A Deep CNN Approach For Predicting Cumulative Incidence Based On Pseudo-Observations

Pablo Gonzalez Ginestet (pablo.gonzalez.ginestet@ki.se)
Karolinska Institutet
Philippe Weitz
Karolinska Institutet
Mattias Rantalainen
Karolinska Institutet
Erin E Gabriel
Karolinska Institutet

Research Article

Keywords: Convolutional neural network, Machine Learning for survival data, Precision medicine, Predicting cumulative incidence, Pseudo-observation

Posted Date: July 14th, 2021

DOI: https://doi.org/10.21203/rs.3.rs-668944/v1

A Deep CNN approach for predicting cumulative incidence based on pseudo-observations

Pablo Matias Gonzalez Ginestet*. Affiliation: Department of Medical Epidemiology and Biostatistics, Karolinska Institutet, Stockholm, Sweden

Philippe Weitz. Affiliation: Department of Medical Epidemiology and Biostatistics, Karolinska Institutet, Stockholm, Sweden

Mattias Rantalainen. Affiliation: Department of Medical Epidemiology and Biostatistics, Karolinska Institutet, Stockholm, Sweden

Erin Gabriel. Affiliation: Department of Medical Epidemiology and Biostatistics, Karolinska Institutet, Stockholm, Sweden

*Corresponding author. Email: pablo.gonzalez.ginestet@ki.se

Keywords*:* Convolutional neural network; Machine Learning for survival data; Precision medicine; Predicting cumulative incidence; Pseudo-observation

Abstract

Background: Prognostic models are of high relevance in many medical application domains. However, many common machine learning methods have not been developed for direct applicability to right-censored outcome data. Recently there have been adaptations of these methods to make predictions based on only structured data (such as clinical data). Pseudo-observations has been suggested as a data pre-processing step to address right-censoring in deep neural network. There is a theoretical backing for the use of pseudo-observations to replace the right-censored response outcome, and this allows for algorithms and loss functions designed for continuous, non-censored data to be used. Medical images have been used to predict time-to-event outcomes applying deep convolutional neural network (CNN) methods using a Cox partial likelihood loss function under the assumption of proportional hazard. We propose a method to predict the cumulative incidence from images and structured clinical data by integrating (or combining) pseudo-observations and convolutional neural networks.

Results: The performance of the proposed method is assessed in simulation studies and a real data example in breast cancer from The Cancer Genome Atlas (TCGA). The results are compared to the existing convolutional neural network with Cox loss. Our simulation results show that our proposed method performs similar to or even outperforms the comparator, particularly in settings where both the dependent censoring and the survival time do not follow proportional hazards in large sample sizes. The results found in the application in the TCGA data are consistent with the results found in the simulation for small sample settings, where both methods perform similarly.

Conclusions: The proposed method facilitates the application of deep CNN methods to time-to-event data and allows for the use of simple and easy to modify loss functions thus contributing to modern image-based precision medicine.

Background

It has recently been demonstrated that contemporary medical image analysis has the potential to improve the diagnostic and prognostic stratification of cancer patients [1-3]. Particularly the analysis of microscopic morphological patterns in histopathological tissue sections is a key component of routine care. For instance, in lung cancer, tumors with predominantly micro-papillary and solid patterns have been associated with a poorer prognosis [4]. With the advent of digital pathology, whole-slide-images (WSIs) of stained tissue sections are becoming increasingly available. This may provide the opportunity to accurately predict individual prognoses using image data paired with other clinical information at scale and provide clinicians with decision support that can guide clinical management decisions to enhance personalized treatment and thus improve patient care [1-2].

Deep convolutional neural network(s) (CNN) are currently at the forefront of image analysis and have become the state-of-the-art in image-based precision medicine [5-6]. Deep CNN models are neural networks with several layers, including convolutional layers that are suitable for modelling of image data. Deep CNNs learn hierarchical representations directly from raw image data given a large dataset of labeled examples.

Few machine learning methods have been developed for survival outcomes originally, and thus, most existing machine learning for survival outcomes are adaptations. This is also true for image analysis methods. CNN methods have been used and adapted to address the task of predicting time-to-event outcomes from WSIs. Recent works [7-11] have used WSIs with CNN for survival predictions. They applied convolutional layers to extract features of the images using convolutional kernels and pooling operations, followed by a sequence of fully connected layers where the terminal layer outputs a predicted risk associated with the image. These risks are plugged in the Cox partial likelihood and the network is trained using a back-propagation procedure and optimization algorithm. These prior works combined modern CNN models with Cox regression for prediction of time-to-event outcomes, keeping the assumption of proportional hazard. This stipulates no effect modification by time, which can be restrictive or even unrealistic [12]. Furthermore, the negative partial log-likelihood is a relatively complicated loss function that can be challenging to implement in existing CNN frameworks.

Recent works in other areas of machine learning have suggested data pre-processing steps that can be used to adapt common classes of machine learning methods to time-to-event outcomes. Pseudo-observations [13] is one of such

methods that has been suggested to adapt random forests [14-15] and more generally all methods for continuous outcomes and ensembles of them [16]. [17] proposed the use of a modified version of pseudo-observations of [13], which they call conditional, to replace the observed survival times to make risk predictions in deep neural network. By using pseudo-observations, [17] avoided the sophisticated loss functions for censored data or the proportional hazard assumption from previous work that modeled survival data using deep neural network [18-22].

We propose to combine classical pseudo-observations with CNN models in order to make risk predictions based on medical images and clinical covariates in a setting of right-censoring. Our proposed method can be applied to any CNN model, and thus, applies to all those publicly available in Pytorch [23] or TensorFlow [24], combining images and structured clinical data used in a deep neural network. After appropriate validation in clinical studies, this risk score could prove valuable for prognostic patient stratification. We demonstrate our method in simulations based on the CIFAR-10 images [25] and in our motivating data example in breast cancer from The Cancer Genome Atlas Breast Invasive Carcinoma data [26].

Methods

Setup and notation

Let $T^{(m)}$ denote the true event-time for an individual m and $C^{(m)}$ the censoring time,

$\tilde{T}^{(m)} = \min(T^{(m)}, C^{(m)})$ the observed survival time and event indicator $\Delta^{(m)} = 1(T^{(m)} \leq C^{(m)})$. In addition, for each individual we observe a p-dimensional vector of clinical covariates at baseline $X^{(m)}$ and a three-dimensional image data denoted as $I^{(m)}$. Each image data is a 3D array of size w ×h×d, where w and h are spatial dimensions and d is the channel dimension, where color images have three channels (red, green and blue (RGB)).

Without censoring, the sample data would be $\mathcal{D}^{ideal} = \{(I^{(m)}, X^{(m)}, T^{(m)}, y^{(m)}(\tau))\}$ for m = 1...N where $y^{(m)}(\tau)$ is the response variable for individual m indicating if the individual has experienced the event at a specific time τ , $y^{(m)}(\tau) = 1(T^{(m)} \leq \tau)$. The goal is to predict the individual risk of experiencing the main event before time τ given his or her information based on sample data $\mathcal{D}^{ideal}$. However, in the presence of censoring, the response variable $y^{(m)}(\tau)$ is not observed for all m. Instead, we observe the sample data $\mathcal{D} = \{(I^{(m)}, X^{(m)}, T^{(m)}, \tilde{y}^{(m)}(\tau))\}$ for m = 1...N where $\tilde{y}^{(m)}(\tau) = \Delta^{(m)}\, 1(\tilde{T}^{(m)} \leq \tau)$.

