insufficient to produce generalizable models, and that combining augmentations is critical ensuring model generalizability across the diverse sources of variability which may be present within the validation cohort.
- Across the evaluated configurations, we observed an inverse relationship between training performance and performance between internal testing and external validation. For instance, enabling the full augmentation suite increased internal testing performance by 13.56% while reducing training performance by

2.02%. This suggests that the proposed strategies ensure optimal performance in hold-out validation while still mitigating overfitting on the training set. This pattern is consistent with augmentation-induced variability constraining the model toward learning more stable, generalizable image cues that transfer to held-out data, rather than exploiting dataset-specific signals which only yield near-perfect performance on the training set.

We do however note some limitations of our study. We evaluated the effect of specific random processes such as data ordering and model initialization on DLS model optimization, but did not evaluate every possible random process involved in training. However, the factors we have examined here were identified as the key gaps in knowledge when developing medical imaging DLS models based on a comprehensive survey of the literature. Our study made the assumption that batch normalization was required for DLS model optimization as it has been shown to be a key element of model generalizability and removing it has been shown to only hurt model performance [55]. An avenue of future work could be examining different strategies for batch normalization. We also limited our evaluation of augmentations to one common augmentation per major category (intensity transforms, noise/blurring transforms, and spatial transforms). Given the immense number of possible choices in this regard [34], an expanded follow-on study is required to understand exactly how each possible augmentation type affects DLS model performance. For instance, examining the impact of augmentations utilizing Poisson noise or metal artifacts could provide more real-world impacts on DLS performance. Additional aspects which remain to be explored include how different inputs (multi-modal data) and different training paradigms affect the generalization of other model architectures, including vision transformers, DenseNets, or EfficientNets. It is expected that model complexity, alternate loss functions, and training requirements could result in different generalizability trends between DLS models. While not examined here due to data limitations, incorporating clinical and demographic data will also be critical for construction of clinically actionable predictors.

5. Conclusion

In this study, we presented the a detailed and comprehensive empirical investigation of DLS models and their training strategies when considering medical imaging data in a multi-institutional setting. Our study demonstrated that intelligent data-partitioning as well as pretraining methods (such as self-supervised learning or foundation models) are critical to ensure optimal train-test-validation partitioning and appropriate model initialization. Careful selection of image augmentations and their intensity levels are also necessary to enhance model generalizability and improve validation performance. Future directions could include examining image acquisition factors impacting DLS model reproducibility, as well as the interplay of image pre-processing approaches commonly implemented in DL frameworks. This could accelerate the clinical translation and implementation of DL models to be used in decision support and practice.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgement

Research reported in this publication was supported by the National Cancer Institute (1R01CA280981-01A1,1U01CA294415-01A1,1F31CA291057-01A1), the National Institute of Nursing Research (1R01NR019585-01A1), the National Institute of Biomedical Imaging and Bioengineering (1R01EB037526-01, 5T32EB007509-19), the National Heart, Lung, and Blood Institute (1R01HL165218-01A1), the National Science Foundation (Award 2320952), the Veterans Affairs Biomedical Laboratory Research and Development Service (1I01BX006439-01), the DOD Peer Reviewed Cancer Research Program (W81XWH-21-1-0725), the Ohio Third Frontier Technology Validation Fund, the JobsOhio Program, and the Wallace H. Coulter Foundation Program in the Department of Biomedical Engineering at Case Western Reserve University. This work made use of the High Performance Computing Resource in the Core Facility for Advanced Research Computing at Case Western Reserve University. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health, the U.S. Department of Veterans Affairs, the Department of Defense, or the United States Government.

Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests:

Satish Viswanath reports that financial support was provided by National Institutes of Health. Satish Viswanath also reports that financial support was provided by National Institute of Diabetes and Digestive and Kidney Diseases. If there are other authors, they declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

All data used in this study is publicly available on the Imaging Data Commons (IDC). Code used in this study is publicly available and can be accessed at https://github.com/brennanFlannery/kca_survival.

Appendix A.: Supplementary data

Supplementary data to this article can be found online at doi:10.1016/j.compbiomed.2026.111603.

References

[1]. Bakator M, Radosav D, Deep learning and medical diagnosis: a review of literature, Multimodal Technol. Interact 2 (3) (2018) 47, number: 3 Publisher: Multidisciplinary Digital Publishing Institute. 10.3390/mti2030047, https://www.mdpi.com/2414-4088/2/3/47.

[2]. Deepa P, Gunavathi C, A systematic review on machine learning and deep learning techniques in cancer survival prediction, Prog. Biophys. Mol. Biol 174 (2022) 62–71, 10.1016/j.pbiomolbio.2022.07.004, https://www.sciencedirect.com/science/article/pii/S0079610722000761. [PubMed: 35933043]

[3]. Wiegrebe S, Kopper P, Sonabend R, Bischl B, Bender A, Deep learning for survival analysis: a review, Artif. Intell. Rev 57 (3) (2024) 65, 10.1007/s10462-023-10681-3

[4]. Cox DR, Regression models and Life-Tables, J. R. Stat. Soc. Ser. B 34 (2) (1972) 187–202, 10.1111/j.2517-6161.1972.tb00899.x

[5]. Kim DW, Lee S, Kwon S, Nam W, Cha I-H, Kim HJ, Deep learning-based survival prediction of oral cancer patients, Sci. Rep 9 (1) (2019) 6994, publisher: Nature Publishing Group. 10.1038/s41598-019-43372-7 [PubMed: 31061433]

[6]. Vale-Silva LA, Rohr K, Long-term cancer survival prediction using multimodal deep learning, Sci. Rep 11 (1) (2021) 13505, publisher: Nature Publishing Group. 10.1038/s41598-021-92799-4 [PubMed: 34188098]

[7]. Matsuo K, Purushotham S, Jiang B, Mandelbaum RS, Takiuchi T, Liu Y, Roman LD, Survival outcome prediction in cervical cancer: Cox models VS deep-learning model, Am. J. Obstet. Gynecol 220 (4) (2019). e381.1–.e381.14, 10.1016/j.ajog.2018.12.030, https://www.sciencedirect.com/science/article/pii/S0002937818322774.

