Predicting Pitcher Injury: A survival analysis approach

Peiyuan Xu
Mentors: Eric Nalisnick, Anton Dahbura
JHU Sports Analytics Research Group
https://sports-analytics.cs.jhu.edu/

JOHNS HOPKINS
WHITING SCHOOL of ENGINEERING

Abstract

Baseball pitchers, at all ability levels, are injuring their arms at an alarming rate. Thus predicting if a pitcher is likely to become injured in the near future would be helpful to the pitcher's career and health. Traditional statistical and machine learning models have made strides in injury forecasting; yet they fail to account for real-world complexities, such as censored data, recurring injuries, and the dynamic nature of player workloads. In this work, we apply survival analysis to the problem, benchmarking traditional statistical approaches with modern recurrent neural networks (RNNs). We apply these models to Statcast data collected from professionals leagues, demonstrating the superiority of survival RNNs.

The Importance of Pitcher Injury Prevention

In baseball, pitchers are responsible for delivering the ball to the catcher without allowing the ball to be hit by the batter. Pitching is extremely strenuous on the arms due to the great torque exerted to throw the pitch at speed that can exceed 90 miles per hour. As a result, pitcher injuries, particularly those requiring Tommy John surgery to repair the ulnar collateral ligament, are increasing dramatically during past several decade at all levels.

Pitcher injuries affect player wellbeing and overall performance. In 2022, there are 30,728 total days lost among all pitchers due to injuries. At the league level, injuries bring significant costs and affect the team's investment in its players. Indeed, $486 million was spent on sidelined players in 2022. Therefore, it's extremely important to dedicate research efforts to the diagnosis and prevention of pitcher injury.

Objectives

The challenges in predicting pitcher injury include data in which the time to event is not observed, time-varying covariates, and injury recurrence, which makes traditional statistical and Machine Learning approaches unsuitable. Therefore, survival analysis methods are necessary.

Our work consists of the following objectives:

1. **Develop survival models to predict injury risk for different pitchers or pitcher groups based on pitch features, demographic characteristics, and injury recurrence**
2. **Investigate the pitch and demographic characteristics that are most predictive of pitcher injury.**

Survival Analysis Formulation

Models the average pitch characteristics for each pitch at game-level as well as the pitcher demographic information

Features:

- Time-invariant game-level analysis considers the average pitch characteristics across all games (average of pitches in game) until injury recurrence
- Time-varying game-level analysis considers the average pitch characteristics for each game until injury recurrence
- Demographic information such as height, weight, batting and throwing hand are recorded for each pitcher.
- Recurrence is encoded as the number of previous injuries for the pitcher in the follow-up season.

Event Time: the number of games until injury recurrence for a pitcher since the previous injury in the follow-up period.

Event Indicator: the pitcher experience the event if injury occurred before the end of the followup period. Otherwise, the instance does not experience the event and the actual time to event is unknown.

