An evaluation of machine learning techniques to predict the outcome of children treated for Hodgkin-Lymphoma on the AHOD0031 trial

A report from the Children's Oncology Group

Cédric Beaulac[a] Jeffrey S. Rosenthal[b] Qinglin Pei[c]
Debra Friedman[d] Suzanne Wolden[e] David Hodgson[f]

March 26, 2021

[a] Department of Statistical Sciences, University of Toronto, Toronto, Canada; [b]Department of Statistical Sciences, University of Toronto, Toronto, Canada; [c]Department of Biostatistics, University of Florida, Gainesville, USA; [d]Department of Pediatrics, Vanderbilt University, Nashville, USA; [e]Department of Radiation Oncology, Memorial Sloan Kettering Cancer Center, New York, USA; [f] Department of Radiation Oncology, University of Toronto, Toronto, Canada.

Abstract

In this manuscript we analyze a data set containing information on children with Hodgkin Lymphoma (HL) enrolled on a clinical trial. Treatments received and survival status were collected together with other covariates such as demographics and clinical measurements. Our main task is to explore the potential of machine learning (ML) algorithms in a survival analysis context in order to improve over the Cox Proportional Hazard (CoxPH) model. We discuss the weaknesses of the CoxPH model we would like to improve upon and then we introduce multiple algorithms, from well-established ones to state-of-the-art models, that solve these issues. We then compare every model according to the concordance index and the brier score. Finally, we produce a series of recommendations, based on our experience, for practitioners that would like to benefit from the recent advances in artificial intelligence.

Keywords : machine learning, case study, survival analysis, Cox proportional hazard, survival trees, neural networks, variational auto-encoders

1 Introduction

There is increasing effort in medical research to applying ML algorithms to improve treatment decisions and predict patient outcomes. In this article, we want to explore the potential of ML algorithms to predict the outcome of children treated for Hodgkin Lymphoma. As we want to minimize the side effects of intensive chemotherapy or radiation therapy, a major clinical concern is how, for a given patient, we can select a treatment that eradicates the disease while keeping the intensity of the treatment, and the associated side effects, to a minimum.

In this article we will introduce multiple ML algorithms adapted to our needs and compare them with the Cox proportional hazard model. As it is the case with many data set within this field, the response variable, time until death or relapse, was right-censored for patients without events and the data set is of relatively small size (n=1712). From a ML perspective, this can be challenging. The response variable is right-censored for multiple observations but many ML techniques are not designed to deal with censored observations and thus it restricts the techniques we can include in our case study. Another challenged previously mentioned is that medical data sets are usually smaller than those used in ML applications and thus we will have to carefully select algorithms that could perform well in this context.

We will introduce the data set in section 2. In section 3 we will introduce the algorithms tested. Then, in section 4 we will present our experimental set up and our results. Finally, in section 5, we will discuss thoroughly the results, recommend further improvements and introduce open questions.

2 Data set

We have a data set of 1,712 patients, treated on the Children's Oncology Group trial AHOD0031, the largest randomized trial of pediatric HL ever conducted. Each observation represents a patient suffering from Hodgkin Lymphoma. For every patient, characteristics and symptoms have been collected as well as the treatment, for a total of 21 predictors. A table containing information on the predictors is in the appendix. The response is a time-to-event variable registered in number of days. We consider events to be either death or relapse. For patients without events, the response variable was right-censored at time of last seen, which is a well-known data structure in survival analysis. This data set and the data collecting technique are presented in detail by Friedman & al. (2014) who previously analyzed the same data set for other purposes.

3 Survival Analysis models

3.1 Benchmark : Cox Proportional Hazard Model

The Cox Proportional Hazard (CoxPH) model (Cox, 1972) serves as our benchmark model. It is widely used in medical sciences since it is robust, easy to use and produce highly interpretable results. It is a semi-parametric model that fits the hazard function, which represents the instantaneous rate of occurrence for the event of interest, using a partial likelihood function (Cox, 1975).

The CoxPH model fits the hazard function which contains two parts, a baseline hazard function of the time and a feature component which is a linear function of the predictors. The proportional hazard assumption assumes the time component and the feature component of the hazard function are proportional. In other words, the effect of the features is fixed through time. In the CoxPH model, the baseline hazard, which contains the time component, is usually unspecified so we cannot use the model directly to compute the hazard or to predict the survival function for a given set of covariates.

The main goal of this analysis is to test whether or not new ML models can outperform the CoxPH model. As ML models have shown great potential in many data analysis applications, it is important to test their potential to improve outcome prediction for cancer patients. We would like our selected models to improve upon at least one of the three following problems that are intrinsic to the CoxPH model. Problem (1): the proportional hazard assumption; we would like models that allow for feature effects to vary through time. Problem (2): the unspecified baseline hazard function; we would like models able to predict the survival function itself. Problem (3): the linear combination of features; we would like to use models that are able to grasp high order of interaction between the variable or non-linear combinations of the features.

3.2 Conventional statistical learning models

3.2.1 Regression models

The first model to be tested is a member of the CoxPH family. One way to capture interactions between predictors in linear models, and thus improve towards problem (3), is to include interaction terms. Since typical medical data sets contains few observations and many predictors, including all interactions usually leads to model saturation.

