Artificial Intelligence In Medicine 160 (2025) 103055

Contents lists available at ScienceDirect

Artificial Intelligence In Medicine

journal homepage: www.elsevier.com/locate/artmed

Research paper

# Concordance-based Predictive Uncertainty (CPU)-Index: Proof-of-concept with application towards improved specificity of lung cancers on low dose screening CT

Yuqi Wang [a], Aarzu Gupta [a], Fakrul Islam Tushar [a], Breylon Riley [b], Avivah Wang [c], Tina D. Tailor [d], Stacy Tantum [a], Jian-Guo Liu [e], Mustafa R. Bashir [d,f], Joseph Y. Lo [a,b,d,g], Kyle J. Lafata [a,b,d,g,h,i,*]

[a] Department of Electrical and Computer Engineering, Duke University, Durham, NC, United States of America
[b] Medical Physics Graduate Program, Duke University, Durham, NC, United States of America
[c] School of Medicine, Duke University, Durham, NC, United States of America
[d] Department of Radiology, Duke University, Durham, NC, United States of America
[e] Departments of Physics and Mathematics, Duke University, Durham, NC, United States of America
[f] Department of Medicine, Division of Gastroenterology, Duke University, Durham, NC, United States of America
[g] Department of Biomedical Engineering, Duke University, Durham, NC, United States of America
[h] Department of Radiation Oncology, Duke University, Durham, NC, United States of America
[i] Department of Pathology, Duke University, Durham, NC, United States of America

## ARTICLE INFO

*Keywords:*
Lung cancer screening
Subgroup analysis
Personalized AI
Time-to-event
Personalized uncertainty quantification

## ABSTRACT

In this paper, we introduce a novel concordance-based predictive uncertainty (CPU)-Index, which integrates insights from subgroup analysis and personalized AI time-to-event models. Through its application in refining lung cancer screening (LCS) predictions generated by an individualized AI time-to-event model trained with fused data of low dose CT (LDCT) radiomics with patient demographics, we demonstrate its effectiveness, resulting in improved risk assessment compared to the Lung CT Screening Reporting & Data System (Lung-RADS). Subgroup-based Lung-RADS faces challenges in representing individual variations and relies on a limited set of predefined characteristics, resulting in variable predictions. Conversely, personalized AI time-to-event models are hindered by transparency issues and biases from censored data. By measuring the prediction consistency between subgroup analysis and AI time-to-event models, the CPU-Index framework offers a nuanced evaluation of the bias–variance trade-off and improves the transparency and reliability of predictions. Consistency was estimated by the concordance index of subgroup analysis-based similarity rank and model prediction similarity rank. Subgroup analysis-based similarity loss was defined as the sum-of-the-difference between Lung-RADS and feature-level 0-1 loss. Model prediction similarity loss was defined as squared loss. To test our approach, we identified 3,326 patients who underwent LDCT for LCS from 1/1/2015 to 6/30/2020 with confirmation of lung cancer on pathology within one year. For each LDCT image, the lesion associated with a Lung-RADS score was detected using a pretrained deep learning model from Medical Open Network for AI (MONAI), from which radiomic features were extracted. Radiomics were optimally fused with patient demographics via a positional encoding scheme and used to train a neural multi-task logistic regression time-to-event model that predicts malignancy. Performance was maximized when radiomics features were fused with positionally encoded demographic features. In this configuration, our algorithm raised the AUC from $0.81 \pm 0.04$ to $0.89 \pm 0.02$. Compared to standard Lung-RADS, our approach reduced the False-Positive-Rate from $0.41 \pm 0.02$ to $0.30 \pm 0.12$ while maintaining the same False-Negative-Rate. Our methodology enhances lung cancer risk assessment by estimating prediction uncertainty and adjusting accordingly. Furthermore, the optimal integration of radiomics and patient demographics improved overall diagnostic performance, indicating their complementary nature.

* Corresponding author.
*E-mail addresses:* yuqi.wang@duke.edu (Y. Wang), kyle.lafata@duke.edu (K.J. Lafata).

https://doi.org/10.1016/j.artmed.2024.103055
Received 9 August 2024; Received in revised form 5 December 2024; Accepted 9 December 2024
Available online 16 December 2024

## 1. Introduction

In medicine, subgroup analysis plays a crucial role across prevention science, intervention research, and early diagnosis efforts [1–3]. In early diagnosis, subgroup analysis allows for examining whether screening tests or diagnostic procedures differentially predict disease status or identify early symptoms across groups defined by demographics, risk factors, biomarkers, or other characteristics. This can inform clinical decision-making and guide targeted treatments for specific subpopulations of patients. Fundamentally, subgroup analysis evaluates whether and how the effects of an intervention or the predictive value of a diagnostic test differ within and between groups defined by baseline characteristics. For example, in patients at risk for lung cancer, such characteristics may include high-level demographic variables, or more complex representations of lung nodule morphology on imaging. Subgroup analysis can be categorical or continuous, and can range from frequently occurring to harder to capture in large samples. Despite its broad application, subgroup analysis has many technical challenges in its implementation (e.g., limitations in capturing individual-level heterogeneity which leads to high variance, high reliance on limited predefined features, low granularity, etc.).

A particularly important area of healthcare that requires robust subgroup analysis is lung cancer screening (LCS). Lung cancer is a major cause of cancer-related deaths [4] and therefore LCS on chest low-dose CT (LDCT) is essential to early detection and prevention. Early implementation of LDCT LCS programs has been a success, most notably the National Lung Screening Trial (NLST), which demonstrated diagnostic benefit over radiography and reported a 20% decrease in lung cancer mortality [5]. However, compared to other image-based screening modalities (e.g., mammography), LDCT diagnostic lexicons are relatively immature and the current Lung Imaging Reporting and Data System (Lung-RADS) may not fully capture the screened population [6]. Thus, despite the proven benefit of LDCT-based LCS, Lung-RADS is an evolving process and efforts to improve its sensitivity and specificity are paramount.

Due to the time dynamics of screening problems, such as LCS, robust time-to-event analysis (i.e., statistical methods used to analyze the time until a specific event of interest occurs) is essential to understanding the timing and probability of events (e.g., diagnosis of lung cancer). In general, time-to-event analyses have wide-reaching applications in medicine and healthcare [7–9] and may guide informed decision support for improved patient care. Furthermore, time-to-event prediction models provide a novel framework for complex screening problems, aiming to improve the accuracy of predicting future events, such as lung cancer. In recent years, the emergence of AI has significantly further enhanced these methods, enabling more sophisticated and accurate time-to-event analyses [10–13].

However, despite their improvements, AI-driven time-to-event models still face significant challenges of uncertainty. The uncertainty stems mainly from two sources. First, there is a lack of interpretability. While AI models can handle complex, high-dimensional data better than simpler methods, they often act like black boxes. This makes it non-trivial for clinicians to understand how decisions are made, potentially limiting their use in patient care. Post-hoc interpretation methods like SHAP [14], are frequently used in time-to-event predictions to enhance interpretability due to its model-agnosticism [15]. However, these methods focus solely on the AI model's perspective, leaving uncertainties unaddressed and not accounting for to integrate complementary information from other sources. Second, these models struggle with censored data (i.e., incomplete information in medical studies when patients drop out or are lost to follow-up). While modern models, such as DeepSurv [10] and its derivatives (e.g., neural multi-task linear regression (NMTLR) [12]), contain specific design choices to handle censored data effectively, certain limitations persist in scenarios involving heavily censored datasets, long-term outcomes, and/or rare events (e.g., LCS), where biases and reduced reliability can impact performance. These limitations underscore the critical need for Uncertainty Quantification (UQ) techniques in time-to-event models.