[8]. She Y, Jin Z, Wu J, Deng J, Zhang L, Su H, Jiang G, Liu H, Xie D, Cao N, Ren Y, Chen C, Development and validation of a deep learning model for non–small cell lung cancer survival, JAMA Network Open 3 (6) (2020) e205842, 10.1001/jamanetworkopen.2020.5842 [PubMed: 32492161]

[9]. Lai Y-H, Chen W-N, Hsu T-C, Lin C, Tsao Y, Wu S, Overall survival prediction of non-small cell lung cancer by integrating microarray and clinical data with deep learning, Sci. Rep 10 (1) (2020) 4679, publisher: Nature Publishing Group. 10.1038/s41598-020-61588-w [PubMed: 32170141]

[10]. Moradmand H, Aghamiri SMR, Ghaderi R, Emami H, The role of deep learning-based survival model in improving survival prediction of patients with glioblastoma, Cancer Med. 10 (20) (2021) 7048–7059, 10.1002/cam4.4230 [PubMed: 34453413]

[11]. Arya N, Saha S, Multi-modal advanced deep learning architectures for breast cancer survival prediction, Knowl.-Based Syst 221 (2021) 106965, 10.1016/j.knosys.2021.106965, https://www.sciencedirect.com/science/article/pii/S0950705121002288.

[12]. Sun T, Wei Y, Chen W, Ding Y, Genome-Wide association study-based deep learning for survival prediction, Stat. Med 39 (30) (2020) 4605–4620, 10.1002/sim.8743 [PubMed: 32974946]

[13]. Wulczyn E, Steiner DF, Moran M, Plass M, Reihs R, Tan F, Flament-Auvigne I, Brown T, Regitnig P, Chen P-HC, Hegde N, Sadhwani A, MacDonald R, Ayalew B, Corrado GS, Peng LH, Tse D, Müller H, Xu Z, Liu Y, Stumpe MC, Zatloukal K, Mermel CH, Interpretable survival prediction for colorectal cancer using deep learning, npj Digital Medicine 4 (1) (2021) 1–13, publisher: Nature Publishing Group. 10.1038/s41746-021-00427-2 [PubMed: 33398041]

[14]. Nam JG, Park S, Park CM, Jeon YK, Chung DH, Goo JM, Kim YT, Kim H, Histopathologic basis for a chest CT deep learning survival prediction model in patients with lung adenocarcinoma, Radiology 305 (2) (2022) 441–451, publisher: Radiological Society of North America. 10.1148/radiol.213262 [PubMed: 35787198]

[15]. Sun L, Zhang S, Chen H, Luo L, Brain tumor segmentation and survival prediction using multimodal MRI scans with deep learning, Front. Neurosci 13, publisher: Frontiers (Aug 2019), 10.3389/fnins.2019.00810

[16]. Wulczyn E, Steiner DF, Xu Z, Sadhwani A, Wang H, Flament-Auvigne I, Mermel CH, Chen P-HC, Liu Y, Stumpe MC, Deep learning-based survival prediction for multiple cancer types using histopathology images, PLOS ONE 15 (6) (2020) e0233678, publisher: Public Library of Science. 10.1371/journal.pone.0233678 [PubMed: 32555646]

[17]. Hao D, Li Q, Feng Q-X, Qi L, Liu X-S, Arefan D, Zhang Y-D, Wu S, SurvivalCNN: a deep learning-based method for gastric cancer survival prediction using radiological imaging data and clinicopathological variables, Artif. Intell. Med 134 (2022) 102424, 10.1016/j.artmed.2022.102424, https://www.sciencedirect.com/science/article/pii/S0933365722001762. [PubMed: 36462894]

[18]. Bychkov D, Joensuu H, Nordling S, Tiulpin A, Kücükel H, Lundin M, Sihto H, Isola J, Lehtimäki T, Kellokumpu-Lehtinen P-L, von Smitten K, Lundin J, Linder N, Outcome and biomarker supervised deep learning for survival prediction in two multicenter breast cancer series, J. Pathol. Inform 13 (2022) 100171, 10.4103/jpi.jpi_29_21, https://www.sciencedirect.com/science/article/pii/S2153353922007659. [PubMed: 35136676]

[19]. Peng Z-H, Huang Z-X, Tian J-H, Chong T, Li Z-L, The application of time-to-event analysis in machine learning prognostic models, J. Transl. Med 22 (1) (2024) 146, 10.1186/s12967-024-04909-1 [PubMed: 38347635]

[20]. Altman DG, Vergouwe Y, Royston P, Moons KGM, Prognosis and prognostic research: validating a prognostic model, BMJ 338 (2009) b605, publisher: British Medical Journal Publishing Group Section: Research Methods & Reporting. 10.1136/bmj.b605 [PubMed: 19477892]

[21]. Riley RD, Ensor J, Snell KIE, Debray TPA, Altman DG, Moons KGM, Collins GS, External validation of clinical prediction models using big datasets from e-health records or IPD meta-analysis: opportunities and challenges, BMJ 353 (2016) i3140, publisher: British Medical Journal Publishing Group Section: Research Methods & Reporting. 10.1136/bmj.i3140 [PubMed: 27334381]

[22]. Katzman JL, Shaham U, Cloninger A, Bates J, Jiang T, Kluger Y, DeepSurv: personalized treatment recommender system using a cox proportional hazards Deep neural network, BMC Med. Res. Methodol 18 (1) (2018) 24, 10.1186/s12874-018-0482-1 [PubMed: 29482517]

[23]. Janowczyk A, Basavanhally A, Madabhushi A, Stain normalization using sparse AutoEncoders (StaNoSA): application to digital pathology, Comput. Med. Imaging Graph 57 (2017) 50–61, 10.1016/j.compmedimag.2016.05.003, https://www.sciencedirect.com/science/article/pii/S0895611116300404. [PubMed: 27373749]

[24]. Howard FM, Dolezal J, Kochanny S, Schulte J, Chen H, Heij L, Huo D, Nanda R, Olopade OI, Kather JN, Cipriani N, Grossman RL, Pearson AT, The impact of site-specific digital histology signatures on deep learning model accuracy and bias, Nature Communications 12 (1) (2021) 4423, publisher: Nature Publishing Group. 10.1038/s41467-021-24698-1, https://www.nature.com/articles/s41467-021-24698-1.

[25]. Kothari S, Phan JH, Stokes TH, Osunkoya AO, Young AN, Wang MD, Removing batch effects from histopathological images for enhanced cancer diagnosis, IEEE J. Biomed. Health Inform 18 (3) (2014) 765–772, conference Name: IEEE Journal of Biomedical and Health Informatics. 10.1109/JBHI.2013.2276766, https://ieeexplore.ieee.org/document/6575092. [PubMed: 24808220]

[26]. Mahootiha M, Qadir HA, Bergsland J, Balasingham I, Multimodal deep learning for personalized renal cell carcinoma prognosis: integrating CT imaging and clinical data, Comput. Methods Programs Biomed 244 (2024) 107978, 10.1016/j.cmpb.2023.107978, https://www.sciencedirect.com/science/article/pii/S0169260723006442. [PubMed: 38113804]

[27]. Lopes AT, de Aguiar E, De Souza AF, Oliveira-Santos T, Facial expression recognition with convolutional neural networks: coping with few data and the training sample order, Pattern Recognition 61 (2017) 610–628, 10.1016/j.patcog.2016.07.026, https://www.sciencedirect.com/science/article/pii/S0031320316301753.