To deal with this issue we will use a variable selection model. Cox-Net (Simon, Friedman, Hastie, & Tibshirani, 2011) is an extension of the now well-know lasso regression (Hastie, Tibshirani, & Friedman, 2009) implemented in the glmnet package (J. Friedman, Hastie, & Tibshirani, 2010) and is the first model we will experiment with. The Cox-Net is a lasso regression-style model that shrinks some model coefficients to zero and thus insures the model is not saturated. The resulting model is as interpretable as the benchmark CoxPH model, but Cox-Net allows us to include all interactions in the base model without losing too many degrees of freedom.

Another approach based on regression models is the Multi-Task Logistic Regression (MTLR). Yu et al. (Yu, Greiner, Lin, & Baracos, 2011) proposed the MTLR model which quickly became a benchmark in the ML community for survival analysis and was cited by many authors (Luck, Sylvain, Cardinal, Lodi, & Bengio, 2017; Fotso, 2018; Zhao & Feng, 2019; Jinga et al., 2019). The proposed technique directly models the survival function by combining multiple local logistic regression models and considers the dependency of these models. By modelling the survival distribution with a sequence of dependent logistic regression, this model captures time-varying effects of features and thus the proportional hazard assumption is not needed. The model also grants the ability to predict survival time for individual patients. This model solves both problem (1) and (2). For our case study, we used the MTLR R-package (Haider, 2019) recently implemented by Haider.

3.2.2 Survival tree models

Decision trees (Breiman, Friedman, Olshen, & Stone, 1984) and random forests (Breiman, 1996, 2001) are known for their ability to detect and naturally incorporate high degrees of interactions among the predictors which is helpful towards problem (3). This family of models is well-established and make very few assumptions about the data set, making it a natural choice for our case study.

Multiple adaptations of decision trees were suggested for survival analysis and are commonly referred as survival trees. The idea suggested by many authors is to modify the splitting criteria of decision trees to accommodate for right-censored data. Based on previously published reviews of survival trees (LeBlanc & Crowley, 1995; Bou-Hamad, Larocque, & Ben-Ameur, 2011), we have selected four techniques for the case study.

One of the oldest survival tree models that was implemented in R (R Core Team, 2013) is the Relative Risk Survival Tree (Leblanc & Crowley, 1992). This survival tree algorithm uses most of the architecture established by CART (Breiman et al., 1984) but also borrows ideas from the CoxPH model. The model suggested by LeBlanc et al. assumes proportional hazards and partitions the data to maximize the difference in relative risk between regions. This technique was implemented in the rpart R-package (Therneau, Atkinson, & Ripley, 2017).

We also selected a few ensemble methods. To begin, Hothorn et al. (2004) proposed a new technique to aggregate survival decision trees that can produce conditional survival function, which solves problem (2). To predict the survival probabilities of a new observation, they use an ensemble of survival trees (Leblanc & Crowley, 1992) to determine a set of observations similar to the one in need of a prediction. They then use this set of observations to generate the Kaplan-Meier estimates for the new one. Their proposed technique is available in the ipred R-package (Peters & Hothorn, 2019). A year later, Hothorn et al. (2005; 2007) proposed a new ensemble technique able to produce log-survival time estimates instead. We will test this technique that is implemented in party R-package (Hothorn, Hornik, & Zeileis, 2006; Hothorn, Hornik, Strobl, & Zeileis, 2019).

Finally, the latest development in random forests for survival analysis is Random Survival Forests (Ishwaran, Kogalur, Blackstone, & Lauer, 2008). This implementation of a random survival forest was shown to be consistent (Ishwaran & Kogalur, 2010) and it comes with high-dimensional variable selection tools (Ishwaran, Kogalur, Gorodeski, Minn, & Lauer, 2010). This model was implemented in the randomForestSRC R-package (Ishwaran & Kogalur, 2019).

3.3 State-of-the-art models

3.3.1 Deep learning models

The first state-of-the-art model we will experiment with is built upon the most popular architecture of models in recent years: deep neural networks. Yu et al. (2011) MTLR model inspired many modifications (Luck et al., 2017; Fotso, 2018; Zhao & Feng, 2019; Jinga et al., 2019) in order to include a deep-learning component to the model. The main purpose is to allow for interactions and non-linear effect of the predictors. For example, Fotso (2018; 2019) suggested an extension of the MTLR where a deep neural networks parameterization replaces the linear parameterization and Luck et al. (2017) proposed a neural network model that produces two outputs: one is the risk and one is the probability of observing an event in a given time bin. Unfortunately, the authors for most of these techniques (Luck et al., 2017; Zhao & Feng, 2019; Jinga et al., 2019) did not provide either their code or a package which causes great reproducibility problems and leads to a serious accessibility issue for practitioners. The DeepSurv architecture (Katzman et al., 2018) proposed by Katzman et. al is a direct extension to the CoxPH model where the linear function of the covariance is replaced by a deep neural network. This allows the model to grasp high-order of interactions between predictors therefore solving problem (3). By allowing for interaction between covariates and the treatment the proposed model provides a treatment recommendation procedure. Finally, the authors provided a Python library available on the first author's GitHub (Katzman, 2017).