As such, researchers have proposed various approaches to address uncertainty from all sources in time-to-event problems. Non-sampling approaches, such as García-Donato et al.'s Bayesian Cox regression [16], and Monte Carlo sampling methods, including Loya et al.'s deep learning-based Bayesian framework [17] and Sokota et al.'s personalized uncertainty representation [18], have shown promise. Additionally, methods exploring intrinsic uncertainty have further advanced the field: Chapfuwa et al.'s adversarial nonparametric models improve calibration and concentration of time-to-event distributions [19], Dubey et al.'s Bayesian Neural Hawkes Process enables uncertainty-aware event prediction [20], and Huh et al.'s joint prognostic frameworks integrate longitudinal and time-to-event uncertainties [21]. However, these approaches still face notable limitations: García-Donato's and Huh et al.'s methods have limited applicability to other parametric survival models, while Chapfuwa et al.'s and Dubey et al.'s techniques are restricted to deep neural network-based models. Sokota's approach is confined to parametric models, and Loya's and Jacob's methods, while capable of identifying out-of-distribution cohorts, fail to provide UQ for individual cases. Furthermore, techniques like those proposed by Chapfuwa et al. [19] and Dubey et al. [20] lack seamless applicability to real-time clinical scenarios. Consequently, model-agnostic, explainable methods are needed to analyze the uncertainty of individual patient predictions in the absence of ground truth—an essential requirement for clinically relevant, real-world applications aimed at prospectively monitoring patients throughout their care.

To address that need, we propose a novel concordance-based predictive uncertainty (CPU)-Index framework for survival analysis that is capable of estimating the uncertainty of individual predictions. The proposed framework addresses limitations of variance vs. bias for both subgroup analysis and personalized AI time-to-event models. We hypothesize that prediction uncertainty for a test patient during inference increases with lower consistency in feature-space-similarity between the test patient and training set patients, as observed across both methodologies. To test this hypothesis, we propose to represent the uncertainty for a particular patient-of-interest by a modified concordance index between a subgroup analysis-based similarity rank and a model prediction similarity rank. Patients are then split into batches to calculate the concordance index and reduce the effect of noise within the outcomes data [22].

To evaluate the efficacy of our proposed approach, we employed the CPU-Index to refine the LCS prediction generated by an individualized AI time-to-event model, specifically, an NMTLR model [12]. In addition, we introduced a positional encoding (PE)-based imaging-EHR fusion technique [23] to further boost model performance. Specifically, we defined subgroup analysis-based similarity loss as the sum-of-the-difference between the Lung-RADS score and the feature-level 0-1 loss, and model prediction similarity loss as squared loss. Subsequently, we computed the CPU-Index for each case in the test dataset. A case is classified as unlikely to receive a cancer diagnosis within one year of the scan if both the model assigns a low-risk prediction and the CPU-Index is low. Our findings indicate that incorporating the CPU-Index improves LCS predictions compared to relying solely on the Lung-RADS score or the NMTLR model.

The structure of the following sections is as follows: Section 2 introduces prior work used in our experiments. In Section 3, we present our proposed CPU-Index methodology. Section 4 introduces the experimental setup and data materials, and reports the findings. The analysis and discussion of the results are reported in Section 5. Finally, Section 6 provides a summary of the key conclusions drawn from the study.

## 2. Prior work

### 2.1. Data fusion quantification and optimization via positional encoding

In medicine, diverse data types such as imaging (CT, MR, etc.), numerical values like lab results are considered throughout the complete healthcare spectrum. These varied data types offer complementary insights, enhancing the overall understanding of the condition. Consequently, the development of computer-assisted models typically involves the integration of diverse data types. Achieving optimal integration requires robust data fusion—a process that combines multiple data sources to generate information that is more consistent, accurate, and valuable than what individual sources provide. This consideration is common in medical imaging scenarios, where the dominance of high-dimensional imaging features over low-dimensional clinical features can result in unequal contributions from individual data sources to the model. Positional encoding has proved to be effective as a vector-growing scheme to increase the dimension of low-dimension data to minimize the source bias [23,24].

The fused feature can be represented as a potential function with a distribution resembling the classical Gibbs measure, conceptualized as random variables influenced by state functions, with these functions representing distinct data sources, i.e., CT imaging, $x = \{x_i\}_{i=1}^{d_1} \in \mathbb{R}^{d_1}$, vs. EHR data, $y = \{y_j\}_{j=1}^{d_2} \in \mathbb{R}^{d_2}$. Using a kernel density estimation technique, the marginal contribution of each source can be quantified as radial basis functions,

$$\phi_1(x) = \frac{1}{d_1}\sum_{i=1}^{d_1} e^{-\frac{1}{2\sigma^2}[x-x_i]^2}, \tag{1}$$

$$\phi_2(y) = \frac{1}{d_2}\sum_{j=1}^{d_2} e^{-\frac{1}{2\sigma^2}[y-y_j]^2}, \tag{2}$$

where $\phi_i$ and $\phi_j$ represent the relative contribution of $x$ and $y$, respectively, to the density of the fused feature,

$$\phi_{12}(x,y) = \alpha\sum_{i=1}^{d_1} e^{-\frac{1}{2\sigma^2}[x-x_i]^2} + \beta\sum_{j=1}^{d_2} e^{-\frac{1}{2\sigma^2}[y-y_j]^2}, \tag{3}$$

where $\alpha = \frac{1}{d_1}$ and $\beta = \frac{1}{d_2}$ represent the contribution weights.

To explore the optimal contributing ratio, a positional encoding vector growing scheme is implemented to transcribe the low-dimensional data into a higher-dimensional space that complements the high-dimensional imaging features. Positional encoding is described as,

$$PE(y, 2k|d_2^*) = \sin\frac{y}{10000^{\frac{2k}{d_2^*}}}, \tag{4}$$

and

$$PE(y, 2k+1|d_2^*) = \cos\frac{y}{10000^{\frac{2k}{d_2^*}}}, \tag{5}$$

where, $y = \{y_j\}_{j=1}^{d_2}$ are the positions in the whole distribution, and $d_2^*$ is the dimension of the encoding space. The scaling coefficient $\gamma = \frac{\alpha}{\beta^*}$ is usually defined to explore the data fusion quality, where $\beta^*$ is the dimension of the positional encoded low dimensional features. Prior studies have demonstrated that the optimal $\gamma$ ratio varies depending on the specific application. For instance, Zhao et al. identified an optimal $\gamma$ ratio of 0.5 for fusing dose-incorporated MRI imaging features and clinical features for predicting brain metastasis stereotactic radiosurgery outcomes [24]. In contrast, Wang et al. reported an optimal $\gamma$ ratio of 1.0 when combining CT imaging features and blood markers to predict portal hypertension [23].