[28]. Narkhede MV, Bartakke PP, Sutaone MS, A review on weight initialization strategies for neural networks, Artif. Intell. Rev 55 (1) (2022) 291–322, 10.1007/s10462-021-10033-z

[29]. Krähenbühl P, Doersch C, Donahue J, Darrell T, Data-dependent Initializations of Convolutional Neural Networks, arXiv:1511.06856 Sep 2016, 10.48550/arXiv.1511.06856

[30]. Wen Y, Chen L, Deng Y, Zhou C, Rethinking pre-training on medical imaging, J. Vis. Commun. Image Represent 78 (2021) 103145, 10.1016/j.jvcir.2021.103145, https://www.sciencedirect.com/science/article/pii/S1047320321000894

[31]. Ray I, Raipuria G, Singhal N, Rethinking Imagenet pre-training for computational histopathology, in: 2022 44th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), 2022, pp. 3059–3062, 10.1109/EMBC48229.2022.9871687

[32]. Zhao Z, Alzubaidi L, Zhang J, Duan Y, Gu Y, A comparison review of transfer learning and self-supervised learning: definitions, applications, advantages and limitations, Expert Syst. Appl 242 (2024) 122807, 10.1016/j.eswa.2023.122807, https://www.sciencedirect.com/science/article/pii/S0957417423033092.

[33]. Rani V, Nabi ST, Kumar M, Mittal A, Kumar K, Self-supervised learning: a succinct review, Arch. Comput. Methods Eng 30 (4) (2023) 2761–2775, 10.1007/s11831-023-09884-2 [PubMed: 36713767]

[34]. Shorten C, Khoshgoftaar TM, A survey on image data augmentation for deep learning, J. Big Data 6 (1) (2019) 60, 10.1186/s40537-019-0197-0

[35]. Survival rates for kidney cancer Jan, 2024, https://www.cancer.org/cancer/types/kidney-cancer/detection-diagnosis-staging/survival-rates.html.

[36]. Linehan WM, Gautam R, Sadow CA, Levine S, The Cancer Genome Atlas Kidney Chromophobe Collection (Tcga-Kich) (Version 3), the Cancer Imaging Archive [Data set], 2016, 10.7937/K9/TCIA.2016.YU3RBCZN

[37]. Akin O, Elnajjar P, Heller M, Jarosz R, Erickson BJ, Kirk S, Lee Y, Linehan WM, Gautam R, Vikram R, Garcia KM, Roche C, Bonaccio E, Filippini J, The Cancer Genome Atlas Kidney Renal Clear Cell Carcinoma Collection (Tcga-Kirc) (Version 3), the Cancer Imaging Archive [Data set], 2016, 10.7937/K9/TCIA.2016.V6PBVTDR

[38]. Linehan WM, Gautam R, Kirk S, Lee Y, Roche C, Bonaccio E, Filippini J, Rieger-Christ K, Lemmerman J, Jarosz R, The Cancer Genome Atlas Cervical Kidney Renal Papillary Cell Carcinoma Collection (Tcga-Kirp) (Version 4), the Cancer Imaging Archive [Data set], 2016, 10.7937/K9/TCIA.2016.ACWOGBEF

[39]. Heller N, Sathianathen N, Kalapara A, Walczak E, Moore K, Kaluzniak H, Rosenberg J, Blake P, Rengel Z, Oestreich M, Dean J, Tradewell M, Shah A, Tejpaul R, Edgerton Z, Peterson M, Raza S, Regmi S, Papanikolopoulos N, Weight C, Data from C4kc-Kits, the Cancer Imaging Archive [Data set], 2019, 10.7937/TCIA.2019.IX49E8NX

[40]. Heller N, Pietrowski M, Beers A, Karkada D, DeVries P, Weight CJ, Papanikolopoulos N, Kits23: kidney tumor segmentation challenge 2023 dataset, 2023, https://kits-challenge.org/kits23/(accessed 12 August 2025).

[41]. Ronneberger O, Fischer P, Brox T, U-Net: Convolutional Networks for Biomedical Image Segmentation, [cs] arXiv:1505.04597, May 2015.

[42]. Schulz S, Woerl A-C, Jungmann F, Glasner C, Stenzel P, Strobl S, Fernandez A, Wagner D-C, Haferkamp A, Mildenberger P, Roth W, Foersch S, Multimodal deep learning for prognosis prediction in renal cancer, Front. Oncol 11, publisher: Frontiers (Nov 2021), 10.3389/fonc.2021.788740

[43]. Goel MK, Khanna P, Kishore J, Understanding survival analysis: Kaplan-Meier estimate, Int. J. Ayurveda Res 1 (4) (2010) 274–278, 10.4103/0974-7788.76794, https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3059453/ [PubMed: 21455458]

[44]. Peng Y, Nagata MH, An empirical overview of nonlinearity and over-fitting in machine learning using COVID-19 data, Chaos Solitons Fractals 139 (2020) 110055, 10.1016/j.chaos.2020.110055, https://www.sciencedirect.com/science/article/pii/S0960077920304525 [PubMed: 32834608]

[45]. Montesinos López OA, Montesinos López A, Crossa J, Overfitting, model tuning, and evaluation of prediction performance, In Montesinos López OA, Montesinos López A, Crossa J (Eds.), Multivariate Statistical Machine Learning Methods for Genomic Prediction, Springer International Publishing, Cham, 2022, pp. 109–139, 10.1007/978-3-030-89010-0_4

[46]. Fan F, Martinez G, DeSilvio T, Shin J, Chen Y, Jacobs J, Wang B, Ozeki T, Lafarge MW, Koelzer VH, Barisoni L, Madabhushi A, Viswanath SE, Janowczyk A, CohortFinder: an open-source tool for data-driven partitioning of digital pathology and imaging cohorts to yield robust machine-learning models, npj Imaging 2 (1) (2024) 1–7, publisher: Nature Publishing Group. 10.1038/s44303-024-00018-2 [PubMed: 40604134]

[47]. Sadri AR, Janowczyk A, Zhou R, Verma R, Beig N, Antunes J, Madabhushi A, Tiwari P, Viswanath SE, Technical note: MRQy - an open-source tool for quality control of MR imaging data, Med. Phys 47 (12) (2020) 6029–6038, 10.1002/mp.14593 [PubMed: 33176026]

[48]. Cardoso MJ, Li W, Brown R, Krishna M, Graff M, Sudre C, Moinuddin SA, Mitcheltree C, Pires R, Jorge Cardoso M, et al. , Monai: An open-source framework for deep learning in healthcare, arXiv preprint arXiv:2211.02701, 2022.