3.3.2 Latent-variable models

The final model is a latent-variable model based on the Variational Auto-Encoder (VAE) (Kingma & Welling, 2013; Kingma, 2017) architecture. Louizos et al. (2017) recently suggested a latent variable model for causal inference. The latent variables allow for a more flexible observed variable distribution and intuitively model the hidden patient status. Inspired by this model and by the recommendation of Nazbal et al. (2018) we implemented a latent variable model (Beaulac, Rosenthal, & Hodgson, 2018) that adapts the VAE architecture for the purposed of survival analysis. This Survival Analysis Variational Auto-Encoder (SAVAE) uses the latent space to represent the patient true sickness status and can produce individual patient survival function based on their respective covariates which should solve problem (1), (2) and (3).

4 Data analysis

4.1 Evaluation metrics

We will use two different metrics to evaluate the various algorithms, both are well established and they evaluate different properties of the models. First, the concordance index (Harrell, Lee, & Mark, 1996) is a metric of accuracy for the ordering of the predicted survival time or hazard. Second, the brier score (Graf, Schmoor, Sauerbrei, & Schumacher, 1999) is a metric similar to the mean squared error but adapted for right-censored observations.

4.1.1 Concordance Index

The concordance index (*c*-index) was proposed by Harrell et al. (1996). It is one of the most popular performance measures for survival problems (Steck, Krishnapuram, Dehing-oberije, Lambin, & Raykar, 2008; Chen, Kodell, Cheng, & Chen, 2012; Katzman, 2017) as it elegantly accounts for the censored data. It is defined as the proportion of all usable patient pairs in which the predictions and outcomes are concordant. Pairs are said to be concordant if the predicted event times have a concordant ordering with the observed event times.

Recently Steck et al. used the *c*-index directly as part of the optimization procedure (Steck et al., 2008), their paper also elegantly presents the *c*-index itself using graphical models as illustrated in figure 1. In their article it is defined as the fraction of all pairs of subjects whose predicted survival times are correctly ordered among all subjects that can actually be ordered. We expect a random classification algorithm to achieves a c-index of 0.5. The further from 0.5 the c-index is the more concordant pairs of predictions the model has produced. A c-index of 1 indicates perfect predicted order.

Figure 1 illustrates when we can compute the concordance for a pair of data points; this is represented by an arrow. We can evaluate the order of events if both events are observed. If one of the data points is censored, then concordance can be evaluated if the censoring for the censored point happens after the event for the observed point. If the reverse happens, if both points are censored or if both events happen exactly at the same time then we cannot evaluate the concordance for that pair.

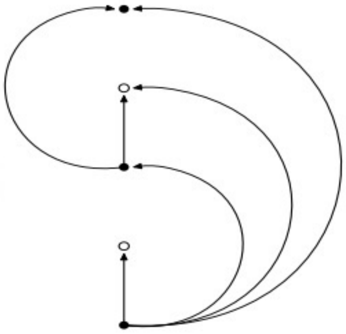


Figure 1: Steck et al.(2008) graphical representation of the c-index computation. Filled circle represents observed points and empty circle represents censored points. This figure illustrates the pairs of points for which an order of events can be established.

4.1.2 Brier Score

The Brier score established by Graf et al. (1999) is a performance metric inspired by the mean squared errors (MSE). For a survival model it is reasonable to try to predict $P(T > t|X = x) = S(t|X = x)$ the survival probabilities a time t for a patient with predictors x. In Graf's notation, $\hat{\pi}(t|x)$ is the predicted probability of survival at time t for a patient with characteristics x. These probabilities are used as predictions of the observed event $y = \mathbf{1}(T > t)$. If the data contains no censoring, the simplest definition of the Brier Score would be

$$\text{BS}(t) = \frac{1}{n}\sum_{i=1}^{n}(\mathbf{1}(T_i > t) - \hat{\pi}(t|x_i))^2. \tag{1}$$

Assuming we have a censoring survival distribution $G(t) = P(C > t)$ and an associated Kaplan-Meier estimated $\hat{G}(t)$. For a given fixed time t we are facing three different scenarios :

Case 1: $T_i > t$ and $\delta_i = 1$ or $\delta_i = 0$

Case 2: $T_i < t$ and $\delta_i = 1$

Case 3: $T_i < t$ and $\delta_i = 0$,

where $\delta_1 = 1$ if the event is observed and 0 if it is censored. For case 1, the event status is 1 since the patient is known to be alive at time t; the resulting contribution to the Brier score is $(1 - \hat{\pi}(t|x_i))^2$. For case 2, the event occurred before t and the event status is equal to $\mathbf{1}(T_i > t) = 0$ and thus the contribution is $(0 - \hat{\pi}(t|x_i))^2$. Finally, for case 3 the censoring occurred before t and thus the contribution to the Brier score cannot be calculated. To compensate for the loss of information due to censoring, the individual contributions have to be reweighed in a similar way as in the calculation of the Kaplan-Meier estimator leading to the following Brier Score

$$\mathrm{BS}^c(t) = \frac{1}{n}\sum_{i=1}^{n}\Big((0-\hat{\pi}(t|x_i))^2\mathbf{1}(T_i < t, \delta_i = 1)(1/\hat{G}(T_i)) + (1-\hat{\pi}(t|x_i))^2\mathbf{1}(T_i > t)(1/\hat{G}(t))\Big). \tag{2}$$