### 2.2. Monai

MONAI [25], which stands for Medical Open Network for AI, is an open-source, community-driven framework designed to facilitate deep learning in healthcare. It is particularly focused on medical imaging applications, providing specialized tools and functionalities that streamline the development of AI models to the specific needs of medical imaging, such as handling various imaging modalities (MRI, CT, X-ray, etc.) and supporting advanced image processing techniques. In this work, we used the LDCT nodule detection model from MONAI.

### 2.3. Neural Multi-task Linear Regression (NMTLR)

Over the past decades, a series of time-to-event models have evolved [10–13,26–28]. The primary goal of the time-to-event model is to estimate the probability of occurrence of the event of interest over time and to investigate the impact of various covariates on the event probability. Therefore, it can be modeled as either a regression problem, a classification problem, or a multi-task problem. In the regression approach, the focus is on predicting the time to the event of interest. The goal is to estimate the hazard function or survival function, which describes the probability of an event occurring at a specific time given the relevant covariates. In contrast, the classification approach aims to predict whether an event will occur within a given time window, without explicitly estimating the exact time to the event. In this case, the response variable is binary, indicating the occurrence or non-occurrence of the event. Combining both ideas, it can also be modeled as a multi-task problem, for example, NMTLR [12], predicting whether and when the event will happen simultaneously. As the focus of this work is not to explore different time-to-event models, we selected the NMTLR as the time-to-event model for this work.

The NMTLR model is a deep learning approach that extends the standard multi-task linear regression model to handle multiple related tasks simultaneously. It is particularly useful in problems where tasks are correlated or share common structures, such as predicting survival rates at different time intervals. The model consists of a shared input layer that learns a common representation across tasks, followed by task-specific output layers to capture task-specific parameters. This architecture enables NMTLR to leverage shared information across tasks while maintaining flexibility for individual predictions, making it well-suited for time-to-event analyses with high-dimensional and heterogeneous data. As highlighted in recent work on hepatocellular carcinoma (HCC) patients using the SEER database [29], NMTLR outperformed other deep learning and machine learning models, such as DeepSurv [10], Cox Proportional Hazards [30] and Random Survival Forest [31], in both calibration and discriminative ability, suggesting NMTLR a highly suitable choice for our study to provide the baseline time-to-event predictions.

## 3. Methodology

### 3.1. Concordance-based Predictive Uncertainty (CPU)-Index

As illustrated in Fig. 1, the CPU-Index is based on the assumption that when the true label of a test case is unknown, the AI model's prediction is less uncertain if it aligns more consistently with subgroup analysis conclusions.

Fig. 2 illustrates our complete methodological workflow for calculating CPU-Index while Algorithm 1 provides a step-by-step pseudo-code workflow. The process is briefly described as follows. Given a particular patient of interest (POI), we first calculate the subgroup analysis-based similarity between each patient in the training set and the POI as shown in Step 1 of Fig. 2. Based on this similarity measurement, patients in the training set are split into k batches as illustrated in Step 2 of Fig. 2. The batches are utilized to obtain a group-level

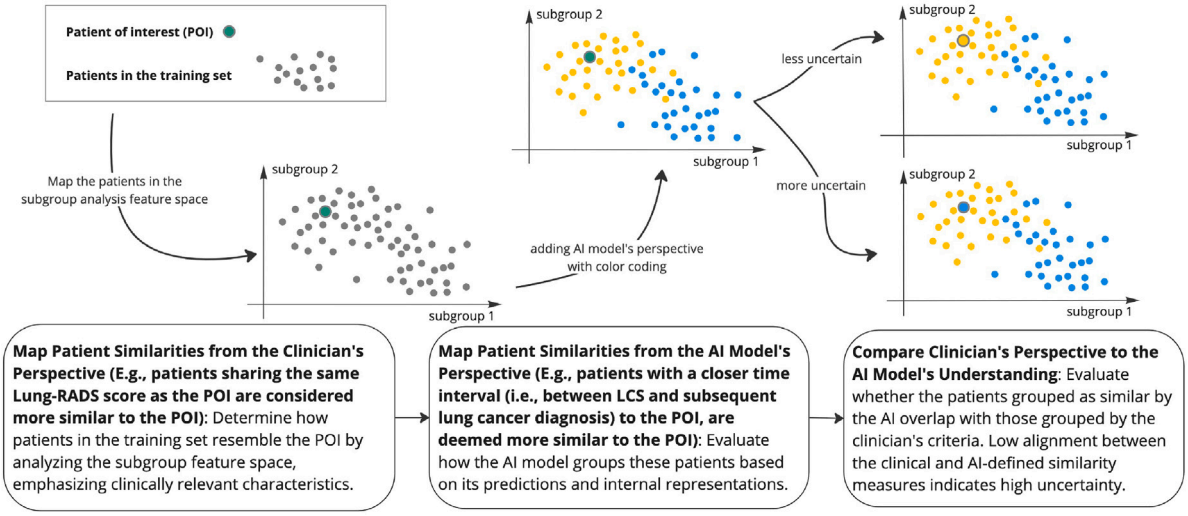


**Fig. 1.** A toy example showcasing the intuition of the CPU-Index. The process involves three key steps: (1) Mapping patient similarities from a clinical feature perspective based on subgroup analysis, where gray dots represent training cases, and the green dot indicates the patient of interest unseen by the model with an unknown true label. (2) Adding the AI model's perspective by color-coding patients based on model predictions (yellow and blue representing different classes). (3) Comparing these two perspectives to determine prediction uncertainty; the model's prediction is more uncertain if it is less consistent with the conclusions of the subgroup analysis.

**Algorithm 1** CPU-Index algorithm

**Ensure:** function $M$ : trained model
$P$ : training set
$POI$ : patient of interest
$k$ : the number of batches
$clinical\_lexicon\_list \leftarrow []$
function $L_{subgroup}(a, b)$ : the loss between a and b calculated from subgroup analysis
function $L_{pred}(a, b)$ : the prediction loss between a and b

**procedure** CALCULATING THE CPU-INDEX
$batched\_subgroup\_similarity\_rank \leftarrow$ generate_linearly_spaced_vector$(1, k)$
**for** $p$ in $P$ **do**
$L_{subgroup}(p, POI) = |clinical\_lexicon(p) - clinical\_lexicon(POI)| + ||POI \oplus p||_0$
Append $L_{subgroup}(p, POI)$ to $clinical\_lexicon\_list$
$P_{sorted} \leftarrow$ sort$(clinical_lexicon_list)$
$batched\_model\_similarity\_loss\_list \leftarrow []$
**for** $i$ in $batched\_subgroup\_similarity\_rank$ **do**
$batch_i \leftarrow$ cases indexed from $(i-1) * k + 1$ to $min(i * k + 1, size(P))$ in $P_{sorted}$
$batch_i\_pred\_list \leftarrow [M(p)$ for $p$ in $batch_i]$
$weights \leftarrow [L_{pred}(p, M(POI))$ for $p$ in $batch_i\_pred\_list]$
$weights \leftarrow softmax(-weights)$
$pred_b \leftarrow$ dot product$(weights, batch_i_pred_list)$
$L_{pred}(POI, batch_i) \leftarrow ||\mathrm{M}(POI) - pred_b||_2$
Append $L_{pred}(POI, batch_i)$ to $batched\_model\_similarity\_loss\_list$
$model\_similarity\_rank \leftarrow$ argsort$(batched_model_similarity_loss_list)$
CPU-Index $\leftarrow 1 -$ concordance index$(batched_subgroup_similarity_rank, model_similarity_rank)$
**return** CPU-Index

calibrated model prediction, which we use to calculate the prediction similarity between the POI and the calibrated in-batch subset of patients as shown in Step 3 of Fig. 2. Finally, the personalized CPU-Index associated with the POI is defined as the complement of the concordance index between subgroup analysis-based similarities and model-prediction-based similarities as illustrated in Step 4-5 of Fig. 2.

