

In with the old, in with the new: Machine Learning for Time to Event Biomedical Research

Ioana Danciu^{1,2,*}, Greeshma Agasthya¹, Janet P. Tate^{3,4}, Mayanka Chandra-Shekar¹, Ian Goethert¹, Olga S. Ovchinnikova¹, Benjamin H. McMahon⁵, Amy C. Justice^{3,4,6}

¹ Advanced Computing for Health Sciences Group, Oak Ridge National Laboratory, Oak Ridge, TN, USA

² Department of Biomedical Informatics, Vanderbilt University, Nashville, TN, USA

³ Department of Veterans Affairs Connecticut Healthcare System, West Haven, CT, USA

⁴ Yale School of Medicine, New Haven, CT, USA

⁵ Theoretical Biology Group, Los Alamos National Laboratory, Los Alamos, NM, USA

⁶ Yale School of Public Health, New Haven, CT, USA

*Corresponding author

Abstract –

Objective: The predictive modeling literature for biomedical applications is dominated by biostatistical methods for survival analysis, and more recently some out of the box machine learning approaches. In this paper we show a presentation of a machine learning method appropriate for time-to-event modeling in the area of prostate cancer long term disease progression.

Materials and Methods: Using XGBoost adapted to long term disease progression, we developed a predictive model for 118,788 patients with localized prostate cancer at diagnosis from the Department of Veterans Affairs (VA). Our model accounted for patient censoring.

Results: Harrell's c-index for our model using only features available at the time of diagnosis was 0.757 95% CI [0.756, 0.757].

Discussion: Our results show that machine learning methods like XGBoost can be adapted to use accelerated failure time with censoring to model long term risk of disease progression. The long median survival justifies and requires censoring.

Conclusion: Overall, we show that an existing ML learning approach can be used for accelerated failure time (AFT) outcome modelling in prostate cancer, and more generally for other chronic diseases with long observation times.

I. INTRODUCTION

Historically, predictive models have been dominated by classical statistical approaches. In recent years, access to an increasing volume and diversity of biomedical data, as well as advances in computational capabilities have generated significant interest in application of machine learning (ML) for biomedical research[1]. Machine learning techniques offer several advantages over statistical approaches: they can account for large numbers of independent variables and multiple outcomes, making fewer assumptions regarding the shape of the predictor and hazard functions. Machine learning approaches also have drawbacks when

they do not include censoring information. For chronic diseases with long observation timeframes, loss to follow-up is of serious concern. Statistical approaches account for these situations by capturing the observation window and censoring time. In this paper, we present a domain-aware predictive modeling approach based on XGBoost[2] for prostate cancer disease progression using machine learning that considers time-to-event with censoring for a direct comparison with traditional biostatistical approaches.

II. RELATED WORK

Prostate cancer is the leading malignancy in male veterans[3] and the second in United States males[4]. Annually, over 13,000 new cases are diagnosed in the veteran population[3]. Fortunately, 80% of new prostate cancer diagnoses are non-metastatic. In this population, treatment options to improve long-term survival and maximize quality of life are extremely complex[5], and there is significant need for data driven predictive models to inform the course of therapy. Over the years, several models of prostate cancer survival have been developed to address the need for survival prediction tools, nomograms, and calculators available to both patients and physicians. The Memorial Sloan Kettering nomograms[6] are widely used in the prostate cancer community and were built using data from 10,000 prostate cancer cases. The PREDICT prostate[7] tool used multivariate Cox models applied to 10,089 non-metastatic prostate cancer cases in England and externally validated on 2,546 cases in Singapore. These tools were mostly developed using traditional biostatistical approaches that have dominated predictive modeling since the 1980s. Recently, the availability of the MIMIC dataset[8] has resulted in more machine and deep learning approaches being used out of the box for biomedical predictive modeling[9]. However, adapting machine and deep learning techniques to time to event with censoring has not received as much attention from the research community, and methods for doing so are still being investigated[10]. Outcome modeling with random survival trees (RSF)[11] is a machine learning non-parametric alternative to the traditional Cox proportional hazards model. Due to the nature of the random forest it is based on, RSF is more sensitive to unbalanced datasets, such as those used for this analysis, compared to XGBoost. DeepSurv[12] is an implementation of a Cox proportional hazards model using a deep neural network. [13] discusses adapting generative adversarial networks for time to event modeling with censoring. In [14] the investigators used XGBoost without time to event with censoring for mortality prediction 10 years after diagnosis in a 76,693 patient cohort recruited as part of a multicenter project in the US. The studies discussed in this section also focus on a singular outcome (mortality).