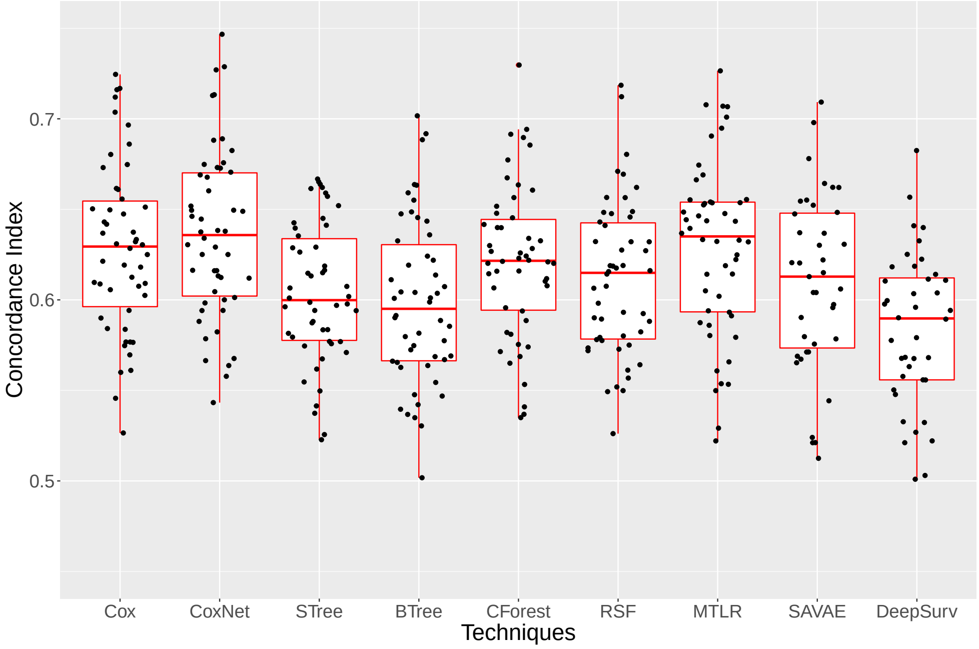
4.2 Comparative results

The data set introduced in section 2 was imported in both R (R Core Team, 2013) and Python (Van Rossum & Drake Jr, 1995). To evaluate the algorithms we randomly divided the data set into 1500 training observations and 212 testing observations. The models were fit using the training observations and the evaluation metrics were computed on the testing observations.

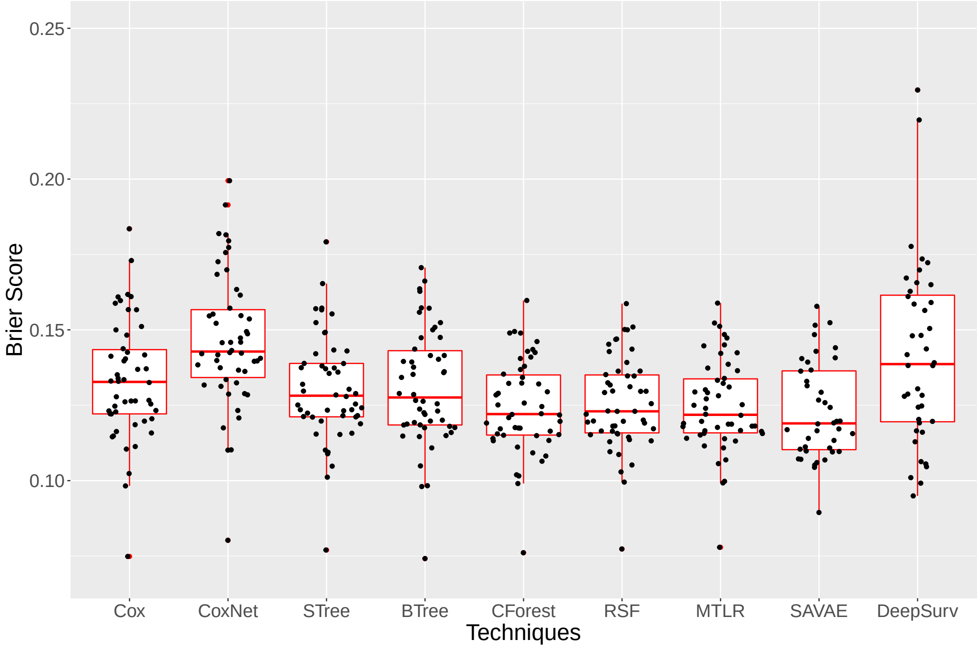
As mentioned in the previous sections, the CoxPH benchmark and the conventional statistical learning models were all tested in the R language (R Core Team, 2013). They were relatively easy to use with very little adjustment needed and clear and concise documentation. The computational speed of these algorithms was fast enough on a single CPU so that we could perform 50 trials. The state-of-the-art techniques needed a deeper understanding of the model as they contain many hyper-parameters that require calibration. They were also slower to run on a single CPU.