### *3.1.1. Subgroup analysis-based similarity loss and model prediction similarity loss*

Subgroup analysis-based similarity is defined based on a clinically relevant lexicon. In this work particularly, subgroup analysis-based similarity was defined based on the lung imaging reporting and data system (Lung-RADS), which is a classification system used by radiologists to report LDCT LCS results on a scale of 0–4. An important limitation of this nomogram is that closer scores may represent different feature combinations, and in some circumstances, the same score is given for different imaging findings. For example, patients with no lung nodules and patients with nodules with benign features will receive the same score. Therefore, we modified the subgroup analysis-based similarity by adding an entry-level zero–one loss term. Thus, the loss function of patient similarity for the patient of interest POI and any patient p in the training set P is defined in Eq. (6),

$$L_{patient}(p, POI) = L_{clinical\_lexicon}(p, POI) + L_{entry}(p, POI), \quad (6)$$

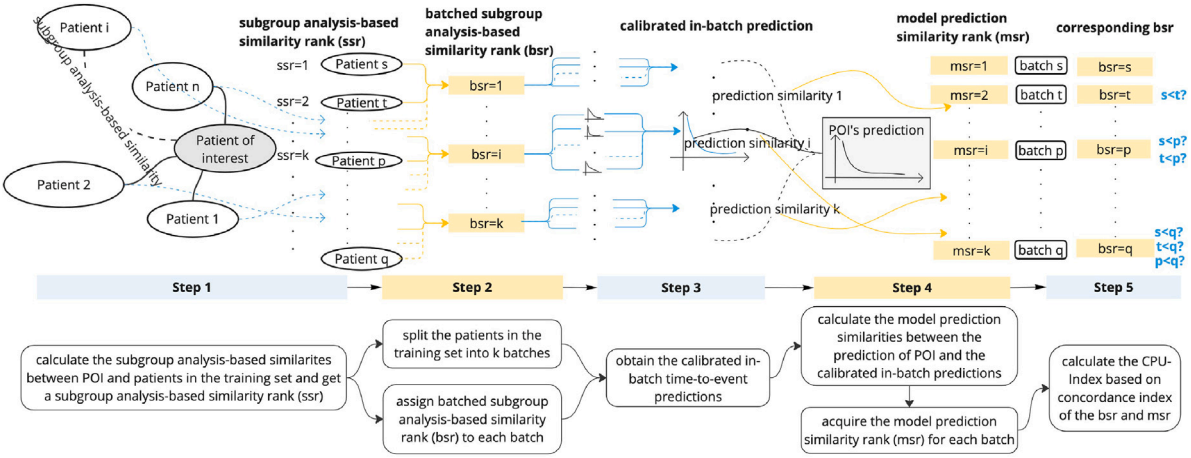


**Fig. 2. Workflow of the proposed personalized uncertainty quantification framework. Step 1:** Step 1: Given a patient of interest (POI), subgroup analysis-based similarity is calculated by comparing the POI to the remaining patients in the training set. We define the loss as the sum-of-the-difference between a clinically relevant lexicon and the feature-level 0-1 loss. Patients are then ranked based on their similarity score. **Step 2:** Patients in the training set are split into batches based on their subgroup analysis-based similarity loss (Step 1) to determine a batch-level patient similarity rank. **Step 3:** Batch-level calibrated model predictions are obtained by calculating the weighted average of all predictions within the batch. **Step 4:** Prediction similarity is measured by calculating the squared loss between each in-batch calibrated prediction and the prediction of the POI. The batches are then ranked based on model prediction loss. **Step 5:** A personalized CPU-Index for POI is defined as the complement of the concordance index of the subgroup analysis-based similarity rank (Step 2) and the model-prediction-based similarity rank (Step 4).

where

$$L_{clinical\_lexicon}(p, POI) = |clinical\_lexicon(POI) - clinical\_lexicon(p)|, \quad (7)$$

and

$$L_{entry}(p, POI) = \|POI \oplus p\|_0. \quad (8)$$

Next, we sort $p$ within $P$ in increasing order of $L_{patient}$ and define the subgroup analysis-based similarity rank $ssr$ for $p$ as the sorted index. Eqs. (6), (7), and (8) mathematically described Step 1 of Fig. 2. To reduce the effect of potential hazards that are not reflected within the data and the noise inside the data [22], instead of considering the similarity between individuals, as shown in Step 2 of Fig. 2, we split $P$ into $k$ ($k = 10$ in our experiment) batches $B$ and assign the batched subgroup analysis-based similarity rank $bsr$ as,

$$(b|bsr(b)) = (\{P_{ssr}, ssr \in [(i-1) * size_b + 1, i * size_b]\}|i), i \leq k, i \in \mathbb{N}, \quad (9)$$

where $size_b$ is the size of each batch,

$$size_b = \frac{size(P)}{k}.$$

As for the model prediction similarity, we define the model-prediction-level loss between $pred_b$ and $pred_{POI}$ as the prediction of each batch and $POI$ using squared error,

$$L_{pred}(pred_{POI}, pred_b) = \|pred_{POI} - pred_b\|_2, \quad (10)$$

where $pred_b$ is the prediction of each batch defined as the weighted average of all predictions within the batch,

$$pred_b = \sum_{p \in b} softmax(-L_{pred}(p, POI)) * pred_p. \quad (11)$$

With the in-batch calibrated prediction defined in Eq. (11) and illustrated in Step 3 of Fig. 2, the model-prediction similarity rank $msr$ shown in Step 4 of Fig. 2 is then acquired by assigning the index of the increasingly sorted $L_{pred}$.