[49]. Gao R, Qin H, Lin P, Ma C, Li C, Wen R, Huang J, Wan D, Wen D, Liang Y, Huang J, Li X, Wang X, Chen G, He Y, Yang H, Development and validation of a radiomic nomogram for predicting the prognosis of kidney renal clear cell carcinoma, Front. Oncol 11, publisher: Frontiers (Jul 2021), 10.3389/fonc.2021.613668

[50]. Mahootiha M, Qadir HA, Aghayan D, Fretland AA, Edwin BVG, Balasingham I, Deep learning-assisted survival prognosis in renal cancer: a CT scan-based personalized approach, Heliyon 10 (2), publisher: Elsevier (Jan 2024), 10.1016/j.heliyon.2024.e24374, https://www.cell.com/heliyon/abstract/S2405-8440(24)00405-5

[51]. Heller N, Isensee F, Trofimova D, Tejpaul R, Zhao Z, Chen H, Wang L, Golts A, Khapun D, Shats D, Shoshan Y, Gilboa-Solomon F, George Y, Yang X, Zhang J, Zhang J, Xia Y, Wu M, Liu Z, Walczak E, McSweeney S, Vasdev R, Hornung C, Solaiman R, Schoephoerster J, Abernathy B, Wu D, Abdulkadir S, Byun B, Spriggs J, Struyk G, Austin A, Simpson B, Hagstrom M,

Virnig S, French J, Venkatesh N, Chan S, Moore K, Jacobsen A, Austin S, Austin M, Regmi S, Papanikolopoulos N, Weight C, The KiTS21 Challenge: Automatic segmentation of kidneys, renal tumors, and renal cysts in corticomedullary-phase CT, [cs] arXiv:2307.01984, Jul 2023.

[52]. Azher ZL, Suvarna A, Chen J-Q, Zhang Z, Christensen BC, Salas LA, Vaickus LJ, Levy JJ, Assessment of emerging pretraining strategies in interpretable multimodal deep learning for cancer prognostication, BioData Min. 16 (1) (2023) 23, 10.1186/s13040-023-00338-w [PubMed: 37481666]

[53]. Gibson NM, Lee A, Bencsik M, A practical method to simulate realistic reduced-exposure CT images by the addition of computationally generated noise, Radiol. Phys. Technol 17 (1) (2024) 112–123, 10.1007/s12194-023-00755-w [PubMed: 37955819]

[54]. Chen T, Kornblith S, Norouzi M, Hinton G, A Simple Framework for Contrastive Learning of Visual Representations, [cs] arXiv:2002.05709, Jul 2020.

[55]. Khan A, Sohail A, Zahoora U, Qureshi AS, A survey of the recent architectures of deep convolutional neural networks, Artif. Intell. Rev 53 (8) (2020) 5455–5516, 10.1007/s10462-020-09825-6

HIGHLIGHTS

- We used an extensive cohort of 525 patients from 4 publicly available collections, 9 institutions, and 46 types of scanners.
- Intelligent covariate-balanced partitioning improved external generalization (c-index 0.74, HR 5.08)
- Model initialization significantly influenced performance variability, while data ordering had little effect.
- A variety of augmentations had the greatest impact on validation performance (c-index +4.76%, HR +44.39%).

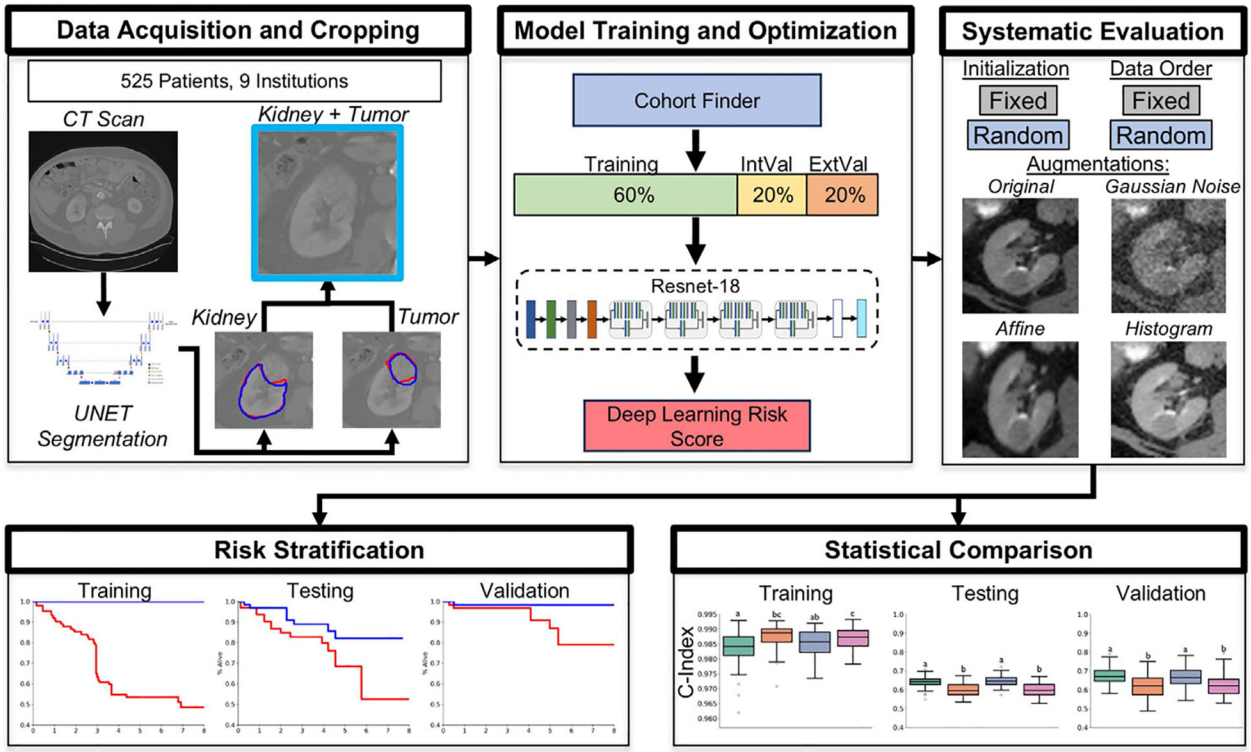

Fig. 1.

Overall experimental workflow of including image data pre-processing, model training, parameter evaluation, and statistical analysis of prognosticating survival for risk stratification.