Figure 2a illustrates Sinaplots (Sidiropoulos, Sohi, Rapin, & Bagger, 2017) with associated Boxplots of the c-index for the CoxPH model and the 8 competitors. We used standard boxplots on the background since they are common and easy to understand. The sinaplots superposed on them represent the actual observed metric values and convey information about the distribution of the metrics for a given technique. As mentioned earlier c-index ranges from 0.5 to 1 where a c-index of 1 indicates perfect predicted order. According to figure 2, it seems no model clearly outperforms another. It seems like MTLR is the best-performing model but the difference is not statistically significant.

Since the Brier score is a metric inspired by the mean squared error, it ranges from 0 to 1 and the lower the Brier score is the better the technique. In figure 2b we once again observe that none of the new techniques drastically outperforms any CoxPH. SAVAE has the lowest Brier



(a) Boxplots and Sinaplots of the *c*-index (higher the better).



(b) Boxplots and Sinaplots of Brier Scores evaluated at 3 years (lower the better)

Figure 2: Results from our experiments.

score but the difference and significantly outperform the CoxPH benchmark. However, as shows in Figure 2b none of the new techniques drastically outperform CoxPH.

5 Takeaways and Recommendations

The previous section demonstrates that the new ML methods offers very little improvement compared to the benchmark CoxPH model according to our two designated performance metrics when patient clinical characteristics that are typically collected in clinical trials are used as predictor variables. This is an important result as we need to evaluate the abilities of ML techniques to solve real-life data problems, and to illuminate the changes in clinical data collection that will have to occur for ML methods to be used to greatest effect in assisting outcome prediction and treatment.

Similar results on real-life data sets are observed in article presenting methodologies (Fotso, 2018; Luck et al., 2017; Jinga et al., 2019) where the proposed techniques provide non-significant improvements over simple models such as CoxPH. Christodoulou et al. (2019) recently performed an exhaustive review of 927 articles that discuss the development of diagnostic or prognostic clinical prediction models for binary outcomes based on clinical data. The authors of the review noted the overall poor comparison methodologies and the lack of significant difference between a simple logistic regression and state-of-the-art ML techniques in most of recent years publications. These results are supported by Hand (2006) who discussed in detail the potential strength of the simple models compared to state-of-the-art ML models. This raises an important question our case study highlights: is it worth using more complex models for a slight improvement?

The alternative we proposed in section 3 are all more complicated than CoxPH in various ways. Most of the new techniques require deeper knowledge of the algorithm behaviors to correctly fix the many hyper-parameters. They can produce less interpretable results due to model complexity, and often require more computing power. Indeed, if the CoxPH model can be fit in seconds, most of the conventional statistical learning models take minutes to fit and the state-of-the-art models take hours. Finally, many of the new techniques are not widely accessible or standardized. As an open language, Python offers very little support to users and the libraries are not maintained, not standardized and come with dependency issues.

Hand (2006) demonstrates the high relative performances of extremely simple methods compared to complex ones and mathematically justifies his argument. He also discusses how these slight improvements over simple models might be undesirable as they might be attributed to overfitting which would cause reproducibility issues on new data sets. These slight improvements might also be artificial as they were achieved only because the inventors of these techniques were able to obtain through much effort the best performance from their own techniques and not the methods described by others. Overall if the improvements over simple techniques are small, perhaps they are simply not an improvement and this argument seems to be supported by both our case study and the recent review of Christodoulou et al. (2019). We recommend that practitioners keep their expectations low when it comes to some of these new models.

In contrast, significant improvements for diagnostic tasks have been accomplished using A.I. in recent years (Liu et al., 2017; Rodriguez-Ruiz et al., 2019; Rodríguez-Ruiz et al., 2019) and thus we ask ourselves what caused this difference ? There is a major difference in the style of data sets that were available. In the cited articles, images (mammographic, gigapixel pathology image, MRI scans) are analyzed using deep convolutional neural networks (CNN) (Goodfellow, Bengio, & Courville, 2016). Models such as CNN were developed because a special type of data was available and none of the current tools were equipped to analyze it. Conventional techniques such as logistic regression or CoxPH are not able to grasp the signal in images, which contains a large number of highly correlated predictors that individually contain close to no information but analyzed together contain a lot. As a matter of fact, the greatest strength of these models is that they are able to extract a lot of information from a rich, but complicated, data set.

In our case study, the *stratum* predictor was a binary predictor indicating if the patient had a rapid early response to the first rounds of chemotherapy. Computed-tomography (CT) scans of the affected regions were analyzed before and after the first round of treatments and this rich information was transformed into a simple binary variable. This practice is common: even in ongoing trials, patients' characteristics continue to be collected manually (often on paper forms), which dramatically limits the capacity to capture the full range of potentially useful data available for analysis. As new tools are established to extract information from ever growing, both in size and complexity, data sets, clinical trialists have to rethink how they gather data and transform it to make sure that no information is lost in order to utilize these new tools. It seems like extracting and keeping as much information as possible and having a data-centric approach where the model is designed to analyze a specific style of data were some of the factors in the success of CNNs.

6 Conclusion

In this article, we have identified a series of statistical and ML techniques that should alleviate some of the flaws of the well-known CoxPH model. These models were tested against a real-life data set and provided little to no improvement according the c-index and the Brier score. Although one might anticipate that these techniques would have increased our prediction abilities, instead the CoxPH performed comparably to modern models. These results are supported by other articles with similar findings.

