Expiration of Machine Learning Predictive Models in healthcare

Aadi Ajmire[1,2], Ghazal Shabestanipour[1], Hamid Ghaednia[1], Joseph Schwab[1]
[1]Cedars Sinai, Los Angeles, CA, [2]University of California - Los Angeles, Los Angeles, CA

DISCLOSURES: Authors have nothing to disclose.

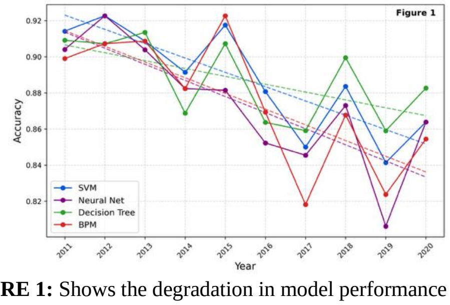
INTRODUCTION: In recent years, thousands of machine learning (ML) predictive models have been developed, validated, and used in the clinical setting. While these models are shown to be extremely useful and very accurate at the time of deployment, the expiration time for these models was never studied. Many clinical models continue to be used throughout the years without any regard for changing socioeconomic factors, healthcare practices, data recording standards, and better algorithms becoming available. These updated outcomes can impact the data that a healthcare model trains and predicts on, causing unreliable predictions and the potential for model decay. In this study, we investigated the decay of one of our groups widely used ML models that was developed for the prediction of mortality in patients with Chondrosarcoma (Thio et al., 2018). We chose Chondrosarcoma because the methods of treatment have not changed significantly. Therefore, one would assume that model prediction would not decay. To observe the impact of temporal change, we trained and experimented with hundreds of ML models to predict 1-year survival outcomes for patients with Chondrosarcoma. These experiments tested the predictive abilities of ML models subjected to time decay and feature selection. Through these experiments, we investigated the optimized models and showed that there are hidden patterns in the data that change with time, which necessitates updating ML models regularly.

METHODS: For this study, the team pulled 17 registries of data from two Surveillance, Epidemiology, and End Results (SEER) database, one from 2022 (old) and another from 2023 (new), for male and female patients with either conventional or dedifferentiated chondrosarcoma. The following variables were preprocessed: age, sex, year of diagnosis, primary site, radiation therapy, chemotherapy, surgical therapy, tumor size, total malignant tumors, total benign tumors, race, origin (Hispanic or non-Hispanic), county median household income, days from diagnosis to treatment, histology, tumor grade, tumor stage, and 1-year survival. These features were chosen based on their clinical relevance from our previous study of predictive Chondrosarcoma survival models. For this experiment, we utilized data from 2000 - 2020 for the old (n=3954) and new (n=3971) datasets. An exploratory data analysis (EDA) was conducted on both datasets. It was noted that the overall 1-year survival fluctuated between 79% (2019) and 94% (2012). Due to heavy class imbalance of our data (89% 1-year average survival) and changing number of year-to-year samples, down sampling and normalization was conducted on the training set to ensure even yearly sampling, and an even number of positive and negative samples per year. We experimented with calibrated (cv=10) and optuna-optimized gradient boosted decision trees (XGB), bayes point machines (BPM), support vector machines (SVM), and neural networks (NN) predicting 1-year survival, and were trained on the following features: age, sex, year of diagnosis, histology, tumor size, primary site, surgical therapy, tumor grade, and tumor stage. A total of 180 samples from 2000 - 2010 from each dataset were extracted to train all models, and models predicted on the full set of samples for each individual year from 2011 to 2020. Models were tested on their accuracy, brier score, specificity, and sensitivity. After observing better overall performance with the XGB architecture for both old and new dataset models, a study was conducted to prevent the decay observed over time in the XGB architecture. Features were individually added to the original feature (OF) set. Thereafter, models were trained and tested in a manner similar to the original study. This allowed for insight into the role of feature selection on model decay. A model with an expanded feature set was also experimented on, with the most clinically relevant features from the additional feature set, which was retrieved based on a Shapley additive explanations (SHAP) analysis of an XGB model trained on all features pulled from SEER. Regression lines were created based on the accuracy of year-to-year predictions on 1-year survival for each feature set, and the slope was used to assess model decay after feature selection.