Pseudo-Observations

[13] introduced a strategy to transform a censored problem into an uncensored one in order to be able to apply standard methods for complete data such as regression models. If $y\ \ (\tau)$ were not subject to censoring, we could use it directly to model the cumulative incidence. In the presence of censoring, the pseudo-observation approach replaces the censored response variable $y^{(m)}(\tau)$ of each individual m by a jackknife pseudo-observation, which can be used as a new response variable to fit models. Pseudo-observations can be based on a number of estimators. We will focus on the nonparametric cumulative incidence estimator of failure before time τ . In the absence of competing risks, a nonparametic estimator of the cumulative incidence of the event of interest is given by $\theta(\tau) = 1 - S_{KM}(\tau)$ θ(τ), where $S_{KM}\ (\tau)$ is the KaplanMeier (KM) survival function. The pseudo-observation (PO) cumulative incidence for individual m at time τ is computed as

$$\hat{\theta}^{(m)}(\tau) = N \times \hat{\theta}\ \ (\tau) - (N-1) \times \hat{\theta}^{(-m)}(\tau)$$

where $\hat{\theta}\ \ (\tau) = 1 - \hat{S}_{KM}(\tau)$ and $\hat{S}_{KM}(\tau)$ is the the KM estimator of the survival function based on all the examples and $\hat{\theta}^{(-m)}(\tau) = 1 - \hat{S}_{KM}^{(-m)}(\tau)$ is obtained by eliminating individual m from the data. The PO are used as a replacement for the incompletely observed random variable $y^{(m)}(\tau)$ for each individual. The asymptotic justification of the pseudo-observation approach requires that $\hat{\theta}\ \ (\tau)$ to be a consistent estimator of $\theta(\tau)$, and that the right-censoring be independent of the survival time and any covariates one intends to include in the model [27].

In cases where the censoring is potentially dependent on covariates, one can model the censoring and use inverse probability of censoring weighted (IPCW) methods to consistently estimate the survival function [28, 29]. In order to perform IPCW, one estimates the conditional censoring survival function at time τ, denoted by $G^{(m)}(\tau) = P(C^{(m)} > \tau | X^{(m)})$ and weights each individual by the inverse of their estimated probability. The IPCW estimator for the survival probability is $\hat{S}^{W}(\tau) = \exp\{\Lambda_W(\tau)\}$, where $\Lambda_W(\tau)$ is the IPCW version of the Nelson-Aalen estimator for the cumulative hazard function [30]. Just like the non-parametric pseudo-observations, there is a large number of different ways to fit the IPCW pseudo-observations [31, 32].

The weighted pseudo-observation, IPCW-PO, cumulative incidence for individual m at time τ uses $\hat{S}^{W}(\tau)$ in Equation 1. Appropriate procedures to estimate G(·) are the Cox proportional hazards model [33] and more flexible models

such as Aalen's linear hazard model [34], boosted Cox regression [35] or random forest [36]. Once pseudo-observations are obtained, the sample data for the analysis is given by $\mathcal{D}^{PO} = \{(I^{(1)}, X^{(1)}, \hat{\theta}^{(1)}(\tau)), \dots, (I^{(N)}, X^{(N)}, \hat{\theta}^{(N)}(\tau))\}$. $\mathcal{D}^{PO}$ can be used to train any CNN model to predict the individual risk of experiencing the main event before time τ, which would have been similar to basing our predictions on $\mathcal{D}^{ideal}$ if this sample data were available.

Convolutional Neural Network

Convolutional neural networks are a class of neural networks that can be applied to data that spatially encodes information in an evenly-spaced grid topology, such as images or time series. As compared to multi-layer perceptrons (MLPs), CNNs share weights between their kernels or filters to drastically reduce the number of model parameters, which is based on the assumption of translational invariance of these filters. Through a hierarchical structure of consecutive convolutional layers, CNNs can learn representations of increasing complexity, such that the first layers typically extract unspecific low-level features such as corners and edges, whereas the last layer encode more abstract concepts that are more specific to the training data. This structure of convolutional layers is typically followed by one or more fully connected layers, which are comparable to an MLP that weights the activation of the final convolutional layers, resulting in a model output. We refer the readers to [5, 37] for a more detailed exposition.

Pseudo-Observation (PO)

CNN Our proposed PO-CNN procedure enables a CNN to be fitted using a PO-based response to predict the cumulative incidence from images and structured clinical data. The pipeline of the proposed framework is shown in Figure 1 and Figure 2 and can be summarized as follows.

i) For a finite number of time points, compute the PO (or IPCW-PO) cumulative incidence for each individual using D to construct DPO.

ii) Choose a CNN model and add additional fully connected layers at the terminal layer of the CNN and the clinical data as intermediate input for multiple outputs (Figure 2.d) OR include the time points as input too for a single output implementation (Figure 2.c).

iii) Train the PO-CNN (or IPCW-PO-CNN) using a mean squared error (MSE) based loss function of your choice.

The implementation that only includes the clinical data as intermediate input is a multi-output regression, which is related to multi-task supervised learning approach [38]. In Figure 2.d, each output at a different time point is regarded as a specific output and each of them have several task-specific layers while sharing all previous layers. Thus, there is no need to add the time points as intermediate predictors. We denote this implementation multi-output. Unlike single output implementation, by default the multi-output minimizes the combined MSE of each output values together. Although we simply sum the different losses, this can be tailored as desired. The single output loss function can also be tailored as desired to more highly weight a particular time point, or can only use a single time point. Although we do not investigate this further, this simple modification of the loss function may be of great advantage over the existing Cox-loss methods. In what follows, we use the default average MSE over all included time points in both PO-CNN approaches.

It is of note that although it is technically possible to use the image information in the fitting of the censoring model, we do not believe this is practical or necessary. Instead, fitting a model for the censoring distribution based solely on the set of available clinical covariates is likely sufficient and much more feasible in practice. Thus, we assume that it is sufficient to condition on the clinical covariates when modeling the censoring in the IPCW-PO. This is slightly more restrictive than a Cox based CNN as all inputs in the CNN are included in some way in the final layer and thus are accounting for censoring; we investigate this in the simulations.

Results

For all methods that follow we used a Residual Network model [39] with 18 layers (ResNet18), although any desired model could be used. The typical block of layer in ResNet is i) convolution; ii) non-linearity activation function; iii) batch normalization; iv) pooling and v) dropout. We used a pre-trained ResNet model on ImageNet [40, 41] to initialize the weights instead of random initialization. We added to the last terminal layer of ResNet a simple fully connected layer followed by $\tanh(\cdot)$ as the activation function. We used Pytorch for the CNN implementation with Adam [42] as optimizer. To obtain the PO and IPCW-PO, we used the R packages prodlim [43] and eventglm [32],

respectively, where the IPC weights were estimated using a Cox regression model based on the set of clinical covariates. We compared our proposed procedures to an existing CNN modelling approach for survival prediction that uses the Cox partial likelihood as loss function to handle censored data. We denoted as Cox-CNN. An example of Cox-CNN is [7]. Cox-CNN is trained using the sample data D. For the Cox-CNN, we incorporated clinical data in the last terminal layer of the CNN model, just as in our proposed methods. Code containing details of all procedures are available at https://github.com/pablogonzalezginestet/POCNN.