It would be advantageous to try to theoretically understand when the new techniques should work and when they should not. As it currently stands, authors are not incentivized to discuss the weakness of their techniques and it actually slows scientific progress. It is imperative that we try to understand when some of the newest technique perform poorly and shed the light on why it is the case. It is also important to understand what made some of these new techniques successful. For example, it seems that CNNs were successful since the model was specifically built for images, a special type of data that was previously hard to handle but contained a large amount of information.

Funding details

Research reported in this work was supported by the Children's Oncology Group; by the National Cancer Institute of the National Institute of Health under the National Clinical Trials Network (NCTN) Operations Center Grant U10CA180886, the NCTN Statistics and Data Center Grant U10CA180899; by St. Baldrick's Foundation; the Natural Sciences and Engineering Research Council of Canada; and by the Ontario Graduate Scholarships.

The content is solely the responsibility of the authors and does not necessarily represent the official views of the Children's Oncology Group, the National Institute of Health, or St. Baldrick's Foundation.

Disclosure statement

No potential conflict of interest was reported by the authors.

Appendix

Variable	Type	Description
agedxyrs	Continuous	Age of the patient at the start of the treatment
gender	Binary	Biological gender
stage	Categorical	Cancer stage ranging from 1 to 4
b_symptoms	Binary	Presence of B symptoms
bulk_disease	Binary	Presence of Bulk disease
extralymphatic_disease	Binary	Presence of Extralymphatic disease
fever	Binary	Presence of recurrent fever
night_sweats	Binary	Presence of night sweats
weight_loss	Binary	Presence of significant weight loss (> 10%)
nodal_aggregate	Binary	Presence of a nodal aggregate
mediastinal_mass	Binary	Presence of a mediastinal mass
esron	Continuous	Erthroctye sedimentation rate (mm/hr)
istnon	Continuous	Number of involved nodal sites
histology	Categorical	Histology (LP,LD,NS,MC, unknown)
albon	Continuous	Albumin (g/dL)
hgbon	Continuous	Hemoglobin(g/dL)
amend	Binary	
stratum	Binary	Rapid early response to first treatment
morpho_icdo	Categorical	ICD-O Morphology codes
RT	Binary	Treatment variable: Radiotherapy
DECA	Binary	Treatment variable: Intensive Chemotherapy

Table 1: Predictor variables and description

References

Beaulac, C., Rosenthal, J. S., & Hodgson, D. (2018). A deep latent-variable model application to select treatment intensity in survival analysis. *Proceedings of the Machine Learning for Health (ML4H) Workshop at NeurIPS 2018, 2018*.

Bou-Hamad, I., Larocque, D., & Ben-Ameur, H. (2011, 01). A review of survival trees. *Statistics Surveys*, *5*. doi: 10.1214/09-SS047

Breiman, L. (1996). Bagging predictors. *Machine Learning*, *24*(2), 123–140. Retrieved from http://dx.doi.org/10.1007/BF00058655 doi: 10.1007/BF00058655

Breiman, L. (2001). Random forests. *Machine Learning*, *45*(1), 5–32. Retrieved from http://dx.doi.org/10.1023/A:1010933404324 doi: 10.1023/A:1010933404324

Breiman, L., Friedman, J., Olshen, R., & Stone, C. (1984). *Classification and Regression Trees*. Monterey, CA: Wadsworth and Brooks.

Chen, H.-C., Kodell, R. L., Cheng, K. F., & Chen, J. J. (2012, Jul 23). Assessment of performance of survival prediction models for cancer prognosis. *BMC Medical Research Methodology*, *12*(1), 102. Retrieved from https://doi.org/10.1186/1471-2288-12-102 doi: 10.1186/1471-2288-12-102

Christodoulou, E., Ma, J., Collins, G. S., Steyerberg, E. W., Verbakel, J. Y., & Calster, B. V. (2019). A systematic review shows no performance benefit of machine learning over logistic regression for clinical prediction models. *Journal of Clinical Epidemiology*, *110*, 12 - 22. Retrieved from http://www.sciencedirect.com/science/article/pii/S0895435618310813 doi: https://doi.org/10.1016/j.jclinepi.2019.02.004

Cox, D. R. (1972). Regression models and life-tables. *Journal of the Royal Statistical Society. Series B (Methodological)*, *34*(2), 187–220. Retrieved from http://www.jstor.org/stable/2985181

Cox, D. R. (1975, 08). Partial likelihood. *Biometrika*, *62*(2), 269-276. Retrieved from https://doi.org/10.1093/biomet/62.2.269 doi: 10.1093/biomet/62.2.269

Fotso, S. (2018, Jan). Deep Neural Networks for Survival Analysis Based on a Multi-Task Framework. *arXiv e-prints*, arXiv:1801.05512.

Fotso, S., et al. (2019). *PySurvival: Open source package for survival analysis modeling*. Retrieved from https://www.pysurvival.io/

Friedman, D. L., Chen, L., Wolden, S., Buxton, A., McCarten, K., FitzGerald, T. J., … Schwartz, C. L.