RESULTS: Through the EDA of the original features, it was observed that the correlative powers of clinical features and survival shift throughout the years. When analyzing model performance, figure 1 demonstrated that the XGB architecture allowed for the most generalizable as well as strongest predictions throughout time out of the set of models used in the initial experiment. Figure 2A and figure 2B show degradation in the XGB architecture regardless of the dataset. However, when comparing model performance based on the dataset, the models trained on the new dataset performed better across all feature sets. Moreover, based on the slopes of the regression lines for yearly model accuracy, feature selection plays a role in model degradation. This is most prominent in the new dataset's XGB model comparing the model trained on the OF set and the OF plus multiple correlative features set, with degradation decreasing by 80% (from a slope of -0.005 to -0.001). Similarly, the old dataset's XGB models degradation decreased by 25% (from a slope of -0.008 to -0.006) when comparing models trained on the OF set and OF plus radiation set.

DISCUSSION: The EDA demonstrated that the relevance of features can differ over the span of a few years as lifestyle and healthcare factors change. Based on the heightened model performance of the XGB architecture, model selection is crucial to ensure strong performance over time. However, models can still observe slight degradation in their predictive abilities. This cements the need to retrain models with more relevant datasets as they become available to curb model expiration. From the feature selection study, the plots give insight regarding updates in data recording, and the need to train models on the most current datasets as data recording practices change. Furthermore, feature selection can play a role in the degradation of a model's predictive powers. Finally, based on the fluctuations in accuracy, it is shown that important features used to train ML models can change in their predictive strength over time.

SIGNIFICANCE/CLINICAL RELEVANCE: These experiments stress the need for regular updates and maintenance of ML models, even in cases which treatment strategies, and medications have not changed. In this study we present that there are hidden patterns in the data that necessitate updating ML models. Updates to data recording process, changes to feature importances such as trends is socioeconomical parameters could all alter the overall performance of survival prediction ML models. This decay of performance will negatively impact the clinical decision-making process; hence we emphasize on the need to not only train models on more recent data, but also to reconduct feature selection on a routine basis to ensure accessibility of the most accurate ML models for clinical decision-making.



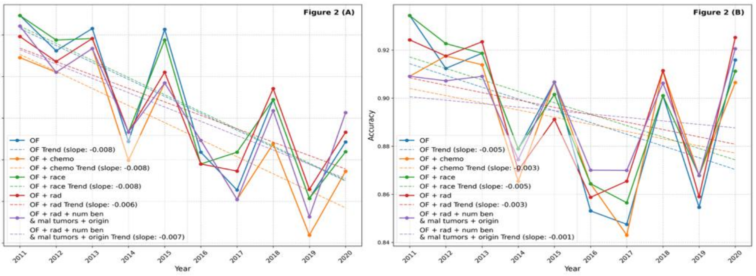


FIGURE 1: Shows the degradation in model performance in 4 different model architectures based on the old dataset, trained on data from 2000 to 2010 and tested yearly from 2011 to 2020. Regression lines were used to compare degradation of models based on model architecture.

FIGURE 2(A, B): Shows the degradation in model performance in XGB models trained on various features (OF or OF & selected features) from 2000 to 2010 and tested yearly from 2011 to 2020. Additionally, regression line slopes were noted to compare degradation based on selected features. Figure 2A is based on model yearly accuracy from the old dataset, while figure 2B is based on model yearly accuracy from the new dataset.

REFERENCES: Thio, Q. C. B. S., et. al. (2018). Can machine-learning techniques be used for 5-year survival prediction of patients with chondrosarcoma? *Clinical Orthopaedics and Related Research, 476*(10), 2040–2048. https://doi.org/10.1097/CORR.0000000000000433