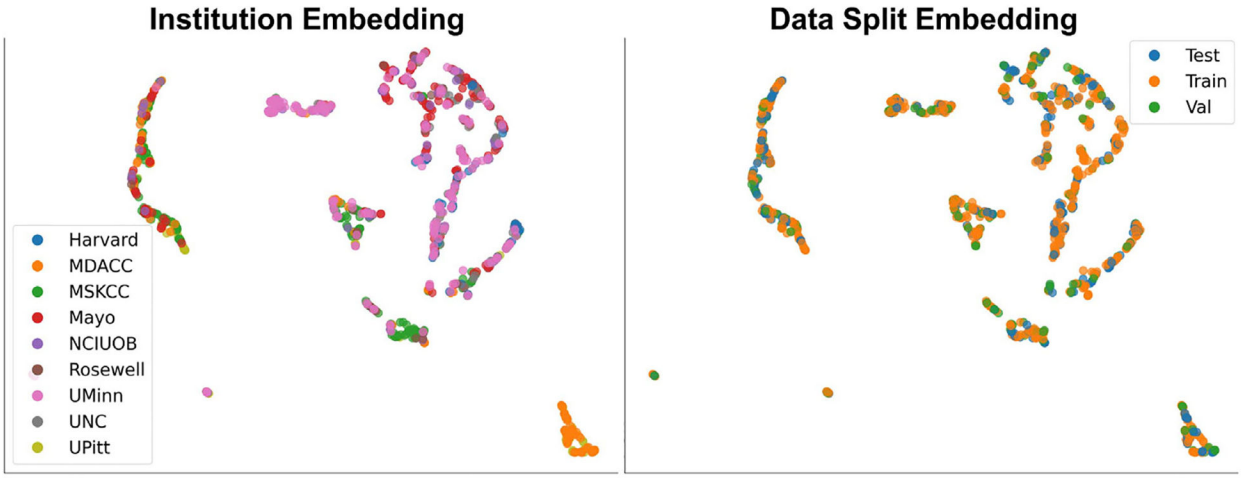

Fig. 2.
2D UMAP embeddings of data using image scanner parameters.

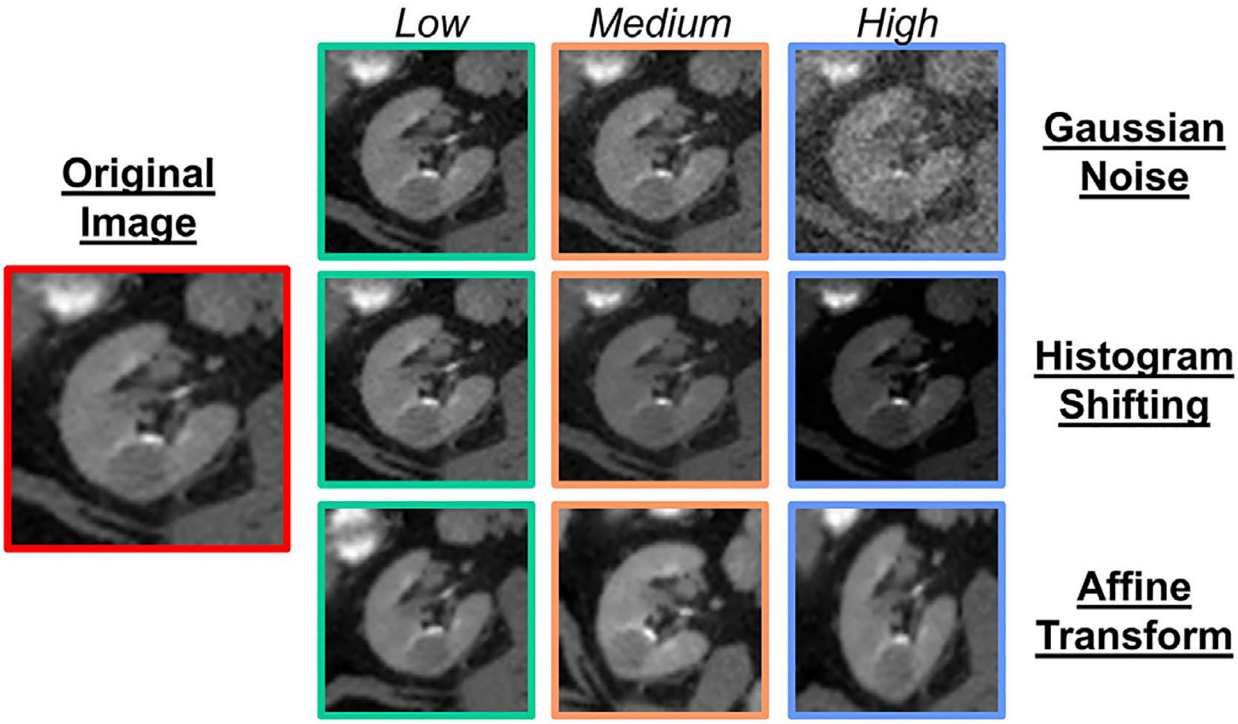

Fig. 3.

Representative images illustrating the effect of each level of augmentation applied to a sample kidney CT scan (localized to the kidney ROI alone).