(2014). Dose-intensive response-based chemotherapy and radiation therapy for children and adolescents with newly diagnosed intermediate-risk hodgkin lymphoma: A report from the children's oncology group study ahod0031. *Journal of Clinical Oncology*, *32*(32), 3651-3658. Retrieved from https://doi.org/10.1200/JCO.2013.52.5410 (PMID: 25311218) doi: 10.1200/JCO.2013.52.5410

Friedman, J., Hastie, T., & Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, *33*(1), 1–22. Retrieved from http://www.jstatsoft.org/v33/i01/

Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep learning*. MIT Press. (http://www.deeplearningbook.org)

Graf, E., Schmoor, C., Sauerbrei, W., & Schumacher, M. (1999, 09). Assessment and comparison of prognostic classification schemes for survival data. *Statistics in medicine*, *18*, 2529-45. doi: 10.1002/(SICI)1097-0258(19990915/30)18:17/183.0.CO;2-5

Haider, H. (2019). Mtlr: Survival prediction with multi-task logistic regression [Computer software manual]. Retrieved from https://CRAN.R-project.org/package=MTLR (R package version 0.2.1)

Hand, D. J. (2006, 02). Classifier technology and the illusion of progress. *Statist. Sci.*, *21*(1), 1–14. Retrieved from https://doi.org/10.1214/088342306000000060 doi: 10.1214/088342306000000060

Harrell, F. E., Lee, K. L., & Mark, D. B. (1996). Multivariable prognostic models: Issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. *Statistics in Medicine*, *15*(4), 361-387. doi: 10.1002/(SICI)1097-0258(19960229)15:4<361::AID-SIM168>3.0.CO;2-4

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning* (2nd ed.). Springer.

Hothorn, T., Bühlmann, P., Dudoit, S., Molinaro, A., & Van Der Laan, M. J. (2005, 12). Survival ensembles. *Biostatistics*, *7*(3), 355-373. Retrieved from https://doi.org/10.1093/biostatistics/kxj011 doi: 10.1093/biostatistics/kxj011

Hothorn, T., Hornik, K., Strobl, C., & Zeileis, A. (2019). party: A laboratory for recursive partytioning [Computer software manual]. Retrieved from https://cran.r-project.org/web/packages/party/index.html (R package version 1.3-3)

Hothorn, T., Hornik, K., & Zeileis, A. (2006). Unbiased recursive partitioning: A conditional inference framework. *Journal of Computational and Graphical Statistics*, *15*(3), 651-674. Retrieved from http://dx.doi.org/10.1198/106186006X133933 doi: 10.1198/106186006X133933

Hothorn, T., Lausen, B., Benner, A., & Radespiel-Tröger, M. (2004). Bagging survival trees. *Statistics in Medicine*, *23*(1), 77-91. Retrieved from https://onlinelibrary.wiley.com/doi/abs/10.1002/sim.1593 doi: 10.1002/sim.1593

Ishwaran, H., & Kogalur, U. (2019). Fast unified random forests for survival, regression, and classification (rf-src) [Computer software manual]. manual. Retrieved from https://cran.r-project.org/package=randomForestSRC (R package version 2.9.1)

Ishwaran, H., & Kogalur, U. B. (2010). Consistency of random survival forests. *Statistics & Probability Letters*, *80*(13), 1056 - 1064. Retrieved from http://www.sciencedirect.com/science/article/pii/S0167715210000672 doi: https://doi.org/10.1016/j.spl.2010.02.020

Ishwaran, H., Kogalur, U. B., Blackstone, E. H., & Lauer, M. S. (2008, 09). Random survival forests. *Ann. Appl. Stat.*, *2*(3), 841–860. Retrieved from http://dx.doi.org/10.1214/08-AOAS169 doi: 10.1214/08-AOAS169

Ishwaran, H., Kogalur, U. B., Gorodeski, E. Z., Minn, A. J., & Lauer, M. S. (2010). High-dimensional variable selection for survival data. *Journal of the American Statistical Association*, *105*(489), 205-217. Retrieved from https://doi.org/10.1198/jasa.2009.tm08622 doi: 10.1198/jasa.2009.tm08622

Jinga, B., Zhangh, T., Wanga, Z., Jina, Y., Liua, K., Qiua, W., … Lia, C. (2019). A deep survival analysis method based on ranking. *Artificial Intelligence in Medicine*, *98*, 1 - 9. Retrieved from http://www.sciencedirect.com/science/article/pii/S0933365718305992 doi: https://doi.org/10.1016/j.artmed.2019.06.001

Katzman, J. (2017). *Deepsurv: Personalized treatment recommender system using a cox proportional hazards deep neural network.* Retrieved from https://github.com/jaredleekatzman/DeepSurv

Katzman, J., Shaham, U., Cloninger, A., Bates, J., Jiang, T., & Kluger, Y. (2018, 12). Deepsurv: Personalized treatment recommender system using a cox proportional hazards deep neural network. *BMC Medical Research Methodology*, *18*. doi: 10.1186/s12874-018-0482-1

Kingma, D. P. (2017). *Variational inference & deep learning : A new synthesis* (Unpublished doctoral dissertation). Universiteit van Armsterdam.

Kingma, D. P., & Welling, M. (2013, December). Auto-Encoding Variational Bayes. *ArXiv e-prints*.