#### 3.1.2. CPU-index

The personalized CPU-Index shown in Step 5 of Fig. 2 is defined as the complement of the concordance index between the batched patient similarity rank $bsr$ and the model prediction similarity rank $msr$,

$$\begin{aligned} \text{CPU-Index}_{POI} &= 1 - concordance\_index(gsr, msr) \\ &= Probability(msr_i < msr_j | bsr_i < bsr_j). \end{aligned} \quad (12)$$

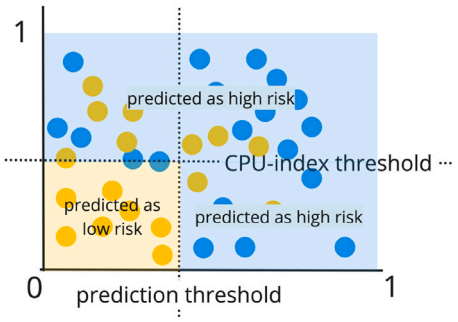


**Fig. 3.** Visualization of the final risk group prediction from combining the model's prediction with the CPU-Index.

### 3.2. CPU-index guided correction

For each case, the final outcome prediction is based on two values: the risk probability prediction from machine learning models and the CPU-Index representing this prediction's uncertainty. In the context of LDS prediction, where false negatives are more detrimental than false positives, we classify a case as negative only when it exhibits both a low predicted risk and a low CPU-Index, as illustrated in Fig. 3.

## 4. Materials and applications

In the following section, we demonstrated the proof of concept by illustrating how the CPU-Index improves the specificity of lung cancer detection in low-dose CT screening CT.

### 4.1. Data acquisition

We evaluated our approach on a dataset comprising 3,326 patients who underwent LCS from January 1, 2015, to June 30, 2020 [32]. Cases with incomplete electronic health records (EHR), unremarkable images with no discernible lesions, and diagnosed lung cancer earlier than the screening were excluded, and only one LDCT series per patient was selected. As a result, 1,767 patients were included and

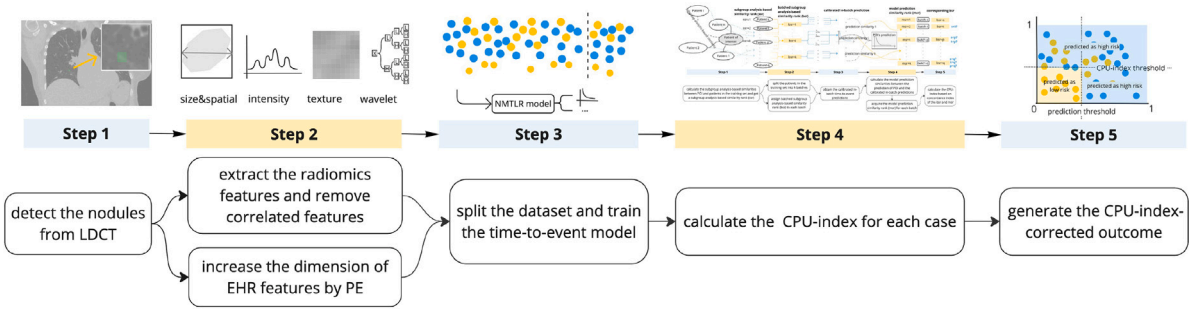


**Fig. 4.** The workflow of the improving specificity of lung cancers on low dose screening CT.

each patient's dataset consists of seven EHR features, including both demographics (i.e., gender, race, ethnicity, and age) and smoking status (i.e., tobacco used years, tobacco packs per day, and years since quitting), one LDCT series, a corresponding Lung-RADS score, subsequent lung cancer diagnosis outcomes, and the diagnosis or last follow-up time. Among these patients, 113 were diagnosed with lung cancer during the entire tracking period, with 90 diagnosed within one year of the LCS.

### 4.2. Experimental design

As shown in Fig. 4, in Step 1, we begin by detecting LDCT nodules using the MONAI lung nodule CT detection model [25]. Since the MONAI model does not guarantee complete accuracy in nodule identification, all detection results were reviewed and quality-controlled by a chest radiologist to ensure reliability. In Step 2, radiomics features-including nodule size-based, spatial statistics-based, intensity-based, texture-based, and wavelet transform-based-were extracted [33]. Features with a correlation higher than 0.95 were clustered, and only the one most correlated with the clinical outcome was retained, resulting in 172 selected radiomics features. Concurrently, low-dimensional EHR features were encoded into higher dimensions using positional encoding. We selected a series of dimensional ratios (0.5, 1.0, and 2.0) between the positionally encoded low-dimensional EHR features and high-dimensional radiomics features. The dataset was then split into training, validation, and test datasets at a ratio of 7:1:2 using Monte Carlo stratified sampling, as shown in Step 3. A time-to-event model was trained on the NMTLR model. In Step 4, after training, the CPU-Index for each case in the test set was calculated. With the model prediction at the one-year time point and the calculated CPU-Index for each case, the final CPU-Index corrected outcome was determined in Step 5.

To validate our approach, we subjected it to a 1000-permutation test (Steps 3 to 5) for robust evaluation. We compared the results with lung cancer predictions based on the Lung-RADS score alone, the NMTLR model alone, and our method at its optimal operating point. Lung-RADS scores of 3 or higher were categorized as predicting a lung cancer diagnosis within a year, while scores below 3 predicted no diagnosis. For the NMTLR model, we evaluated its performance using traditional metrics, such as the concordance index (C-index) and integrated Brier score (IBS). Additionally, at the one-year mark from the initial scan, we assessed model classification performance using metrics such as the area under the receiver operating characteristic curve (AUC), accuracy (ACC), false positive rate (FPR), and false negative rate (FNR). For Lung-RADS, serving as a baseline, we reported ACC, FPR, and FNR. To estimate the variance of each procedure, we performed permutation tests with 1000 iterations. To evaluate statistically significant differences in performance metrics across these permutations, we utilized Wilcoxon signed-rank tests, where $p < 0.05$ was considered statistically significant.

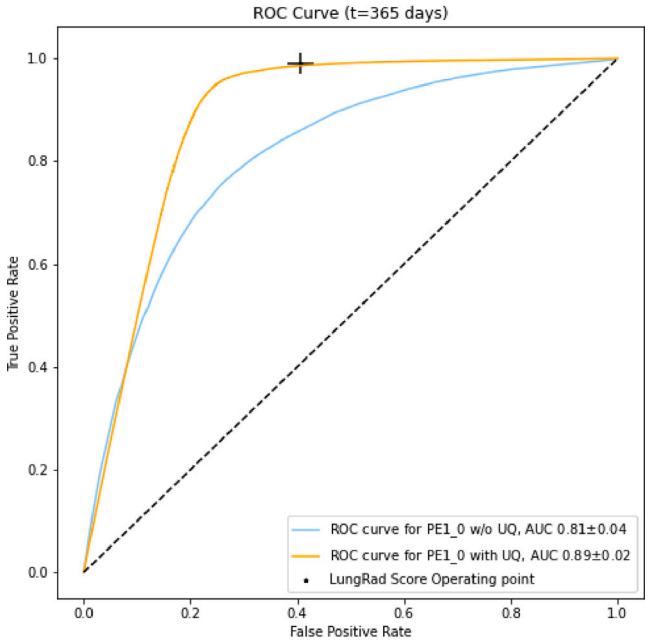


**Fig. 5.** The average ROC curves for the time-to-event model with and without UQ correction of 1000 permutation tests, comparing with Lung-RADS operating point. The blue line represents the ROC curve for the NMTLR model performance without CPU-Index correction, while the orange curve represents the overall performance adding CPU-Index correction.