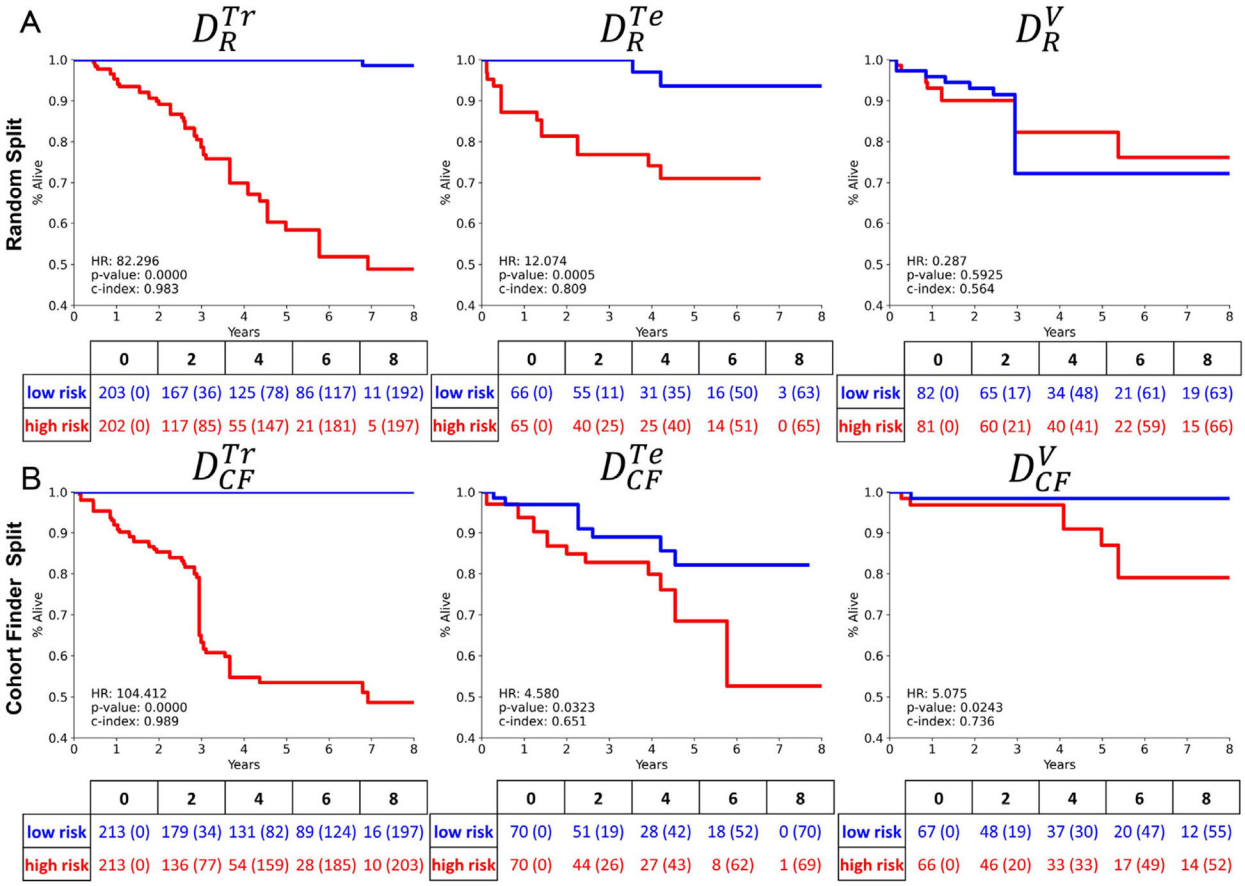

Fig. 4.

Survival curves for training, testing, and validation datasets for different data partitioning strategies, for (A) M_R (random partitioning, top row), and (B) M_{CF} (CohortFinder partitioning, bottom row). Tables indicate the number of patients alive and the number of patients dead or censored at that time point, note there may be numerical differences between models as a result of the partitioning approach. M_{CF} achieves statistically significant separation in all three splits, while M_R can be seen demonstrate significant differences in training and testing but not in holdout validation.

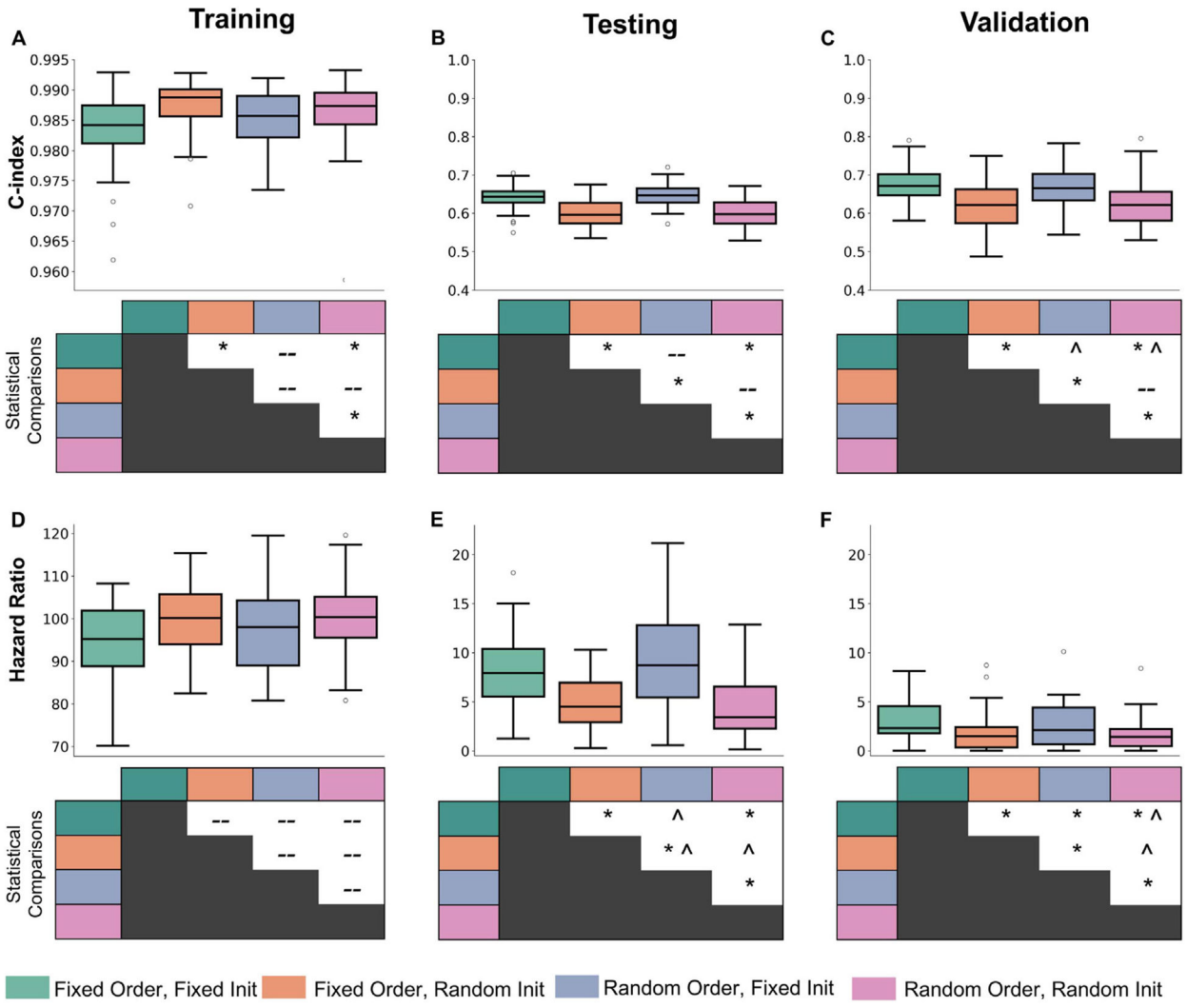

Fig. 5.

Box plots of model metrics at different configurations of data ordering and model weight initialization in DLS model training. These configurations include fixed data order (fixed order), random data order (random order), fixed model weight initialization (fixed init), and random model weight initialization (random init). Statistical analyses are presented in tables below their respective box plots. * indicates significant differences between group means (ANOVA and Tukey Tests), and ^ indicates significant differences between group variances (F-test); based on Bonferroni-corrected p-values.