Simulations

We used images from the CIFAR-10 dataset as the basis of our simulated data. The CIFAR-10 dataset consists of color images (32 × 32 × 3) in 10 classes. We denote the classes with y where $y_i \in \{0, \ldots, 9\}$ is the class for the image I_i. We generated the true survival time based on the classes that each image represents as well as independent covariates. We generated nine independent covariates $X_1, \ldots, X_9$ from the standard normal distribution and one binary covariate X_{10}. We presented six cases corresponding to different survival and censoring time models. For each case, we consider a sample size of N = 1000 and 5000. From each sample size, 80% of the observations were randomly sampled for training, while the remaining 20% were set aside as a test set. The simulations were repeated 100 times each. The accuracy of the prediction of the cumulative incidence at the four percentiles observed times was assessed using the area under the ROC curve (AUC). The prediction at each time point were compared to the true binary outcome of having an event prior to a given time point of interest. This latter variable is no censored since we know the exact survival time. The pseudo-observations PO and IPCW-PO were computed for a grid of time points corresponding to 20th, 30th, 40th and 50th percentiles of the overall time distribution. We do not tune any hyper-parameter. All simulations were run using a learning rate of 0.0001. For each simulation, we applied both approaches: single output and multi-output to both IPCW-PO and unweighted PO.

The six cases are as follow:

Case 1. The true survival time was generated from a proportional hazard model. T was generated with hazard function:

$$\lambda_T(t \mid y, X) = \lambda_{T,0}(t) \exp\{1.7y + (0.3 + 0.6\cos(y))X_{10} + 0.2X_1\}$$

where $\lambda_{T,0}(\mathrm{t}) = 2\mathrm{t}$. We randomly selected around 30% observations to be rightcensored at time C generated from a uniform distribution on (0, T).

Case 2. The true survival time was generated under a proportional hazard model as Case 1 but the censoring time is generated from

$$\lambda_C\ (\mathrm{t} \mid \mathrm{X}) = \lambda_{C,0}(\mathrm{t}) \exp\{\ 1.4\mathrm{X}_{10} + 2.6\mathrm{X}_1 - 0.2\mathrm{X}_2\ \}$$

where $\lambda_{C,0}(\mathrm{t}) = 12\mathrm{t}$. The censoring percentage is around 20%.

Case 3. The true survival time was generated from a proportional hazard model where

$$\lambda_T\ (\mathrm{t} \mid \mathrm{y}, \mathrm{X}) = \lambda_{T,0}(\mathrm{t}) \exp\{\ \mathrm{y} - 1.6\cos(\mathrm{y})\mathrm{X}_{10} + 0.3\mathrm{X}_1\ \mathrm{X}_{10}\}$$

and $\lambda_{T,0}(\mathrm{t}) = 0.7\mathrm{t}$. The censoring time was generated using a gamma distribution with shape parameter equal to $\exp\{-1.8\mathrm{X}_{10} + 1.4\mathrm{X}_1 + 1.5\mathrm{X}_{10}\mathrm{X}_1\}$ and scale parameter equal to y. The censoring percentage is around 43%.

Case 4. The true model for survival time was generated using a gamma distribution with shape parameter equal to $\exp\{0.5\mathrm{y} + 0.2\mathrm{X}_{10}\cos(\mathrm{y}) + 1.5\mathrm{X}_1 + 1.2\mathrm{X}_{10}\}$. We randomly selected 30% observations to be right-censored at time C generated from a uniform distribution on (0, T).

Case 5. The true survival times are non-proportional hazard as Case 4 and the censoring time was generated

$$\lambda_C\ (\mathrm{t} \mid \mathrm{X}) = \lambda_{C,0}(\mathrm{t}) \exp\{-3.4\mathrm{X}_{10} + 0.6\mathrm{X}_1 - 2.2\mathrm{X}_2\ \}$$

Where $\lambda_{C,0}(\mathrm{t}) = 0.01\mathrm{t}$. The censoring percentage is around 60%.

Case 6. The true survival times and the censoring times are both generated using a gamma distribution. The shape parameter is $\exp\{0.7\mathrm{y} + 0.4\mathrm{X}_{10}\mathrm{y} - 0.1\mathrm{X}_1\ \mathrm{X}_{10} + 0.1\mathrm{y}\mathrm{X}_1\}$ and $\exp\{3.8\mathrm{X}_{10} + 5.2\mathrm{X}_1 - 3.3\mathrm{X}_{10}\mathrm{X}_1\}$ for the survival and censoring time, respectively. The shape parameter was set equal to y. The censoring percentage is around 65%.

Figure 3 and Figure 4 show the simulations results. When the sample size is small, Figure 3, IPCW-PO-CNN single output had the best performance across the different cases even in cases that they did not require IPC-weights, with a median AUC above of all other models. However, this method and the multi-output version tended to show high variability when the censoring time was generated from a non-proportional hazard model (*Case 3 and Case 6*). The latter behavior was expected since the censoring model is misspecified. The second best median performance was the unweighted PO-CNN single output. The Cox-CNN demonstrated similar results to PO-CNN, even in the non-proportional hazard cases, something that was surprising and that we suspect is due to the sample size.

Figure 4 depicts a clearer and more stable pattern, which we believe is due to the sample size increase, from N = 1000 to N = 5000. Firstly, the prediction accuracy increased across all methods and time points. In the case where the censoring model is misspecified, *Case 6*, the impact on the performance of IPCW-PO-CNN was accentuated. However, except for this case, the PO-CNN and IPCW-PO-CNN performed similarly. The unweighted PO-CNN had the best median performance across the different cases and the single output performed better than multi-output, weighted or not.

Over all sample sizes and time points we see minor improvements or similar results in predictive accuracy using unweighted single output PO-CNN as compared to the Cox-CNN. Except for when both the event time and the censoring time have large deviations from proportional hazards, *Case 6*, where the Cox-CNN does not perform well, the multi-output PO-CNN performs similarly to the Cox-CNN. Although IPC weighting seems to improve results slightly when the censoring is dependent and the censoring model is correct in larger sample sizes and overall in smaller sample sizes, the potential losses due to incorrect modelling, as demonstrated in *Case 6* and sample size 5000, likely makes chasing these minor improvements inadvisable. Therefore, based on the simulations, one would expect that there is little to lose using either the unweighted single output or multi-output PO-CNN over the Cox-CNN, while there may be slight improvements in predictive accuracy.