III. METHODS

This study has been approved by the VA Central Institutional Review Board (IRB).

A. Cohort selection

Our cohort consisted of patients with localized prostate cancer at diagnosis from the Department of Veterans Affairs (VA) Corporate Data Warehouse (CDW) cancer registry. We defined localized prostate cancer using the tumor, node, metastases (TNM) staging system: having clinical N0 and M0. We included only patients with a Gleason score clinical between 6 and 10, and restricted our study to individuals with a baseline prostate specific antigen (PSA) between 1 and 100. We also excluded patients without a biopsy confirmed diagnosis and those who had another primary cancer either before or after the prostate cancer diagnosis.

B. Predictors

The 13 predictors measured at or before primary prostate cancer diagnosis, fell in two categories: (i) A set of independent variables abstracted from the CDW cancer registry characterizing patient demographics (age at diagnosis, race, ethnicity) and disease staging (Gleason score clinical, AJCC stage group, SEER summary stage and a computed stage value from the registry (TNM) variables); (ii) Prostate specific antigen (PSA) values, aggregated into minimum, maximum, average, density, standard deviation values across the 5-year period prior to diagnosis. In addition, we included the last PSA before diagnosis adjudicated over the last year, the penultimate PSA over the last 5 years, and their rate of change.

C. Outcome

The outcome was a composite in days from primary diagnosis, consisting of cancer related death from the National Death Index as well as registry documentation, a PSA > 50, and inpatient and outpatient ICD codes indicative of metastatic disease. In our dataset more than half of the population was censored, therefore we could not calculate a median survival/disease progression, and we reported survival in 5 year intervals.

D. Censoring

We censored all patients one year after their last PSA value or 12/31/2017, whichever came first.

E. Computational methods

We used a well-known machine learning algorithm, XGBoost[15] adapted for accelerated failure time[2]. The modification uses XGBoost to fit a good tree ensemble $\tau(x)$. It expresses the AFT model of the form:

$$\ln(Y) = \tau(x) + \text{noise},$$

where Y is the output label, x is the input feature vector. Since the noise is a random variable, the goal of xgboost is maximizing $\ln(Y)$ by fitting a good tree ensemble $\tau(x)$. Unlike traditional survival analysis methodologies that compute a hazard ratio, our machine learning algorithm calculated time to event survival. The model training and 5-fold cross validation used 80% of the

dataset. The remaining 20% was used solely for independent testing of the trained model to compute the performance metrics. We performed grid search to determine the `max_depth`, `learning_rate`, `aft_loss_distribution_scale`, `min_child_weight`, `n_estimators` hyperparameters needed for the model. Our step by step methodology followed the approach detailed in[16].

F. Explainability

For improving the explainability of the machine learning approach, we derived feature importance measures using a tree solution for SHAP (SHapley Additive explanation)[17], a method that is both consistent and accurate[18]. The SHAP value of each feature for each prediction is the marginal contribution of that feature to the output. In addition to averaging these contributions for all patients in our dataset, we also generated instance level predictions for two scenarios: a patient with more aggressive disease at diagnosis and one with lower grade disease.

G. Metrics

To determine the discrimination power of our algorithms, we used two metrics: the modified Harrell’s c-statistic, and the accuracy. We modified the Harrell’s c-statistic calculation[19] to use the time to event survival predictions generated by our model instead of the risk to generate the list of concordant pairs and ties.

IV. RESULTS

A. Cohort selection

Figure 1 below shows the cohort selection procedure.