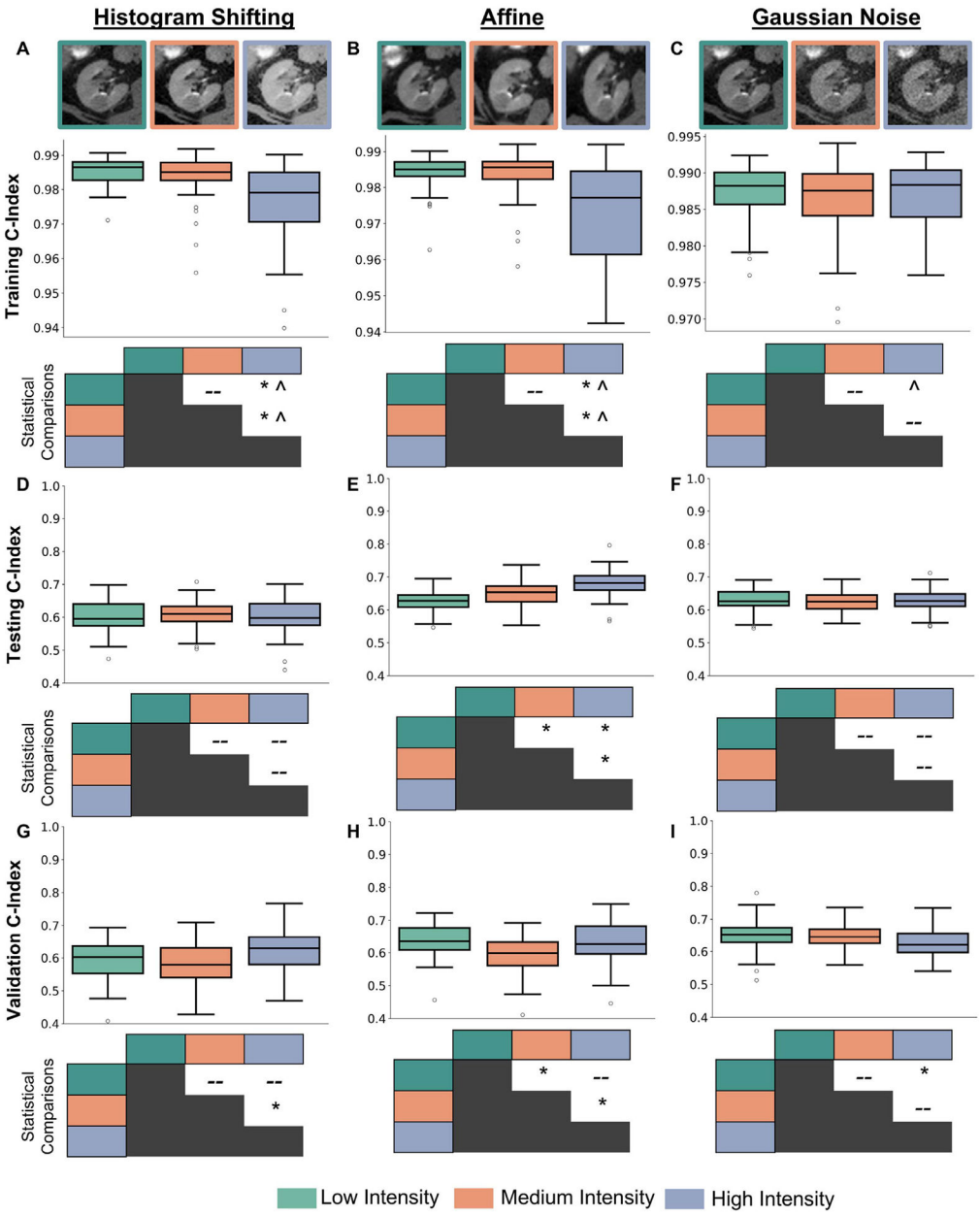

Fig. 6.

Boxplots of c-index achieved by DLS models using different configurations of image augmentations during model training. Statistical analyses are presented in tables below their respective box plots. * indicates significant differences between group means (ANOVA and Tukey Tests), and ^ indicates significant differences between group variances (F-test); based on Bonferroni-corrected p-values.