### 4.3. Results

The average ROC curves from the 1000 permutation tests are shown in Fig. 5, with the Lung-RADS operating point included for comparison. Table 1 summarizes the key performance metrics.

Positional encoding appears to moderately enhance the model's performance, with a $\gamma$ ratio of 1.0 yielding slightly better results across multiple metrics. This suggests that a $\gamma$ ratio of 1.0 may be near optimal for balancing the predictive capabilities of the model, as shown in Table 1. As shown in Fig. 5, under this input setting, the UQ algorithm increased the model's AUC from 0.81±0.04 to 0.89±0.02. Compared with the Lung-RADS lexicon, the pipeline could improve the operating point by lowering the FPR from 0.41±0.02 to 0.30±0.13 while maintaining a nearly zero FNR.

## 5. Discussion

Our proposed CPU-Index assumes that predictions are more uncertain at high discordance of similarity metrics between subgroup

**Table 1**
Statistical results of the 1000 permutation tests.

| | Model | C-index | IBS | AUC | ACC | FPR | FNR |
|---|---|---|---|---|---|---|---|
| Lung-RADS | N/A | N/A | N/A | N/A | 0.615 ± 0.023 | 0.406 ± 0.024 | 0.01 ± 0.021 |
| **radiomics only** | NMTLR | 0.747 ± 0.058 | 0.065 ± 0.006 | 0.815 ± 0.045 | 0.744 ± 0.083 | 0.261 ± 0.092 | 0.173 ± 0.104 |
| | NMTLR with CPU-Index | N/A | N/A | 0.889 ± 0.02 | 0.707 ± 0.119 | 0.309 ± 0.126 | 0.0 ± 0.0 |
| **radiomics + EHR** | NMTLR | 0.746 ± 0.061 | 0.065 ± 0.007 | 0.814 ± 0.048 | 0.749 ± 0.085 | 0.255 ± 0.093 | 0.179 ± 0.106 |
| | NMTLR with CPU-Index | N/A | N/A | 0.888 ± 0.021 | 0.7 ± 0.126 | 0.316 ± 0.134 | 0.0 ± 0.0 |
| **radiomics + PE EHR (PE ratio = 0.5)** | NMTLR | 0.754 ± 0.057 | 0.065 ± 0.007 | 0.820 ± 0.045 | 0.766 ± 0.082 | 0.237 ± 0.090 | 0.188 ± 0.101 |
| | NMTLR with CPU-Index | N/A | N/A | 0.888 ± 0.021 | 0.704 ± 0.127 | 0.312 ± 0.134 | 0.0 ± 0.0 |
| **radiomics + PE EHR (PE ratio = 1.0)** | NMTLR | 0.755 ± 0.056 | 0.068 ± 0.007 | 0.813 ± 0.044 | 0.753 ± 0.081 | 0.250 ± 0.089 | 0.185 ± 0.100 |
| | NMTLR with CPU-Index | N/A | N/A | 0.888 ± 0.021 | 0.716 ± 0.119 | 0.3 ± 0.126 | 0.0 ± 0.0 |
| **radiomics + PE EHR (PE ratio = 2.0)** | NMTLR | 0.744 ± 0.056 | 0.069 ± 0.007 | 0.806 ± 0.047 | 0.741 ± 0.088 | 0.263 ± 0.097 | 0.184 ± 0.102 |
| | NMTLR with CPU-Index | N/A | N/A | 0.886 ± 0.021 | 0.712 ± 0.12 | 0.304 ± 0.127 | 0.0 ± 0.0 |

*Note:* Metrics including AUC, ACC, FPR, and FNR were compared using the Wilcoxon signed-rank test, with $p < 0.001$ for all comparisons.

analysis conclusions and personalized AI model predictions. Our approach enables us to provide each patient an uncertainty quantification score of the personalized AI model predictions, without knowing the ground truth event status. Our method conceptually works by combining a naive data-driven model with clinically relevant domain knowledge, which serve as orthogonal axes of model response. Uncertainty is estimated as high when these axes are adversarial. By applying PE to balance the dimension between high-dimensional radiomics and low-dimensional demographics, we demonstrated an improvement in prediction performance [23].

In comparison to previous work, our proposed CPU-Index addresses several limitations in existing diagnostic approaches based solely on either subgroup analysis [1–3] or personalized AI time-to-event models [7–9]. In the context of the LCS problem, subgroup analysis-based Lung-RADS encounters challenges such as accurately representing individual variations, an over-reliance on a restricted set of predefined characteristics, and insufficient temporal granularity, resulting in explainable yet highly variable predictions [6]. To address these challenges, subgroup analysis features for LCS are evolving, with alternative criteria to the U.S. Preventive Services Task Force (USPSTF) guidelines demonstrating improved sensitivity and specificity for identifying high-risk populations, including racial and ethnic minorities [34]. Further refinements in socio-demographic criteria also aim to improve adherence to follow-up screening and better target underrepresented groups [35], which may help combat social determinants of health and other disparities. In addition, improvements in liquid biopsy technology are on the rise [36], in particular, cell-free DNA and circulating tumor DNA, which have been linked to treatment response dynamics [37] and CT imaging [38] for lung cancer. As these technologies improve, they will enable better similarity metrics to improve the fidelity of our CPU-Index formalism.

Personalized AI time-to-event models suffer from uncertainties arising from their lack of transparency and biases stemming from censored data, hindering their applicability in real-world scenarios [16–18,39]. By combining insights from both AI modeling and domain expertise, the CPU-Index framework offers a nuanced evaluation of this bias and variance trade-offs. Additionally, from the data perspective, viewing subgroup analysis-based predictions as reflections of conclusions drawn from distinct datasets complements the training data used for AI model development.

In clinical practice, the CPU-Index provides actionable guidance for interpreting AI model predictions. Importantly, we note that the CPU-Index complements clinical judgment by a trained radiologist and can help them better interpret the diagnostic scenario. For example, higher CPU-Index values indicate greater concordance between clinical features (i.e., based on Lung-RADS and demographics) of patient similarity and the AI model's predictions, suggesting more reliable predictions. Conversely, lower CPU-Index values signal potential discrepancies between clinical features and AI predictions, warranting additional scrutiny by a human in the loop. Our findings suggest that an improvement in AUC from 0.81 to 0.89 indicates that the model is better at distinguishing between positive and negative cases. A reduction in FPR from 0.41 to 0.30 means about 110 fewer false positive results per 1000 screenings, assuming the prevalence and screening threshold remain constant. This is important for the efficacy of LCS programs, because reductions in false positives could significantly reduce unnecessary follow-up procedures, patient anxiety, and healthcare costs. We note that this initial FPR is slightly higher than previously reported [32], which is due to our exclusion of unremarkable images with no discernible lesions from which to define radiomic expression.

Regarding the UQ problem in time-to-event analysis, while previous methods predominantly focus on non-sampling or Monte Carlo sampling techniques to estimate uncertainty at a population level [16–18,39], our approach uniquely provides individualized UQ. This innovation is particularly relevant in clinical settings where personalized predictions are crucial but ground truth data may be unavailable or incomplete. By integrating subgroup analysis-based similarity metrics with personalized AI time-to-event models, our framework leverages both data-driven modeling and clinical domain knowledge, enhancing the interpretability and reliability of predictions. Moreover, the partitioning of patients into batches for concordance index calculation mitigates noise and improves the robustness of uncertainty estimates [22].