Figure 1: Cohort selection

B. Population

Table 1 shows the population demographics in detail. Our cohort is dominated by 55-74 year olds, consistent with previous studies and current screening guidelines. The dates of diagnosis span a period of 14 years from 2002 to 2016. Our population consists of a higher percentage of African Americans compared to the VA overall[20]. The Gleason scores, stage, and PSA at diagnosis are consistent with localized disease. More than 50% of the population is outcome-free at 5 years with 14.19% observed outcome-free 10 years after diagnosis.

Table 1: Population demographics

		Train-validation set (n=94,608)	Test set (n=24,180)
Demographics			
Age at diagnosis <i>Patient count (% of total)</i>			
	<55	5745 (6.07%)	1540 (6.36%)
	55-64	36323 (38.39%)	9231 (38.17%)
	65-74	39674 (41.93%)	10134 (41.91%)
	>=75	12866 (13.59%)	3275 (13.54%)
Race <i>Patient count (% of total)</i>			
	African-American	27239 (28.79%)	7038 (29.1%)
	East Asian	185 (0.19%)	47 (0.19%)
	South Asian	28 (0.02%)	5 (0.02%)
	Polynesian/Hawaiian	153 (0.16%)	36 (0.14%)
	White	63903 (67.54%)	16264 (67.26%)
	Other/Unknown	3100 (3.27%)	790 (3.26%)
Ethnicity <i>Patient count (% of total)</i>			
	Hispanic or Latino	5268 (5.56%)	1318 (5.45%)
	Not Hispanic or Latino	89340 (94.43%)	22862 (94.54%)
Staging variables			
T-stage <i>Patient count (% of total)</i>			
	1A	726 (0.76%)	188 (0.19%)
	1B	419 (0.44%)	119 (0.49%)
	1C	64691 (68.37%)	16597 (68.63%)
	1	335 (0.35%)	79 (0.32%)
	2A	10087 (10.66%)	2549 (10.54%)
	2B	4230 (4.47%)	1115 (4.61%)
	2C	8588 (9.07%)	2167 (8.96%)
	2	3749 (3.96%)	936 (3.87%)
	3	1598 (1.68%)	385 (1.59%)
	4	181 (0.19%)	44 (0.18%)
	missing	4 (0%)	1 (0%)
Gleason score <i>Patient count (% of total)</i>			
	6	38937 (41.15%)	9919 (41.02%)
	7	39828 (42.09%)	10332 (42.72%)
	8	9223 (9.74%)	2280 (9.42%)
	9	6078 (6.42%)	1509 (6.24%)

	10	542 (0.57%)	140 (0.57%)
PSA at diagnosis <i>mean (std)</i>		9.777 (10.844)	9.706 (10.733)
Outcome			
Composite outcome <i>Patient Count (% of total)</i>			
	<5 years	5545 (5.86%)	1419 (5.86%)
	5-10 years	2501 (2.64%)	632 (2.61%)
	>10 years	594 (0.62%)	144 (0.59%)
Censoring <i>Patient Count (% of total)</i>			
	<5 years	39766 (42.03%)	10197 (42.17%)
	5-10 years	32768 (34.63%)	8331 (34.45%)
	>10 years	13434 (14.19%)	3457 (14.29%)

C. Performance metrics

The model shows robust performance from the validation to the test dataset. The test c-index is 0.757, with a narrow confidence interval. Table 2 shows the accuracy and c-index metrics on the validation and test datasets.

Table 2: Performance metrics

Validation accuracy [95% CI]	Validation c-index [95% CI]	Test accuracy [95% CI]	Test c-index [95% CI]
0.889	0.764	0.891	0.757
[0.888, 0.889]	[0.759, 0.768]	[0.890, 0.891]	[0.756, 0.757]

D. Explainability

Figure 2 shows the average impact of the independent variables on model output magnitude. The Gleason score is the most important overall feature, and the Hispanic/Latino ethnicity has the lowest impact.

Figure 2. SHAP predictor importance (average impact on model output magnitude)

At an individual patient level, the feature importance varies, as shown in Figures 3a and 3b. For a hypothetical patient 1 (Figure 3a), with more aggressive disease at diagnosis, the higher Gleason score, PSA and stage reduce the survival time ($f(x)$) from the average (base value).