SurvivalRNN Model

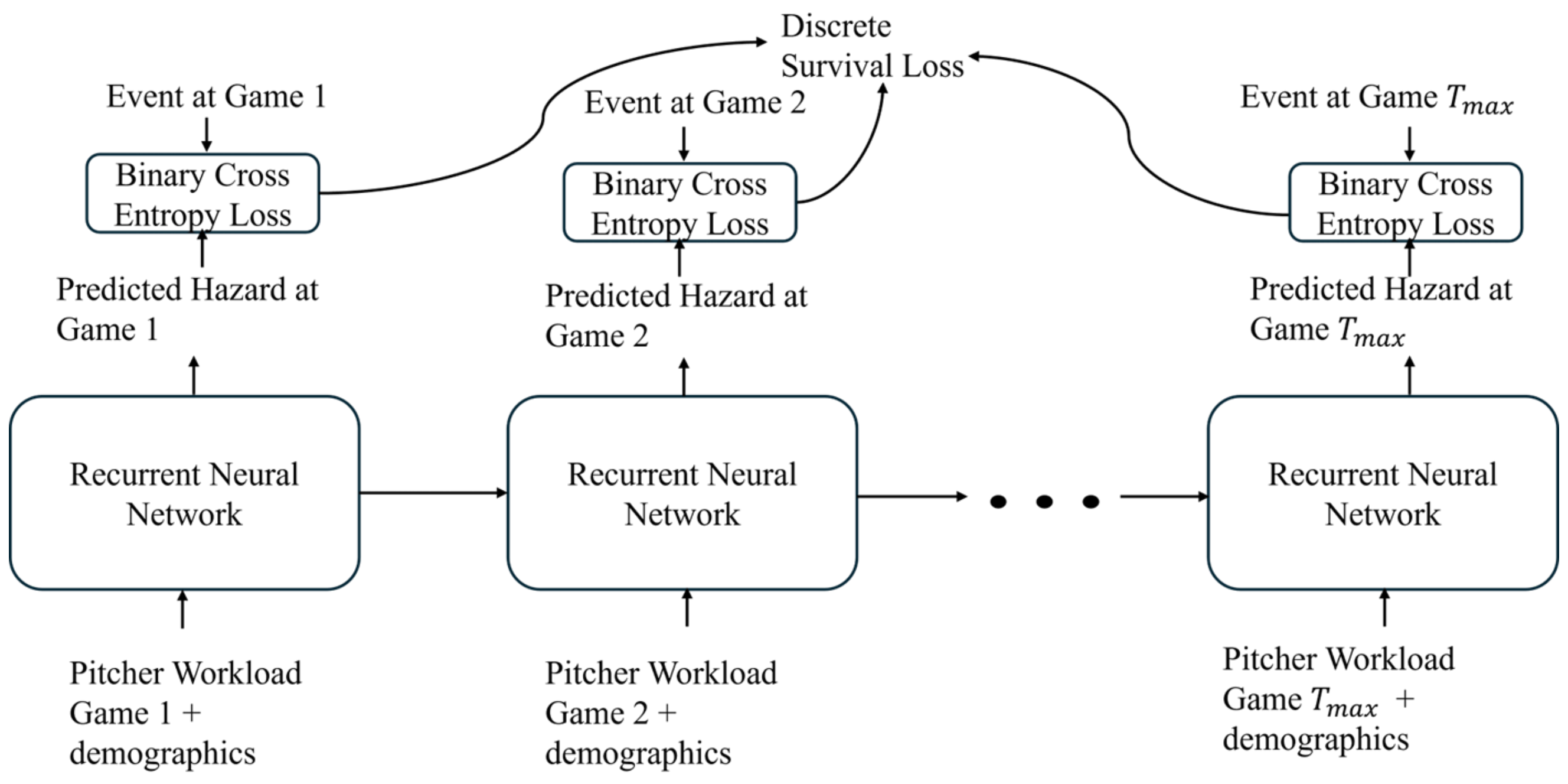
Model the time-varying effects of pitch features and demographic characteristics on injury risk through SurvivalRNN models with weighted discrete survival loss.

Survival RNN Formulation:

- The timesteps represent the games until injury recurrence.
- The time-varying workload features for each game include average release spin rate, release speed, effective speed, and velocity in the x, y, and z directions across all its pitches as well as the number of pitches.
- Each timestep also passes in information such as the pitcher demographics.
- The events represent whether the injury occurred after a certain number of games.
- The SurvivalRNN model is trained with the weighted discrete survival loss where the weights are tuned to optimize the predictive accuracy of the minority injured instances.

SurvivalRNN Evaluation:

- Integrated Brier Score: how well-calibrated the predicted survival probabilities are against the actual survival outcomes?
- Concordance Index: how well does the model predict the relative risk of different individuals with different survival times?