While our CPU-Index framework advances toward more transparent and clinically relevant UQ tailored for time-to-event analysis, it is not without limitations. First, the reliance on subgroup analysis-based conclusions may introduce biases or inaccuracies, particularly if the subgroup characteristics are not sufficiently diverse or representative of the broader population. Additionally, this reliance may limit the framework's applicability to true time-to-event predictions. In this study, although we employed a time-to-event model, we focused on event-at-time predictions, specifically assessing the probability of lung cancer occurrence within one year. This focus was driven by the clinical need to evaluate short-term risk for prioritizing follow-up screening and earlier intervention. Extending the framework to true time-to-event predictions would require two key conditions: (1) A model capable of predicting the survival function or cumulative hazard function over continuous or discrete time horizons; and (2) a clinically relevant subgroup analysis formalism from which to measure similarity. The time-to-event model we used (NMTLR) satisfies the first condition and can estimate time-to-event probabilities. However, the second condition is currently non-trivial for this application. Lung-RADS, which we used for uncertainty estimation, focuses on short-term event-at-time risk evaluation and does not directly estimate long-term risk. Therefore, while our CPU-Index framework may not directly apply to time-to-event predictions for the application studied in this paper, it is a generalized concept that could be extended to other scenarios where long-term subgroup similarity measures are more readily-available in clinical practice.

Second, the CPU-Index framework's computational complexity and resource requirements may limit its applicability. Specifically, the framework requires evaluating all cases in the training set for each test case, which can pose challenges in resource-constrained healthcare settings or with large, incomplete training datasets. Future work could

explore optimization strategies, such as reducing computational overhead through customizing a subset of representative samples instead of spanning all samples in the training set or leveraging parallel processing and cloud computing. Additionally, developing efficient preprocessing pipelines tailored to specific clinical datasets may enhance its scalability and practicality.

Third, while the framework aims to enhance transparency and interpretability, the complexity of its algorithms and calculations may hinder its usability for non-expert users. Simplifying the interpretive framework while maintaining transparency and predictive accuracy is critical to bridging this gap. For example, future work could focus on creating visual decision-support tools [40] or integrating interpretable AI techniques [41] to present CPU-Index outputs in a more intuitive and actionable format for clinical use.

Finally, the performance of CPU-Index is influenced by various hyperparameters, for example, the choice of the number of batches. In our implementation, we heuristically set the number of batches to create batches of approximately 100 patients each (given our training set of 1237 samples divided into 10 groups). While this choice yielded reasonable results, it was based on practical considerations rather than systematic optimization. The batch partitioning involves an inherent trade-off: too small a group may lead to unstable concordance calculations and high variance in uncertainty estimates, while too large a group may smooth over important local patterns and reduce sensitivity to individual patient characteristics. Future work could explore adaptive methods or develop theoretical frameworks for optimal batch size selection based on data characteristics and clinical requirements.

## 6. Conclusion

In summary, we developed a CPU-Index to estimate the uncertainty of individual predictions by AI models without knowing the true labels by incorporating subgroup analysis conclusions. Our method contributes to improved lung cancer risk assessment, and the incorporation of positional encoding further enhances overall performance.

## CRediT authorship contribution statement

**Yuqi Wang:** Writing – original draft, review & editing, Visualization, Software, Formal analysis, Methodology, Investigation, Conceptualization. **Aarzu Gupta:** Writing – review & editing, Software, Formal analysis, Conceptualization. **Fakrul Islam Tushar:** Writing – review & editing, Data curation. **Breylon Riley:** Writing – review & editing. **Avivah Wang:** Writing – review & editing, Data curation. **Tina D. Tailor:** Writing – review & editing, Supervision, Investigation, Funding acquisition . **Stacy Tantum:** Writing – review & editing, Supervision, Methodology. **Jian-Guo Liu:** Writing – review & editing, Supervision, Methodology. **Mustafa R. Bashir:** Writing – review & editing, Supervision. **Joseph Y. Lo:** Writing – review & editing, Supervision, Funding acquisition. **Kyle J. Lafata:** Writing – original draft, review & editing, Supervision, Project administration, Conceptualization, Funding acquisition.

## Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: Kyle J. Lafata, Joseph Y. Lo, Fakrul Islam Tushar reports financial support was provided by National Institute of Biomedical Imaging and Bioengineering. Kyle J. Lafata, Joseph Y. Lo reports financial support was provided by National Cancer Institute. 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.

## Acknowledgments

This work was funded in part by the Duke University Department of Radiology Charles E. Putman Vision Award, NIH/NIBIB P41-EB028744, NIH/NCI R01-CA261457, and NIH/NCI R01-CA289261.