Real Data Application

We illustrate the proposed method using whole-slide histopathology images of breast tumors and clinical structured data obtained from The Cancer Genome Atlas Breast Invasive Carcinoma (TCGA-BRCA) [26]. The event of interest was time to death from first diagnosis of breast cancer at four time points: 2 year, 3.5 year, 5 year and 8 year. We

implement our proposed method by computing pseudo-observations for these four time points. We selected the following clinical predictors from the clinical data: race, ethnicity, age, pathologic stage and molecular subtype [44, 45]. Table 3 in the appendix summarizes these variables. The breast histopathology image dataset are composed of 710 WSIs and each of them was tiled into image patches that span 512 x 512 pixels at 20X magnification. Tiling WSIs into smaller image patches and assigning the patient-level label to each image patch is a common strategy in digital pathology due to current memory constraints. Models are then fitted to these image patches in a weakly supervised manner. In order to only predict risk scores from cancer tissue regions, we deployed a cancer detection model to exclude benign tissue. Each WSI, which is associated to a patient, is linked to the clinical data. Patients were divided randomly into training (64%), validation (16%) and test (20%) data sets, respectively. The number of tiles in the train/val/test was 4,494,472/1,061,601/1,417,900, respectively. Due to differences in tumor size and variations in the sectioning, patients have differing numbers of tiles. To sample equivalent numbers of tiles per patient, we decided to augment the original number of tiles of all patients to the extent of balancing the number of tiles per patient.

We performed data augmentation as a form of regularization, including random horizontal flip and random rotation from -90º to 90º. For all models, we only tuned the learning rate using the package Ray Tune [46, 47] for a maximum of 30 epoch in each trial. The mean absolute error was used as evaluation metric in the validation set for PO-CNN and IPCW-PO-CNN, whereas the average of the AUC for each time point was used for the Cox-CNN. We trained the CNN model on per-tile basis. The final per-slide prediction, which is our interest, was obtained by applying a tile aggregation method. We considered the average and the 75th percentile of the per-tile scores across all tiles as a patient-level prediction. As for the evaluation for the test dataset, we used the time-dependent area under the ROC curve (AUC) for right-censored time-event data [48, 49]. The AUC is estimated at the four different time points of interest.

Due to the superior performance of single output over multi-output shown in the simulations, we did not refer to the latter in this analysis. Instead, we included PO-CNN that uses as response variable a PO cumulative incidence computed only for one time point. We denoted PO-CNN one-time-point. This model is trained separately for each time point considered in the analysis and for this reason we only present the non-weighted version of it. This model

would act as a lower bound in terms of accuracy for the PO-CNN that is computed for a grid of time points. Theoretically, multiple time points should perform as well or better than single time points, due to the fact that information across the time points is shared.

Comparative results are presented in Table 1 and Table 2. The results obtained using the averaging criteria are similar to those using the 75th percentile aggregation criteria. The unweighted PO-CNN and Cox-CNN had similar performance for the prediction at later years, while Cox-CNN resulted in better accuracy for the earliest year and PO-CNN for the middle two time points. PO-CNN resulted in better accuracy than its weighted version across all years except for year five. Lastly, POCNN one-time-point performed as expected, except for the earliest time.

Discussion

Improved prognostic models, including those based on routine histopathology image data, are of high clinical relevance as they can provide information that is important for clinical decision making. The proposed method, based on pseudo-observation, provides an efficient approach to fit deep CNN models to right-censored time-to-event outcomes using standard loss functions, making implementation straight forward while providing comparable or improved model performance in comparison to alternative approaches.

We showed over a large set of simulated scenarios that our proposed method of PO-CNN performed similar to or even outperformed the existing CNN for survival analysis that uses the Cox partial likelihood, while having a simpler and more easily modified loss function. We found this was particularly true in settings where both the dependent censoring and the survival time did not follow the assumption of proportional hazards in large sample size. Although in the real data example, the proposed PO-CNN that performed best in the simulations was outperformed by the Cox-CNN for one time point, it performed similarly or slightly better at all other time points. These results are also consistent with the results found in the simulation for a small sample size, where the Cox-CNN and the PO-CNN performed more similarly over all scenarios. The superior performance of the Cox-CNN for this time point paired with the best performance being obtained by the single time point PO-CNN, suggests that the model for this time point may differ from the other time points.

Despite the fact that training a CNN model requires a large amount of images, in the area of medical research it often happens that the real application dataset is based on a small number of whole slide images as training samples. This fact may be a limitation of our approach. However, as shown in the simulations, this is not the case when using a large dataset. Additionally, the individual tiles used to train the model in the real application may not be discriminative and thus biasing the predictions [50]. This may be accommodated by modifying a loss function to more highly weight the earlier time point.

Lastly, we have not investigated tuning hyper-parameters in great detail other than the learning rate. The CNN model as well as the two implementations (single output and multi-output) can be tuned and we think that with further hyper-parameter tuning, better performance might be achieved. In the multi-output implementation, the criteria used to combine the loss of the multiple outputs is an important hyper-parameter to tune. Our suspicion is that the poor performance seen there in the simulation is caused by a lack of tuning. Furthermore, the real application poses extra challenges. For instance, one should tune the aggregation procedure applied to get a per-slide prediction per individual. Also, one might consider weighting the loss function at each time point to take account for the heterogeneity across time points. When these later factors are tuned, a higher performance can potentially be achieved. Although this is a future line of work for the authors, this in no way detracts from the fact that the PO-CNN is clearly a useful alternative to the Cox-CNN model that allows for simple and easier to modify loss function.

Conclusions

In this work, we proposed a method that uses classical pseudo-observations as the outcome in deep CNN methods to predict the cumulative incidence using images and structured clinical data. Compared with Cox regression based model, the proposed method is more flexible as it does not assume proportional hazards. The proposed method facilitates the application of deep CNN methods to time-to-event data with a simple and easily modified loss functions. This works contributes to modern image-based precision medicine be providing an alternative to Cox loss in CNN image analysis for prediction of cumulative incidence or risk before a given time point.

Declarations

Ethics approval and consent to participate

Not applicable

Consent for publication

Not applicable

Availability of data and materials

The images and clinical data used in the real application are publicly available through The Cancer Genome Atlas (TCGA) Program (https://portal.gdc.cancer.gov/projects/TCGA-BRCA). The simulation dataset generated and all the program codes used in this project are available in the POCNN repository, https://github.com/pablogonzalezginestet/POCNN.

Competing interests

The authors declare that they have no competing interests.

Funding

This work has been supported by The Swedish research council grants 2017-01898, 2018-03056 and ERAPERMED2019-224-ABCAP, The Swedish Cancer Society grants 20 0714 Pj and 20 0906 PjF, and The Swedish e-science Research Centre (SeRC) - eCPC.

Authors' contributions

PGG and EEG developed the methods. PGG wrote the code for the methods and the simulations, ran and interpreted the simulations and analyzed and interpreted the real data example. PGG was the primary contributor to the writing, with major contributions from EEG and useful contributions from MR and PW. MR and PW aided in the understanding and analysis of the real data example and PW in the use of the CNN architecture. All authors have read and approved the manuscript.

Author details

Department of Medical Epidemiology and Biostatistics, Karolinska Institutet, Stockholm, Sweden.