Discrete Survival Loss

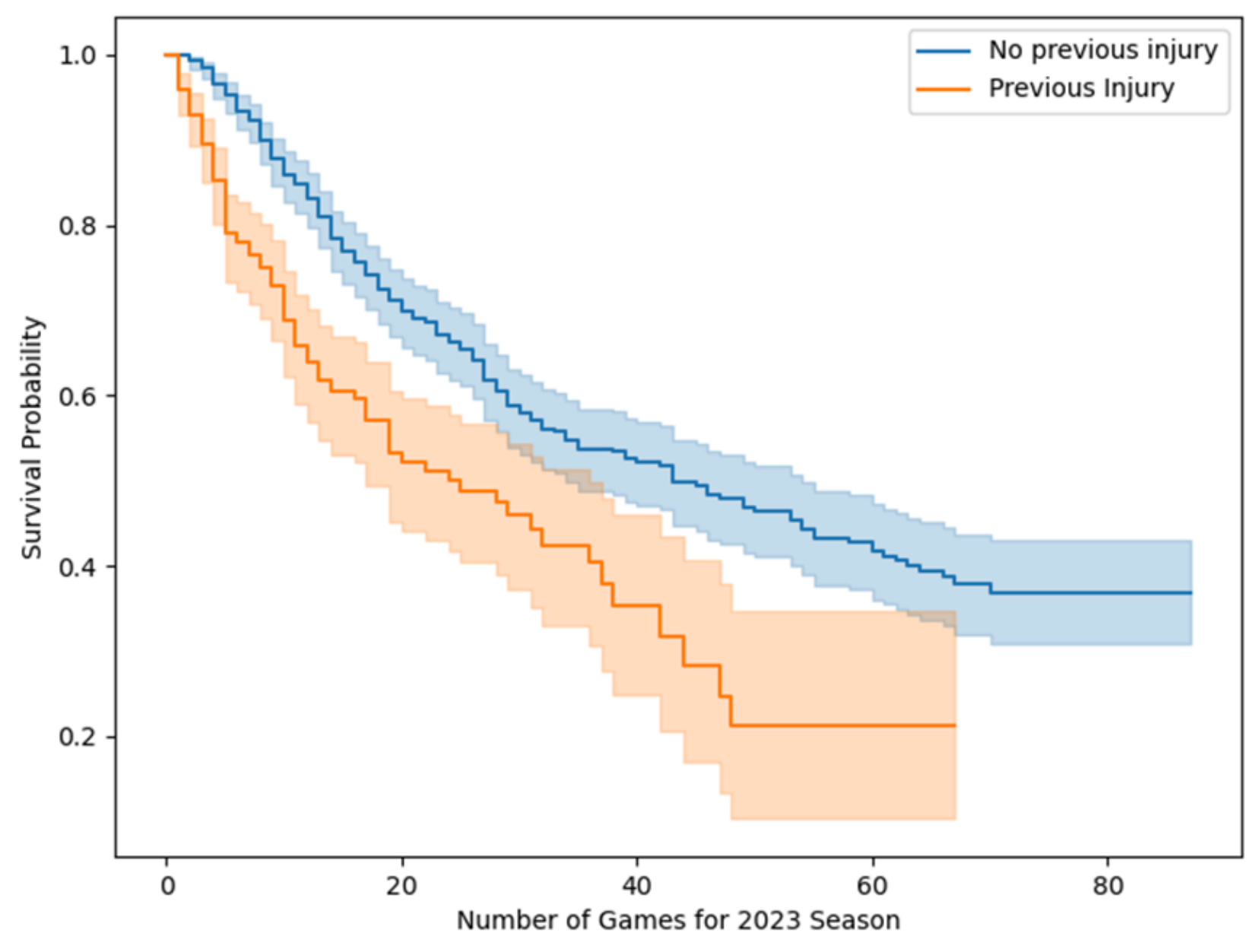
Event at Game 1
Binary Cross Entropy Loss
Predicted Hazard at Game 1
Recurrent Neural Network
Pitcher Workload Game 1 + demographics

Event at Game 2
Binary Cross Entropy Loss
Predicted Hazard at Game 2
Recurrent Neural Network
Pitcher Workload Game 2 + demographics