Figure 3a. Instance level predictor importance for hypothetical patient 1 with more aggressive disease at diagnosis- **Hypothetical patient 1 timeline**

Figure 3a. Instance level predictor importance for hypothetical patient 1 with more aggressive disease at diagnosis- **Outcome prediction**

Figure 3b. Instance level predictor importance for hypothetical patient 2 with less aggressive disease at diagnosis-**Hypothetical patient 2 timeline**

Figure 3b. Instance level predictor importance for hypothetical patient 2 with less aggressive disease at diagnosis **Outcome prediction**

For hypothetical patient 2 (Figure 3b) with less aggressive disease at diagnosis, the age, Gleason score, stage, minimum, maximum and average PSA values before diagnosis drive an increase in the survival, while the last PSA and the standard deviation of the PSA values contribute to a reduction.

V. DISCUSSION

Overall, we show an approach that uses an existing machine learning method adapted for accelerated failure time with censoring to model long-term risk of disease progression. In our study, 14% of the population is censored more than 10 years after diagnosis indicating that loss to follow-up is an important consideration. Especially when they align more closely with traditional biostatistical methods, the value proposition of machine learning over other approaches are multifold. First, unlike Cox, we are not making any assumptions regarding the contributions of predictors in our survival function over time. Secondly, machine learning algorithms like XGBoost are well-suited computationally for problems with large numbers of predictors. Our model was intentionally simpler with only 13 predictors, but our methods are scalable to problems with hundreds of independent variables. Machine learning approaches also provide better incorporation of interaction terms: while they can be included in regression, this process is cumbersome as the range of possibilities is broad, even for pairwise interactions. Additionally, having the capability to analyze the predictor importance, will allow us to factor in these hundreds of predictors across the entire electronic health record.

Since SHAP values depend on the choice of predictors, their degree of collinearity and ultimately the type of machine learning algorithm, correctly contextualizing the predictor importance given the clinical problem, will allow us to find the most relevant ones.

The predictor importance plots in Figures 2, and 3 highlight another benefit of machine learning models. Whereas in traditional regression models, the coefficients for each feature are static, our modeling paradigm allows for recommendations tailored to the individual patient. Traditional biostatistical approaches try to account for more personalized predictions using subgroup analysis, capability also accounted for by our machine learning approach that generates the decision tree splits as part of the algorithm training.

Our c-statistic is lower than other studies of long term outcomes for prostate cancer patients: 0.84[7] and 0.8[14]. We hypothesize this is due to a number of factors including (1) our prediction happens at the time of diagnosis not accounting for any treatment modalities that could significantly impact disease progression, as reported in other studies[14][21]. The long follow-up time creates a long window of opportunity for patients with similar profiles at diagnosis to follow different treatment arms, and therefore have different outcomes; (2) our intent was to build a simple model using a large cohort with long follow-up times, that would enable clinicians to make informative decisions with the limited data available to them at the time of diagnosis. Consequently, we used a small set of predictors compared to other studies[14], not accounting for the medical history, physical activity or socio-economic status; (3) unlike other studies that focus on a single outcome, mortality (cancer related or all-cause), we used a composite that consists of cancer mortality and metastatic disease. In our scenario two patients with similar profiles at diagnosis, could have significant times to outcome if one is missing an ICD code indicating metastasis; (4) our dataset includes patients diagnosed over a period of 14 years, time in which coding standards changed for our independent and dependent variables; (5) training also suffered from class imbalance problems because the cohort who had the outcome before censoring was about 10% of the total number of patients.

VI. CONCLUSION

Machine and deep learning methodologies that can capture the full complexities of datasets are both needed and underrepresented in the literature and current biomedical practice. Our methods scale to complex problems with thousands of predictors and are applicable to a variety of chronic conditions with long observation periods, for which loss to follow-up is an important consideration. In the current data-rich biomedical environment, domain-specific computational methods such as machine and deep learning that scale to large longitudinal datasets hold the promise for future discoveries.

VII. ACKNOWLEDGEMENTS

Funding: This research was supported by award MVP017 from the Million Veteran Program, Office of Research and Development, Veterans Health Administration.