Acknowledgements

Not applicable

Abbreviations

AUC: Area under the ROC curve

CNN: Convolutional Neural Network

Cox-CNN: CNN that uses the Cox partial likelihood as loss function

IPCW: Inverse probability of censoring weighted

IPCW-PO: Inverse probability of censoring weighted pseudo-observation

KM: Kaplan-Meier

MLP: Multi-layer perceptron

MSE: Mean squared error

PO: Pseudo-observation

PO-CNN: Proposed method that integrates PO and CNN

IPCW-PO-CNN: Proposed method that integrates IPCW-PO and CNN

RGB: Red, green and blue color model

ROC: Receiver operating curve

TCGA: The Cancer Genome Atlas

WSIs: Whole-slide-images

References

1. Colling R, Pitman H, Oien K, Rajpoot N, Macklin P, in Histopathology Working Group CPA, et al. Artificial intelligence in digital pathology: a roadmap to routine use in clinical practice. The Journal of Pathology. 2019;249(2):143–150.

2. Ehteshami Bejnordi B, Veta M, Johannes van Diest P, van Ginneken B, Karssemeijer N, Litjens G, et al. Diagnostic assessment of deep learning algorithms for detection of lymph node metastases in women with breast cancer. JAMA. 2017;318(22):2199–2210.

3. Yoo H, Kim KH, Singh R, Digumarthy SR, Kalra MK. Validation of a deep learning algorithm for the detection of malignant pulmonary nodules in chest radiographs. JAMA Network Open. 2020;3(9):e2017135–e2017135.

4. Ma M, She Y, Ren Y, Dai C, Zhang L, kang Xie H, et al. Micropapillary or solid pattern predicts recurrence free survival benefit from adjuvant chemotherapy in patients with stage IB lung adenocarcinoma. Journal of thoracic disease. 2018;10(9):5384–5393.

5. LeCun Y, Bengio Y, Hinton G. Deep Learning. Nature. 2015;521:436–44.

6. Lu L, Zheng Y, Carneiro G, Yang L. Deep learning and convolutional neural networks for medical image computing: precision medicine, high performance and large-scale datasets. Advances in Computer Vision and Pattern Recognition. Springer; 2017.

7. Mobadersany P, Yousefi S, Amgad M, Gutman DA, Barnholtz-Sloan JS, Velázquez Vega JE, et al. Predicting cancer outcomes from histology and genomics using convolutional networks. Proceedings of the National Academy of Sciences. 2018;115(13):E2970–E2979. Available from: https://www.pnas.org/content/115/13/E2970.

8. Wulczyn E, Steiner DF, Xu Z, Sadhwani A, Wang H, Flament-Auvigne I, et al. Deep learning-based survival prediction for multiple cancer types using histopathology images. PLOS ONE. 2020;15(6):1–18. Available from: https://doi.org/10.1371/journal.pone.0233678.

9. Li H, Boimel P, Janopaul-Naylor J, Zhong H, Xiao Y, Ben-Josef E, et al. Deep convolutional neural networks for imaging data based survival analysis of rectal cancer. In: 2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019); 2019. p. 846–849.

10. Hao J, Kosaraju SC, Tsaku N, Song D, Kang M. PAGE-Net: Interpretable and integrative deep learning for survival analysis using histopathological images and genomic data. Pacific Symposium on Biocomputing. 2020;25:355–366.

11. Zhu X, Yao J, Huang J. Deep convolutional neural network for survival analysis with pathological images. In: 2016 IEEE International Conference on Bioinformatics and Biomedicine (BIBM); 2016. p. 544–547.

12. Hernán M. The Hazards of Hazard Ratios. Epidemiology. 2010;21(1):13–5.

13. Andersen PK, Klein JP, Rosthøj S. Generalised linear models for correlated pseudo-observations, with applications to multi-state models. Biometrika. 2003;90(1):15–27.

14. Ishwaran H, Kogalur UB, Blackstone EH, Lauer MS. Random survival forests. Annals of Applied Statistics. 2008;2(3):841–860.

15. Hothorn T, Bühlmann P, Dudoit S, Molinaro A, Van Der Laan MJ. Survival ensembles. Biostatistics. 2006;7(3):355–373.

16. Sachs MC, Discacciati A, Everhov AH, Olen O, Gabriel EE. Ensemble prediction of time-to-event outcomes with competing risks: a case-study of surgical complications in Crohn's disease. Journal of the Royal Statistical Society: Series C (Applied Statistics). 2019;68(5):1431–1446.

17. Zhao L, Feng D. Deep neural networks for survival analysis using pseudo values. IEEE Journal of Biomedical and Health Informatics. 2020.

18. Faraggi D, Simon R. A neural network model for survival data. Statistics in Medicine. 1995;14(1):73–82. Available from: https://onlinelibrary.wiley.com/doi/abs/10.1002/sim.4780140108.

19. 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 Medical Research Methodology. 2018;18(24).

20. Ching T, Zhu X, Garmire LX. Cox-nnet: An artificial neural network method for prognosis prediction of high-throughput omics data. PLOS Computational Biology. 2018; 14(4):1–18. Available from: https://doi.org/10.1371/journal.pcbi.1006076.

21. Luck M, Sylvain T, Cardinal H, Lodi A, Bengio Y. Deep learning for patient-specific kidney graft survival analysis. ArXiv. 2017;abs/1705.10245.

22. Hao J, Kim Y, Mallavarapu T, Oh JH, Kang M. Interpretable deep neural network for cancer survival analysis by integrating genomic and clinical data. BMC Medical Genomics. 2019;12:189.

23. Paszke A, Gross S, Massa F, Lerer A, Bradbury J, Chanan G, et al. PyTorch: An imperative style, high-performance deep learning library. In: NeurIPS; 2019.

24. Abadi M, Agarwal A, Barham P, Brevdo E, Chen Z, Citro C, et al.. TensorFlow: Large-scale machine learning on heterogeneous systems; 2015. Software available from tensorflow.org.

25. Krizhevsky A. Learning multiple layers of features from tiny images; 2009.

26. Gutman D, Cobb J, Somanna D, Yuna P, Wang F, Kurc T, et al. Cancer Digital Slide Archive: an informatics resource to support integrated in silico analysis of TCGA pathology data. Journal of the American Medical Informatics Association : JAMIA. 2013;20.

27. Graw F, Gerds T, Schumacher M. On pseudo-values for regression analysis in competing risks models. Lifetime Data Anal. 2009;15:241–255.

28. Robins J, Finkelstein D. Correcting for noncompliance and dependent censoring in an AIDS clinical trial with inverse Probability of Censoring Weighted (IPCW) Log-Rank Tests. Biometrics. 2000;56(3):779–788.

29. Satten GA, Datta S, Robins J. Estimating the marginal survival function in the presence of time dependent covariates. Statistics and Probability Letters. 2001;54(4):397 – 403.

30. Xiang F, Murray S. Restricted mean models for transplant benefit and urgency. Statistics in Medicine. 2012;31(6):561–576.

31. Binder N, A GT, Andersen PK. Pseudo-observations for competing risks with covariate dependent censoring. Lifetime Data Analysis. 2014;(20):303–315.