## References

[1] Supplee LH, Kelly BC, MacKinnon DM, Barofsky MY. Introduction to the special issue: Subgroup analysis in prevention and intervention research. Prev Sci 2013;14:107—110.
[2] Rothwell PM. Subgroup analysis in randomised controlled trials: importance, indications, and interpretation. Lancet 2005;365(9454):176–86.
[3] Wang R, Lagakos SW, Ware JH, Hunter DJ, Drazen JM. Statistics in medicine — Reporting of subgroup analyses in clinical trials. N Engl J Med 2007;357(21):2189–94.
[4] Society AC. Cancer facts & figures 2021. 2021, URL https://www.cancer.org/content/dam/cancer-org/research/cancer-facts-and-statistics/annual-cancer-facts-and-figures/2021/cancer-facts-and-figures-2021.pdf. [Accessed 24 January 2023].
[5] National Lung Screening Trial Research Team, Aberle DR, Adams AM, Berg CD, Black WC, Clapp JD, et al. Reduced lung-cancer mortality with low-dose computed tomographic screening. N Engl J Med 2011;365(5):395–409.
[6] Sundaram V, Gould MK, Nair VS. A comparison of the PanCan model and lung-RADS to assess cancer probability among people with screening-detected, solid lung nodules. Chest 2021;159(3):1273–82.
[7] Beer DG, Kardia SLR, Huang C-C, Giordano TJ, Levin AM, Misek DE, et al. Gene-expression profiles predict survival of patients with lung adenocarcinoma. Nature Med 2002;8(8):816–24.
[8] Antonov AV, Krestyaninova M, Knight RA, Rodchenkov I, Melino G, Barlev NA. PPISURV: a novel bioinformatics tool for uncovering the hidden role of specific genes in cancer survival outcome. Oncogene 2014;33(13):1621–8, Number: 13 Publisher: Nature Publishing Group.
[9] Delen D, Walker G, Kadam A. Predicting breast cancer survivability: a comparison of three data mining methods. Artif Intell Med 2005;34(2):113–27.
[10] Katzman J, Shaham U, Bates J, Cloninger A, Jiang T, Kluger Y. DeepSurv: Personalized treatment recommender system using a Cox proportional hazards deep neural network. BMC Med Res Methodol 2018;18(1):24.
[11] Pölsterl S, Navab N, Katouzian A. Fast training of support vector machines for survival analysis. In: Appice A, Rodrigues PP, Santos Costa V, Gama Ja, Jorge Ap, Soares C, editors. Machine learning and knowledge discovery in databases. Lecture notes in computer science, Cham: Springer International Publishing; 2015, p. 243–59.
[12] Fotso S. Deep neural networks for survival analysis based on a multi-task framework. 2018, ArXiv.
[13] Lee C, Yoon J, Schaar Mvd. Dynamic-DeepHit: A deep learning approach for dynamic survival analysis with competing risks based on longitudinal data. IEEE Trans Biomed Eng 2020;67(1):122–33.
[14] Lundberg SM, Lee S-I. A unified approach to interpreting model predictions. In: Guyon I, Luxburg UV, Bengio S, Wallach H, Fergus R, Vishwanathan S, Garnett R, editors. In: Advances in neural information processing systems, vol. 30, Curran Associates, Inc.; 2017, p. 4765–74.
[15] Jiang L, Xu C, Bai Y, Liu A, Gong Y, Wang Y-P, et al. Autosurv: interpretable deep learning framework for cancer survival analysis incorporating clinical and multi-omics data. npj Precis Oncol 2024;8:4.
[16] García-Donato G, Cabras S, Castellanos ME. Model uncertainty quantification in Cox regression. Biometrics 2023;n/a(n/a). _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/biom.13823.
[17] Loya H, Anand D, Poduval P, Kumar N, Sethi A. A bayesian framework to quantify survival uncertainty. Ann Oncol 2019;30:vii32–3, Publisher: Elsevier.
[18] Sokota S, D'Orazio R, Javed K, Haider H, Greiner R. Simultaneous prediction intervals for patient-specific survival curves. In: Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence. 2019, p. 5975–81, Conference Name: Twenty-Eighth International Joint Conference on Artificial Intelligence {IJCAI-19} ISBN: 9780999241141 Place: Macao, China Publisher: International Joint Conferences on Artificial Intelligence Organization.
[19] Chapfuwa P, Tao C, Li C, Khan I, Chandross KJ, Pencina MJ, et al. Calibration and uncertainty in neural time-to-event modeling. IEEE Trans Neural Netw Learn Syst 2023;34(4):1666–80.
[20] Dubey M, Palakkadavadath R, Srijith P. Bayesian neural hawkes process for event uncertainty prediction. Int. J. Data Sci. Anal. 2023.
[21] Huh YK, Kim M, Liu K, Zhou S. An integrated uncertainty quantification model for longitudinal and time-to-event data. IEEE Trans Autom Sci Eng 2024;1–14.
[22] Stagge P. Averaging efficiently in the presence of noise. 1998, p. 188–200. http://dx.doi.org/10.1007/BFb0056862.
[23] Wang Y, Li X, Konanur M, Konkel B, Seyferth E, Brajer N, et al. Towards optimal deep fusion of imaging and clinical data via a model-based description of fusion quality. Med Phys 2022;50.

[24] Zhao J, Vaios E, Wang Y, Yang Z-Y, Cui Y, Reitman Z, et al. Dose-incorporated deep ensemble learning for improving brain metastasis SRS outcome prediction. Int J Radiat Oncol*Biol*Phys 2024.
[25] Cardoso MJ, Li W, Brown R, Ma N, Kerfoot E, Wang Y, et al. Monai: an open-source framework for deep learning in healthcare. 2022, http://dx.doi.org/10.48550/arXiv.2211.02701.
[26] Wright MN, Dankowski T, Ziegler A. Unbiased split variable selection for random survival forests using maximally selected rank statistics: M. WRIGHT, T. DANKOWSKI AND A. ZIEGLER. Stat Med 2017;36(8):1272–84.
[27] Li Y, Liang D, Ma S, Ma C. Spatio-temporally smoothed deep survival neural network. J Biomed Inform 2023;137:104255.
[28] Li Y, Yang AY, Marelli A, Li Y. MixEHR-SurG: A joint proportional hazard and guided topic model for inferring mortality-associated topics from electronic health records. J Biomed Inform 2024;153:104638.
[29] Wang S, Shao M, Fu Y, Zhao R, Xing Y, Zhang L, et al. Deep learning models for predicting the survival of patients with hepatocellular carcinoma based on a surveillance, epidemiology, and end results (SEER) database analysis. Sci Rep 2024;14:13232.
[30] Cox DR. Regression models and life tables. J R Stat Soc Ser B Stat Methodol 1972;34(2):187–202, _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.2517-6161.1972.tb00899.x.
[31] Wang H, Zhou L. Random survival forest with space extensions for censored data. Artif Intell Med 2017;79:52–61.
[32] Lafata KJ, Read C, Tong BC, Akinyemiju T, Wang C, Cerullo M, et al. Lung cancer screening in clinical practice: A 5-year review of frequency and predictors of lung cancer in the screened population. J Am College Radiol 2024;21(5):767–77, Focus on Interventional Radiology.
[33] Lafata KJ, Wang Y, Konkel B, Yin F-F, Bashir MR. Radiomics: a primer on high-throughput image phenotyping. Abdom Radiol 2022;47:2986–3002.
[34] Kearney LE, Belancourt P, Katki HA, Tanner NT, Wiener RS, Robbins HA, et al. The development and performance of alternative criteria for lung cancer screening. Ann Intern Med 2024;177(9):1222–32, PMID: 39159457.
[35] Silvestri GA, Goldman L, Burleson J, Gould M, Kazerooni EA, Mazzone PJ, et al. Characteristics of persons screened for lung cancer in the united states. Ann Intern Med 2022;175(11):1501–5, PMID: 36215712.
[36] Yu W, Hurley J, Roberts D, Chakrabortty SK, Enderle D, Noerholm M, et al. Exosome-based liquid biopsies in cancer: opportunities and challenges. Ann Oncol 2021;32(4):466–77.
[37] Corradetti MN, Torok JA, Hatch AJ, Xanthopoulos EP, Lafata K, Jacobs C, et al. Dynamic changes in circulating tumor DNA during chemoradiation for locally advanced lung cancer. Adv Radiat Oncol 2019;4(4):748–52.
[38] Lafata KJ, Corradetti MN, Gao J, Jacobs CD, Weng J, Chang Y, et al. Radiogenomic analysis of locally advanced lung cancer based on CT imaging and intratreatment changes in cell-free DNA. Radiol Imaging Cancer 2021;3(4):e200157.
[39] Zhang J. Modern Monte Carlo methods for efficient uncertainty quantification and propagation: A survey. WIREs Comput. Stat. 2021;13(5):e1539, _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/wics.1539.
[40] Yang Z, Hu Z, Ji H, Lafata K, Vaios E, Floyd S, Yin F-F, Wang C. A neural ordinary differential equation model for visualizing deep neural network behaviors in multi-parametric mri-based glioma segmentation. Medical Physics 2023;50(8):4825–38.
[41] Rudin C, Chen C, Chen Z, Huang H, Semenova L, Zhong C. Interpretable machine learning: fundamental principles and 10 grand challenges. ArXiv 2021;abs/2103.11251. https://api.semanticscholar.org/CorpusID:232307437.