This research used resources of the Knowledge Discovery Infrastructure at the Oak Ridge National Laboratory, which is supported by the Office of Science of the U.S. Department of Energy under Contract No. DE-AC05-00OR22725 and the Department of Veterans Affairs Office of Information Technology Inter-Agency Agreement with the Department of Energy under IAA No. VA118-16-M-1062.

Competing Interests: The authors have no competing interests to declare.

Contributorship: ID proposed the study design, methods, wrote the first draft and oversaw subsequent revisions of the paper. ID and IG generated the data used for this study. All authors contributed to iterative study design and manuscript editing.

Data Availability: Final data sets underlying this study cannot be shared outside the VA, except as required under the Freedom of Information Act (FOIA), per VA policy. However, upon request through the formal mechanisms in place and pending approval from the VHA Office of Research Oversight (ORO), a de-identified, anonymized dataset underlying this study can be created and shared.

General acknowledgements: This manuscript has been authored by UT-Battelle, LLC under Contract No. DE-AC05-00OR22725 with the U.S. Department of Energy. This publication does not represent the views of the Department of Veteran Affairs, the Department of Energy or the United States Government. This project used data from the Center of Excellence for Mortality Data Repository, Joint Department of Veterans Affairs (VA) and Department of Defense (DoD) Suicide Data Repository– National Death Index (NDI). <http://vawww.virec.research.va.gov/Mortality/Overview.htm>; Extract < November 20, 2020>.

The authors also wish to acknowledge the support of the larger partnership. Most importantly, the authors would like to thank and acknowledge the veterans who chose to get their care at the VA.

REFERENCES

- [1] J. Goecks, V. Jalili, L. M. Heiser, and J. W. Gray, “How Machine Learning Will Transform Biomedicine,” *Cell*, vol. 181, no. 1, pp. 92–101, Apr. 2020, doi: 10.1016/j.cell.2020.03.022.
- [2] A. Barnwal, H. Cho, and T. D. Hocking, “Survival regression with accelerated failure time model in XGBoost,” *ArXiv200604920 Cs Stat*, Jun. 2020, Accessed: Mar. 05, 2021. [Online]. Available: <http://arxiv.org/abs/2006.04920>
- [3] L. L. Zullig *et al.*, “Cancer Incidence among Patients of the United States Veterans Affairs (VA) Healthcare System: 2010 Update,” *Mil. Med.*, vol. 182, no. 7, pp. e1883–e1891, Jul. 2017, doi: 10.7205/MILMED-D-16-00371.
- [4] CDCBreastCancer, “Prostate Cancer Statistics,” *Centers for Disease Control and Prevention*, Jun. 08, 2021. <https://www.cdc.gov/cancer/prostate/statistics/index.htm> (accessed Aug. 14, 2021).
- [5] T. J. Wilt *et al.*, “Follow-up of Prostatectomy versus Observation for Early Prostate Cancer.,” *N. Engl. J. Med.*, vol. 377, no. 2, pp. 132–142, Jul. 2017, doi: 10.1056/NEJMoa1615869.
- [6] “Prostate Cancer Nomograms | Memorial Sloan Kettering Cancer Center.” <https://www.mskcc.org/nomograms/prostate> (accessed Aug. 16, 2021).
- [7] D. R. Thurtle, D. C. Greenberg, L. S. Lee, H. H. Huang, P. D. Pharoah, and V. J. Gnanapragasam, “Individual prognosis at diagnosis in nonmetastatic prostate cancer: Development and external validation of the PREDICT Prostate multivariable model,” *PLoS Med.*, vol. 16, no. 3, p. e1002758, Mar. 2019, doi: 10.1371/journal.pmed.1002758.
- [8] A. E. W. Johnson *et al.*, “MIMIC-III, a freely accessible critical care database,” *Sci. Data*, vol. 3, p. 160035, May 2016, doi: 10.1038/sdata.2016.35.