LeBlanc, M., & Crowley, J. (1995). A review of tree-based prognostic models. *Recent Advances in Clinical Trial Design and Analysis*, *75*, 113-124.

Leblanc, M. E., & Crowley, J. P. (1992). Relative risk trees for censored survival data. *Biometrics*, *48 2*, 411-25.

Liu, Y., Gadepalli, K., Norouzi, M., Dahl, G. E., Kohlberger, T., Boyko, A., ... Stumpe, M. C. (2017, Mar). Detecting Cancer Metastases on Gigapixel Pathology Images. *arXiv e-prints*, arXiv:1703.02442.

Louizos, C., Shalit, U., Mooij, J., Sontag, D., Zemel, R., & Welling, M. (2017, May). Causal Effect Inference with Deep Latent-Variable Models. *ArXiv e-prints*.

Luck, M., Sylvain, T., Cardinal, H., Lodi, A., & Bengio, Y. (2017). Deep learning for patient-specific kidney graft survival analysis. *CoRR*, *abs/1705.10245*. Retrieved from `http://arxiv.org/abs/1705.10245`

Nazábal, A., Olmos, P. M., Ghahramani, Z., & Valera, I. (2018). Handling incomplete heterogeneous data using vaes. *ArXiv*, *abs/1807.03653*.

Peters, A., & Hothorn, T. (2019). ipred: Improved predictors [Computer software manual]. Retrieved from `https://CRAN.R-project.org/package=ipred` (R package version 0.9-9)

R Core Team. (2013). R: A language and environment for statistical computing [Computer software manual]. Vienna, Austria. Retrieved from `http://www.R-project.org/`

Rodríguez-Ruiz, A., Krupinski, E., Mordang, J.-J., Schilling, K., Heywang-Köbrunner, S. H., Sechopoulos, I., & Mann, R. M. (2019). Detection of breast cancer with mammography: Effect of an artificial intelligence support system. *Radiology*, *290*(2), 305-314. Retrieved from `https://doi.org/10.1148/radiol.2018181371` (PMID: 30457482) doi: 10.1148/radiol.2018181371

Rodriguez-Ruiz, A., Lång, K., Gubern-Merida, A., Broeders, M., Gennaro, G., Clauser, P., ... Sechopoulos, I. (2019, 03). Stand-Alone Artificial Intelligence for Breast Cancer Detection in Mammography: Comparison With 101 Radiologists. *JNCI: Journal of the National Cancer Institute*, *111*(9), 916-922. Retrieved from `https://doi.org/10.1093/jnci/djy222` doi: 10.1093/jnci/djy222

Sidiropoulos, N., Sohi, S. H., Rapin, N., & Bagger, F. O. (2017). *sinaplot: an enhanced chart for simple and truthful representation of single observations over multiple classes.* Retrieved from `https://cran.r-project.org/web/packages/sinaplot/`

vignettes/SinaPlot.html

Simon, N., Friedman, J., Hastie, T., & Tibshirani, R. (2011). Regularization paths for cox's proportional hazards model via coordinate descent. *Journal of Statistical Software, Articles*, *39*(5), 1–13. Retrieved from https://www.jstatsoft.org/v039/i05 doi: 10.18637/jss.v039.i05

Steck, H., Krishnapuram, B., Dehing-oberije, C., Lambin, P., & Raykar, V. C. (2008). On ranking in survival analysis: Bounds on the concordance index. In J. C. Platt, D. Koller, Y. Singer, & S. T. Roweis (Eds.), *Advances in neural information processing systems 20* (pp. 1209–1216). Curran Associates, Inc. Retrieved from http://papers.nips.cc/paper/3375-on-ranking-in-survival-analysis-bounds-on-the-concordance-index.pdf

Strobl, C., Boulesteix, A.-L., Zeileis, A., & Hothorn, T. (2007). Bias in random forest variable importance measures: Illustrations, sources and a solution. *BMC Bioinformatics*, *8*(1), 25. Retrieved from http://dx.doi.org/10.1186/1471-2105-8-25 doi: 10.1186/1471-2105-8-25

Therneau, T., Atkinson, B., & Ripley, B. (2017). rpart: Recursive partitioning and regression trees [Computer software manual]. Retrieved from https://CRAN.R-project.org/package=rpart (R package version 4.1-11)

Van Rossum, G., & Drake Jr, F. L. (1995). *Python tutorial*. Centrum voor Wiskunde en Informatica Amsterdam, The Netherlands.

Yu, C.-N., Greiner, R., Lin, H.-C., & Baracos, V. (2011). Learning patient-specific cancer survival distributions as a sequence of dependent regressors. In J. Shawe-Taylor, R. S. Zemel, P. L. Bartlett, F. Pereira, & K. Q. Weinberger (Eds.), *Advances in neural information processing systems 24* (pp. 1845–1853). Curran Associates, Inc. Retrieved from http://papers.nips.cc/paper/4210-learning-patient-specific-cancer-survival-distributions-as-a-sequence-of-dependent-regressors.pdf

Zhao, L., & Feng, D. (2019, Aug). DNNSurv: Deep Neural Networks for Survival Analysis Using Pseudo Values. *arXiv e-prints*, arXiv:1908.02337.