32. Sachs MC, Gabriel EE, Overgaard M, Gerds TA, Therneau T. eventglm: Regression models for event history outcomes.; 2020. R package version 2020.11.10. Available from: https://sachsmc.github.io/eventglm.

428 33. Cox DR. Regression models and life-tables. Journal of the Royal Statistical Society Series B (Methodological).
429 1972;34(2):187–220.

430 34. Aalen OO. A linear regression model for the analysis of life times. Statistics in Medicine. 1989;8(8):907–925.

431 35. Binder H. CoxBoost: Cox models by likelihood based boosting for a single survival endpoint or competing risks;
432 2013. R package version 1.4. Available from: https://cran.r-project.org/package=CoxBoost.

433 36. Ishwaran H, Kogalur UB. randomForestSRC: Fast Unified Random Forests for Survival, Regression, and
434 Classification (RF-SRC). 2020. R package version 2.9.3. Available from: https://cran.r-
435 project.org/package=randomForestSRC.

436 37. Goodfellow I, Bengio Y, Courville A. Deep Learning. MIT Press; 2016.

437 38. Zhang Y, Yang Q. An overview of multi-task learning. National Science Review. 2017 09;5(1):30–43. Available
438 from: https://doi.org/10.1093/nsr/nwx105.

439 39. He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. IEEE Conference on Computer Vision
440 and Pattern Recognition (CVPR). 2015.

441 40. Russakovsky O, Deng J, Su H, Krause J, Satheesh S, Ma S, et al. ImageNet large scale visual recognition challenge.
442 International Journal of Computer Vision. 2015;115:211-252.

443 41. Raghu M, Zhang C, Kleinberg J, Bengio S. Transfusion: understanding transfer learning for medical imaging.
444 NeurIPS. 2019; 3342-3352.

445 42. Kingma D, Ba J. Adam: A method for stochastic optimization. International Conference on Learning
446 Representations. CoRR. 2015; abs/1412.6980.

447 43. Gerds TA. prodlim: Product-limit estimation for censored event history analysis.; 2020. R package version
448 2019.11.13. Available from: https://cran.r-project.org/package=prodlim.

449 44. Russo J, Frederick J, Ownby HE, Fine G, Hussain M, Krickstein HI, et al. Predictors of recurrence and survival of
450 patients with breast cancer. American Journal of Clinical Pathology. 1987;88(2):123–131.

45. Lee J, Kim S, Kang B. Prognostic factors of disease recurrence in breast cancer using quantitative and qualitative magnetic resonance imaging (MRI) parameters. Scientific Reports. 2020.

46. Liaw R, Liang E, Nishihara R, Moritz P, Gonzalez JE, Stoica I. Tune: A research platform for distributed model selection and training. ArXiv. 2018;abs/1807.05118.

47. Moritz P, Nishihara R, Wang S, Tumanov A, Liaw R, Liang E, et al. Ray: A distributed framework for emerging AI applications. In: Proceedings of the 13th USENIX Conference on Operating Systems Design and Implementation. OSDI'18. USA: USENIX Association; 2018. p. 561–577.

Uno H, Cai T, Tian L, Wei L. Evaluating prediction rules for t-year survivors with censored regression models. Journal of the American Statistical Association. 2007;102:527 - 537.

49. Kamarudin AN, Cox T, Kolamunnage-Donà R. Time-dependent ROC curve analysis in medical research: Current methods and applications. BMC Medical Research Methodology. 2017;17(1):53.

50. Hou L, Samaras D, Kurç T, Gao Y, Davis JE, Saltz J. Patch-based convolutional neural network for whole slide tissue image classification. 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2016;p. 2424–2433.

51. Otsu N. A threshold selection method from gray-level histograms. IEEE Transactions on Systems, Man, and Cybernetics. 1979;9(1):62–66.

52. Macenko M, Niethammer M, Marron JS, Borland D, Woosley JT, Guan X, et al. A method for normalizing histology slides for quantitative analysis. In: 2009 IEEE International Symposium on Biomedical Imaging: From Nano to Macro; 2009. p. 1107–1110.

Figures Legends

Figure 1. Pseudo-observations cumulative incidence are computed at a finite number of time points for each individual to be used as the new response variable.

Figure 2. (b) Medical images are passed-through the CNN model chosen; (c) single output, and (d) multi-output.

Figure 3. Boxplots of AUC values for the prediction of the cumulative incidence at 20th, 30th, 40th and 50th percentile of the overall time across 100 simulated datasets of sample size 1000 using different methods: cox, CNN

with Cox PH layer; po1, PO-CNN single output; po2, PO-CNN multi-output; po3, IPCW-PO-CNN single output, and po4, IPCW-PO-CNN multi-output.

Figure 4. Boxplots of AUC values for the prediction of the cumulative incidence at 20th, 30th, 40th and 50th percentile of the overall time across 100 simulated datasets of sample size 5000 using different methods: cox, CNN with Cox PH layer; po1, PO-CNN single output; po2, PO-CNN multi-output; po3, IPCW-PO-CNN single output, and po4, IPCW-PO-CNN multi-output.

497 Tables

498 Table 1. Estimated AUCs for predicting death at 2-year, 3.5-year, 5-year and 8-year, using the average.

CNN model	AUC(t=2)	AUC(t=3.5)	AUC(t=5)	AUC(t=8)
Proposed PO-CNN (one-time-point)	0.802	0.630	0.688	0.674
Proposed PO-CNN (single output)	0.696	0.760	0.740	0.832
Proposed IPCW-PO-CNN (single output)	0.685	0.693	0.750	0.678
Cox-CNN	0.785	0.718	0.737	0.832

499

500

501 Table 2. Estimated AUCs for predicting death at 2-year, 3.5-year, 5-year and 8-year, using the average.

CNN model	AUC(t=2)	AUC(t=3.5)	AUC(t=5)	AUC(t=8)
Proposed PO-CNN (one-time-point)	0.806	0.623	0.725	0.691
Proposed PO-CNN (single output)	0.690	0.727	0.703	0.804
Proposed IPCW-PO-CNN (single output)	0.696	0.714	0.753	0.729
Cox-CNN	0.774	0.707	0.733	0.830

502

503

504
505

506

Appendix

Description of the clinical information of the TCGA-BRCA dataset

The Cancer Genome Atlas (TCGA) Program provides publicly-available clinical data for different types of cancers. For this analysis, we use the Breast Cancer (BRCA) clinical dataset. From the entire dataset, we selected the following clinical predictors:

- age: continuous variable
- race: categorical variable that takes on three categories ('black', 'white' and 'other').
- ethnicity: categorical variable that takes on two categories ('not hispanic latino' and 'other')
- pathologic stage: categorical variable that takes on four categories (Stage I', 'Stage II', 'Stage III', 'StageX'). This is the classification of cancer stages based on tumor, lymph and metastasis.
- molecular subtype: categorical variable that takes on five categories ('Basal', 'HER2', 'Luminal A', 'Luminal B', 'Normal')

The next table summarizes the predictor variables as well as the time-to-event variable, time to death where time is measured by days, and the vital status of the patient, if it is alive/censored (status=0) or dead (status=1).