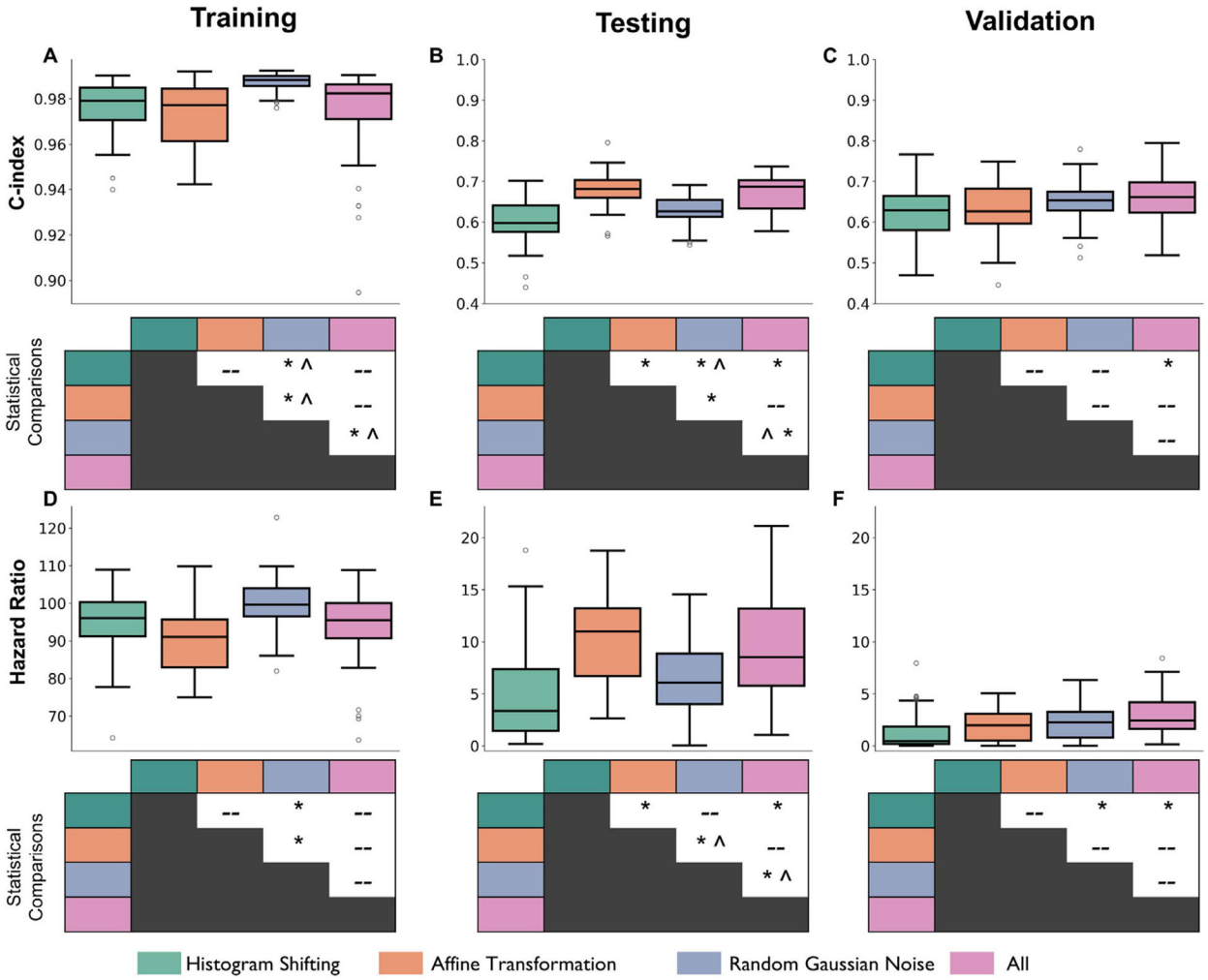

Fig. 7.

Boxplots of c-index and hazard ratio for individual augmentation types as well as for using all augmentations in combination. Statistical analyses are presented in tables below their respective box plots. * indicates significant differences between group means (ANOVA and Tukey Tests), and ^ indicates significant differences between group variances (F-test).

	Initialization	Random	Random	Fixed	Fixed	Fixed	Fixed	Fixed	Fixed
	Order	Random	Fixed	Random	Fixed	Fixed	Fixed	Fixed	Fixed
	Augmentation	None	None	None	None	Histogram	Affine	Noise	All
C-index	Train	0.00%	0.00%	0.00%	0.00%	-1.01%	-2.02%	0.00%	-2.02%
	Test	0.00%	+1.69%	+8.47%	+8.47%	+1.69%	+15.25%	+6.78%	+13.56%
	Validation	0.00%	-3.17%	+4.76%	+6.35%	-1.59%	0.00%	+3.17%	+4.76%
Hazard Ratio	Train	0.00%	+0.22%	-1.08%	-2.82%	-4.46%	-8.81%	+0.41%	-4.95%
	Test	0.00%	+4.72%	+102.36%	+107.61%	+22.83%	+170.34%	+71.13%	+147.51%
	Validation	0.00%	-20.92%	+36.73%	+33.16%	-29.59%	+2.55%	+23.47%	+44.39%

Fig. 8.

Configuration performance improvements compared to a baseline random initialization, random order, no augmentation model (corresponding to the first column). Green indicates increased relative model performance, red indicates a decrease in relative model performance, and yellow indicates little change in relative model performance.

Table 1

Summary of publicly available multi-institutional cohorts of renal CT scans used in this study. Acronyms for contrast phases: Art=Arterial Contrast Phase, Neph=Nephrogenic Contrast Phase, Del=Delay, Ven=Venous Contrast Phase.

Dataset	Sites	Task	Contrast phases	Patients (deaths)	Scans used
KITS23 [40]	1	Kidney segmentation	Art, Neph	489 (N/A)	489
KITS19 [39]	1		Art	210 (21)	210
KIRC [37]	8	Survival prediction	Art, Ven, Neph, Del, Unknown	267 (84)	728
KICH [36]	3		Art, Ven, Neph, Del, Unknown	15 (1)	29
KIRP [38]	2		Art, Ven, Neph, Del, Unknown	33 (2)	48

Table 2

Mean and standard deviation of survival model performance metrics for different configurations of data ordering and model weight initialization. Detailed statistical analyses of these results are presented in Fig. 5.

Fixed order	Fixed weights	Train	Test	Val
		C-Index		
Yes	Yes	0.99 ±0.00	0.64 ±0.03	0.67 ±0.05
Yes	No	0.99 ±0.00	0.60 ±0.04	0.61 ±0.05
No	Yes	0.99 ±0.00	0.64 ±0.03	0.68 ±0.04
No	No	0.99 ±0.00	0.59 ±0.04	0.63 ±0.06
		Hazard Ratio		
Yes	Yes	96.02 ±8.66	7.91 ±4.15	2.61 ±1.77
Yes	No	99.03 ±8.39	3.99 ±2.66	1.55 ±1.79
No	Yes	97.74 ±8.57	7.71 ±4.39	2.68 ±2.00
No	No	98.81 ±7.76	3.81 ±2.99	1.96 ±1.69

Table 3

Mean and standard deviation of survival model performance metrics for different intensity levels of image augmentations. Detailed statistical analyses of these results are presented in Figs. 6 and 7 in addition to the Supplementary Materials.

Transform	Intensity	Train	Test	Val
		C-Index		
Histogram Shifting	Low	0.99 ±0.00	0.60 ±0.05	0.59 ±0.06
	Medium	0.98 ±0.01	0.61 ±0.04	0.58 ±0.06
	High	0.98 ±0.01	0.60 ±0.05	0.62 ±0.06
Affine	Low	0.98 ±0.00	0.63 ±0.03	0.64 ±0.05
	Medium	0.98 ±0.01	0.65 ±0.04	0.59 ±0.06
	High	0.97 ±0.01	0.68 ±0.04	0.63 ±0.06
Noise	Low	0.99 ±0.00	0.63 ±0.03	0.65 ±0.05
	Medium	0.99 ±0.01	0.62 ±0.03	0.65 ±0.04
	High	0.99 ±0.00	0.63 ±0.03	0.63 ±0.04
All	–	0.97 ±0.02	0.67 ±0.04	0.66 ±0.06
		Hazard Ratio		
Histogram Shifting	Low	100.79 ±7.59	5.21 ±4.07	0.73 ±0.75
	Medium	100.32 ±7.43	5.27 ±3.38	0.89 ±1.13
	High	94.40 ±8.91	4.68 ±4.13	1.38 ±1.78
Affine	Low	98.05 ±9.17	5.53 ±3.14	1.91 ±1.67
	Medium	97.73 ±7.25	7.32 ±3.52	1.61 ±1.12
	High	90.10 ±8.05	10.30 ±4.59	2.01 ±1.64
Noise	Low	99.22 ±7.07	6.52 ±3.61	2.42 ±1.74
	Medium	99.12 ±9.29	6.59 ±4.10	2.08 ±1.91
	High	96.83 ±8.57	7.52 ±3.67	1.46 ±1.47
All	–	93.92 ±9.69	9.43 ±4.77	2.83 ±1.92