- [9] A. E. Johnson, D. J. Stone, L. A. Celi, and T. J. Pollard, “The MIMIC Code Repository: enabling reproducibility in critical care research,” *J. Am. Med. Inform. Assoc. JAMIA*, vol. 25, no. 1, pp. 32–39, Jan. 2018, doi: 10.1093/jamia/ocx084.
- [10] D. M. Vock *et al.*, “Adapting machine learning techniques to censored time-to-event health record data: A general-purpose approach using inverse probability of censoring weighting,” *J. Biomed. Inform.*, vol. 61, pp. 119–131, 2016, doi: <https://doi.org/10.1016/j.jbi.2016.03.009>.
- [11] H. Ishwaran, U. B. Kogalur, E. H. Blackstone, and M. S. Lauer, “Random survival forests,” *Ann. Appl. Stat.*, vol. 2, no. 3, pp. 841–860, Sep. 2008, doi: 10.1214/08-AOAS169.
- [12] J. L. Katzman, U. Shaham, A. Cloninger, J. Bates, T. Jiang, and Y. Kluger, “DeepSurv: personalized treatment recommender system using a Cox proportional hazards deep neural network,” *BMC Med. Res. Methodol.*, vol. 18, no. 1, p. 24, Feb. 2018, doi: 10.1186/s12874-018-0482-1.
- [13] P. Chapfuwa *et al.*, “Adversarial Time-to-Event Modeling,” *Proc. Mach. Learn. Res.*, vol. 80, pp. 735–744, Jul. 2018.
- [14] J.-E. Bibault *et al.*, “Development and Validation of an Interpretable Artificial Intelligence Model to Predict 10-Year Prostate Cancer Mortality,” *Cancers*, vol. 13, no. 12, p. 3064, Jun. 2021, doi: 10.3390/cancers13123064.
- [15] T. Chen and C. Guestrin, “XGBoost: A Scalable Tree Boosting System,” *Proc. 22nd ACM SIGKDD Int. Conf. Knowl. Discov. Data Min.*, pp. 785–794, Aug. 2016, doi: 10.1145/2939672.2939785.
- [16] “Survival Analysis with Accelerated Failure Time — xgboost 1.5.2 documentation.” https://xgboost.readthedocs.io/en/stable/tutorials/aft_survival_analysis.html (accessed Apr. 10, 2022).
- [17] S. M. Lundberg, G. G. Erion, and S.-I. Lee, “Consistent Individualized Feature Attribution for Tree Ensembles,” Feb. 2018, Accessed: Aug. 19, 2021. [Online]. Available: <https://arxiv.org/abs/1802.03888v3>
- [18] S. M. Lundberg and S.-I. Lee, “A Unified Approach to Interpreting Model Predictions,” in *Advances in Neural Information Processing Systems*, 2017, vol. 30. Accessed: Aug. 19, 2021. [Online]. Available: <https://proceedings.neurips.cc/paper/2017/hash/8a20a8621978632d76c43dfd28b67767-Abstract.html>
- [19] F. E. Harrell, “Evaluating the Yield of Medical Tests | JAMA | JAMA Network.” <https://jamanetwork.com/journals/jama/article-abstract/372568> (accessed Mar. 06, 2021).
- [20] O. of P. and Planning, “Veteran Population - National Center for Veterans Analysis and Statistics.” https://www.va.gov/vetdata/veteran_population.asp (accessed Sep. 02, 2021).
- [21] F. C. Hamdy *et al.*, “10-Year Outcomes after Monitoring, Surgery, or Radiotherapy for Localized Prostate Cancer,” *N. Engl. J. Med.*, vol. 375, no. 15, pp. 1415–1424, Oct. 2016, doi: 10.1056/NEJMoa1606220.

Figure legend

Figure 1: Cohort selection

Figure 2. SHAP predictor importance (average impact on model output magnitude)

Figure 3a. Instance level predictor importance for hypothetical patient 1 with more aggressive disease at diagnosis- **Hypothetical patient 1 timeline**

Figure 3a. Instance level predictor importance for hypothetical patient 1 with more aggressive disease at diagnosis- **Outcome prediction**

Figure 3b. Instance level predictor importance for hypothetical patient 2 with less aggressive disease at diagnosis-**Hypothetical patient 2 timeline**

Figure 3b. Instance level predictor importance for hypothetical patient 2 with less aggressive disease at diagnosis **Outcome prediction**