Table 3: Summary of the clinical information of breast cancer patients included in the analysis

	Overall
Sample size	710
Days to death (mean (SD))	1351.32 (1267.36)
Status (mean (SD))	0.16 (0.37)
Age (mean (SD))	58.33 (12.97)
Race	
Black	87 (12.3)
white	567 (79.9)
Other	& 56 (7.9)
Ethnicity = other (%)	87 (12.3)
Pathologic stage (%)	
Stage I	& 62 (8.7)
Stage II	397 (55.9)
Stage III	151 (21.3)
StageX	100 (14.1)

Molecular subtype (%)	
Basal	129 (18.2)
Her2	57 (8.0)
LumA	369 (52.0)
LumB	132 (18.6)
Normal	23 (3.2)

Image pre-processing

Each WSI was preprocessed before inclusion into this study. As a first step, tissue masks were generated. To this end, WSIs were down-sampled by a factor of 32 and converted to the HSV color space. Tissue masks were generated by applying a pixel-wise logical operation between a mask that was generated by applying the Otsu threshold [51] to the saturation channel and by applying a cutoff of 0.75 to the hue channel. We subsequently performed morphological opening and closing to remove salt-and-pepper noise from the binary masks. WSIs were then tiled with 50% overlap at 20X magnification into image patches spanning 598x598 pixels, where 598 pixels correspond to 271µm of tissue section, while discarding tiles with less than 50% tissue content. Tiles were then color-normalized using the method described by [52]. We subsequently applied a cancer detection CNN to identify cancer regions and excluded all tiles that were predicted to belong to a benign tissue region. Furthermore, out-of-focus tiles with a low variance were excluded, which was computed by filtering each tile with a Laplacian.

Figures

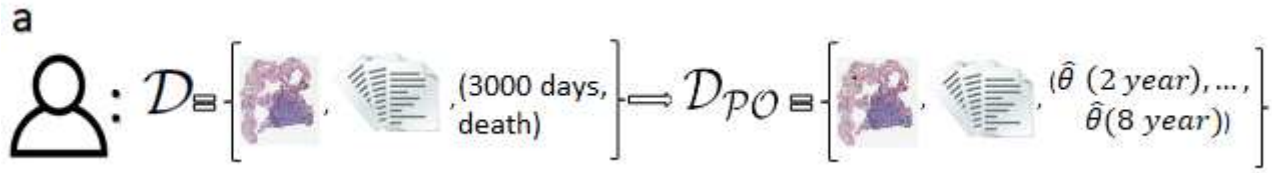

Figure 1

Pseudo-observations cumulative incidence are computed at a finite number of time points for each individual to be used as the new response variable.

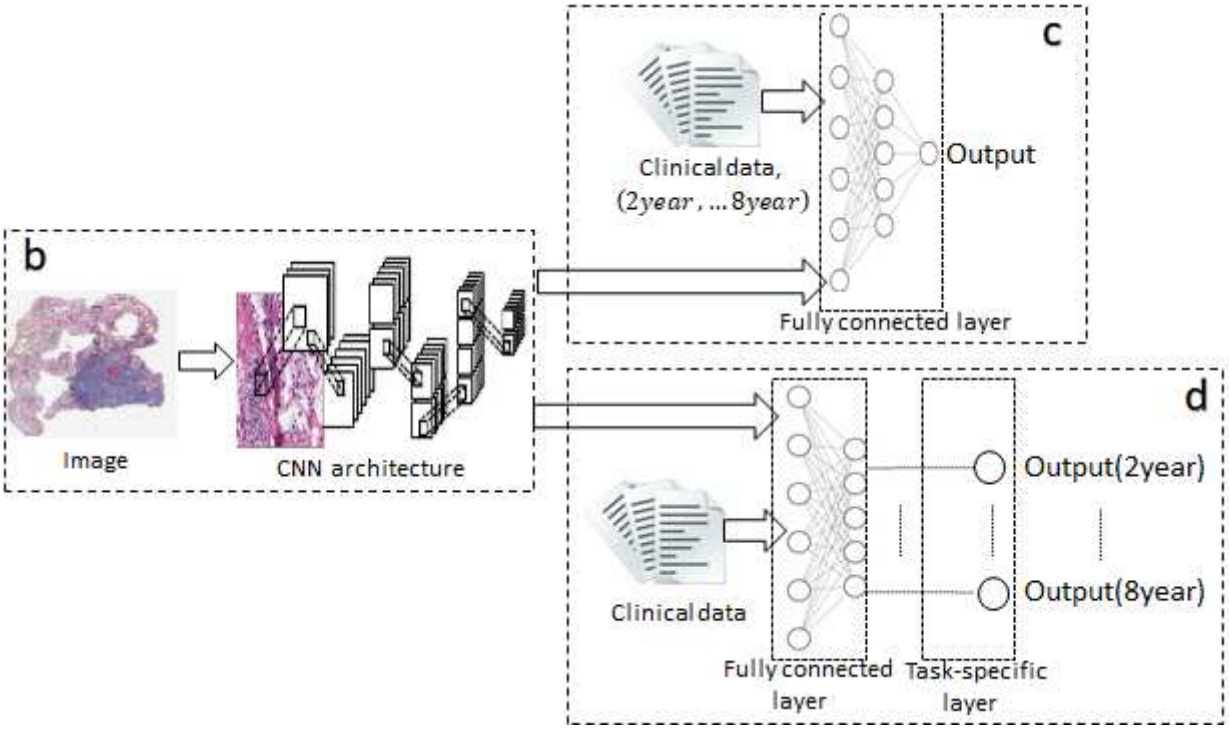

Figure 2

(b) Medical images are passed-through the CNN model chosen; (c) single output, and (d) multi-output.

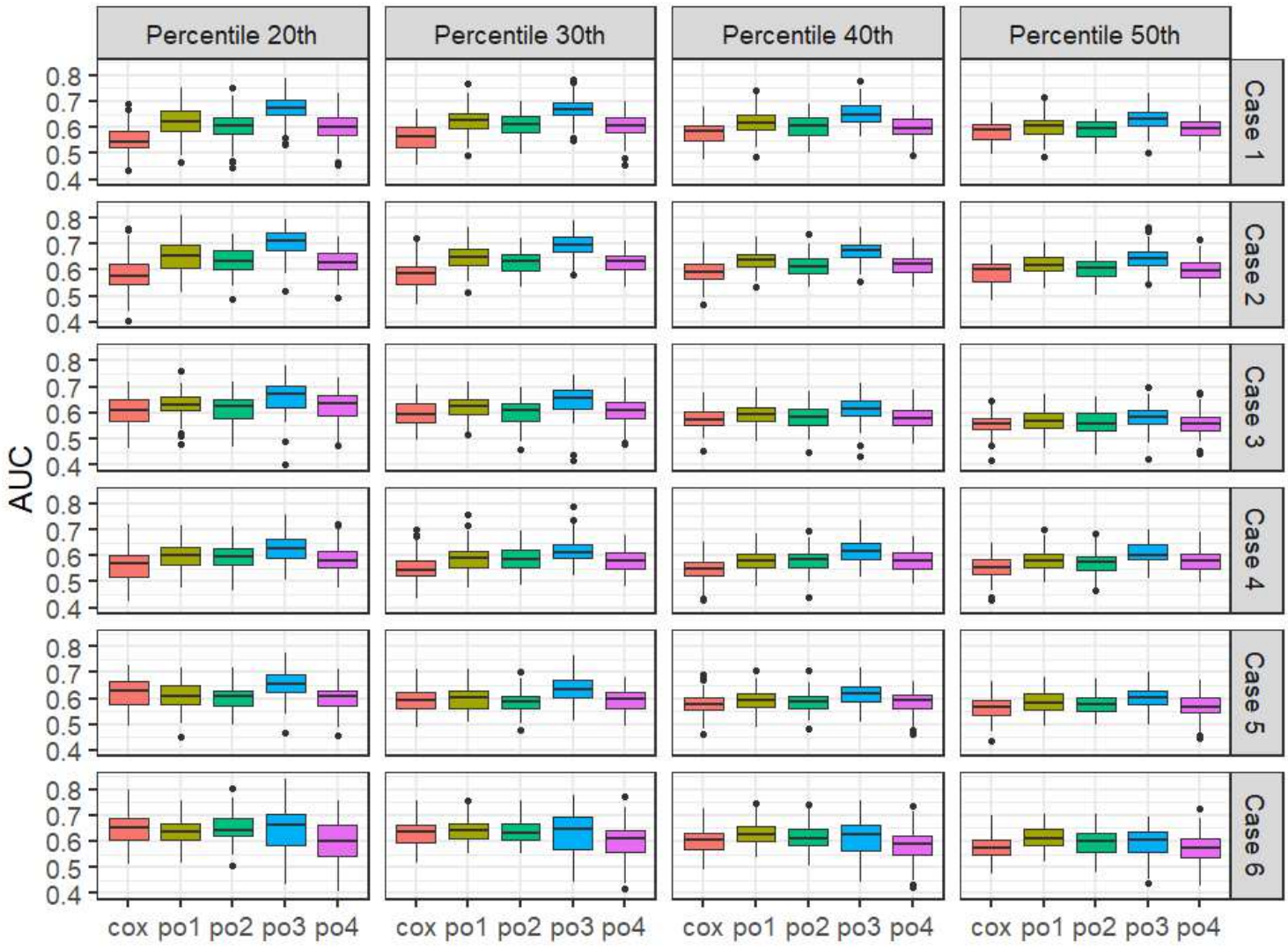

Figure 3

Boxplots of AUC values for the prediction of the cumulative incidence at 20th, 30th, 40th and 50th percentile of the overall time across 100 simulated datasets of sample size 1000 using different methods: cox, CNN with Cox PH layer; po1, PO-CNN single output; po2, PO-CNN multi-output; po3, IPCW-PO-CNN single output, and po4, IPCW-PO-CNN multi-output.

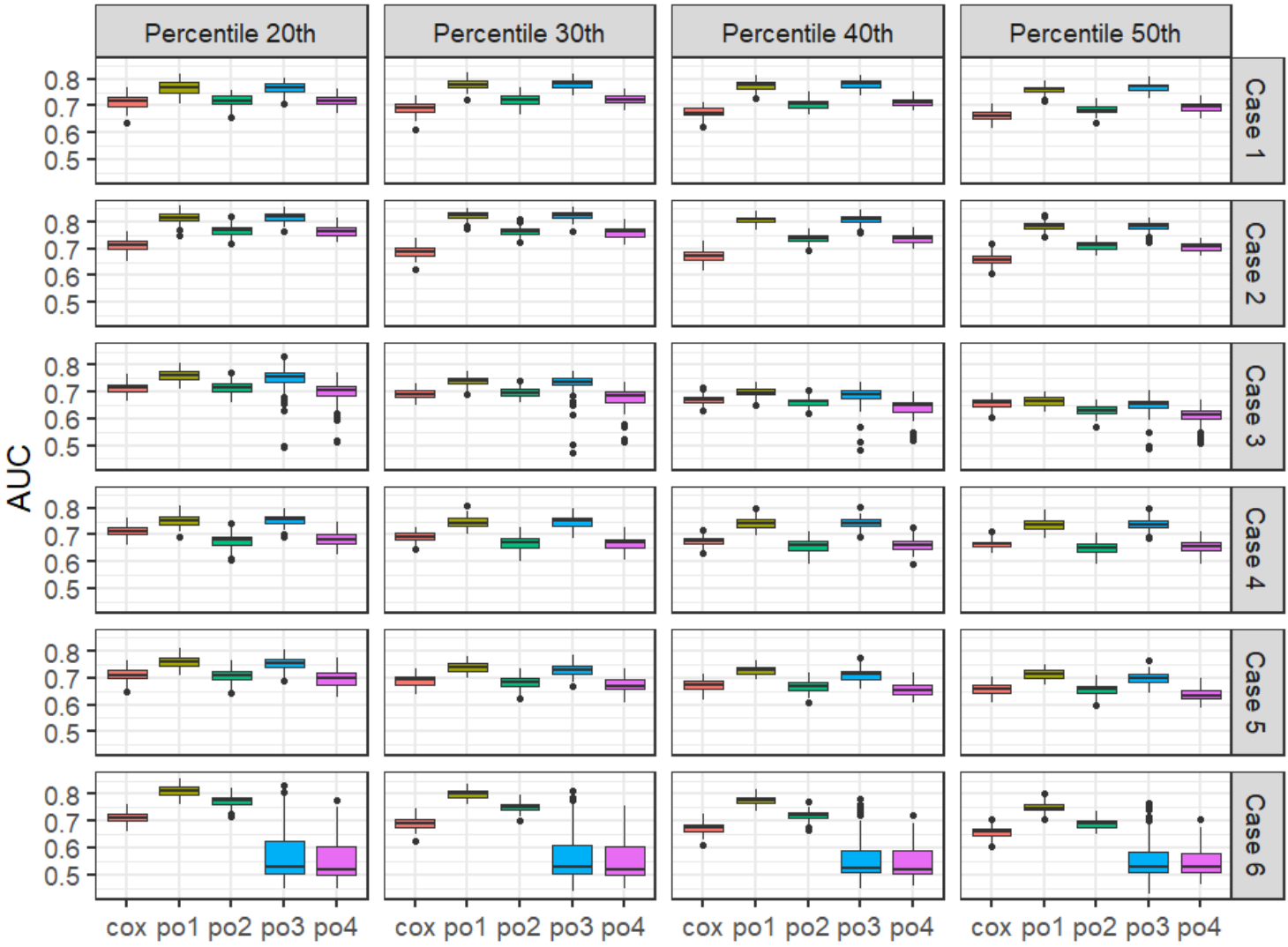

Figure 4

Boxplots of AUC values for the prediction of the cumulative incidence at 20th, 30th, 40th and 50th percentile of the overall time across 100 simulated datasets of sample size 5000 using different methods: cox, CNN with Cox PH layer; po1, PO-CNN single output; po2, PO-CNN multi-output; po3, IPCW-PO-CNN single output, and po4, IPCW-PO-CNN multi-output.