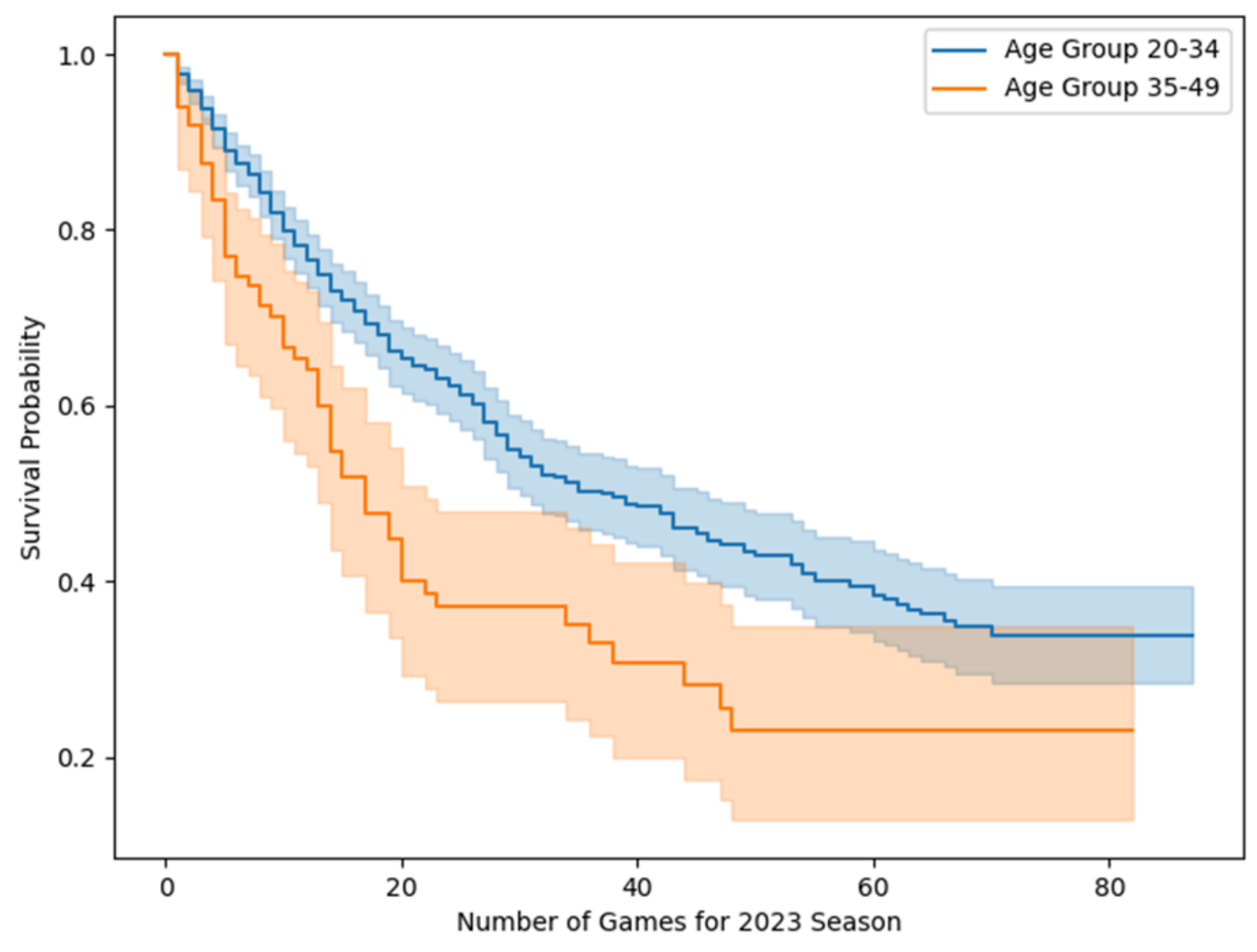
Event at Game T_{max}
Binary Cross Entropy Loss
Predicted Hazard at Game T_{max}
Recurrent Neural Network
Pitcher Workload Game T_{max} + demographics

Proposed SurvivalRNN Model

Empirical Results



Difference in Estimated Kaplan-Meier Curves for Previous Injury (p-value: < 0.005)



Difference in Estimated Kaplan-Meier Curves for Age (p-value: < 0.005)

Real-world datasets

- Prosports Transactions Baseball: injury dates, outcomes, and types for Major League Baseball pitchers
- Statcast Pitching Data: characteristics for each thrown pitch within a particular period
- Lahman Demographic Database: demographic information for Major League Baseball pitchers

Baseline Survival Models

- Kaplan-Meier curve: nonparametric survival model that estimates the survival fucntions.
- Cox Proportional Hazard Model: semi-parametric survival model with the proportional hazard assumption.
- Weibull Accelerated Failure Time Model: parametric model with accelerating or decelerating hazard.
- Gradient Boosting Trees: fits regression trees on the negative gradient of the Cox Proportional Hazards loss for each state starting from a base learner.
- Random Survival Forest: averages multiple fitted survival trees on various subsamples of the dataset.

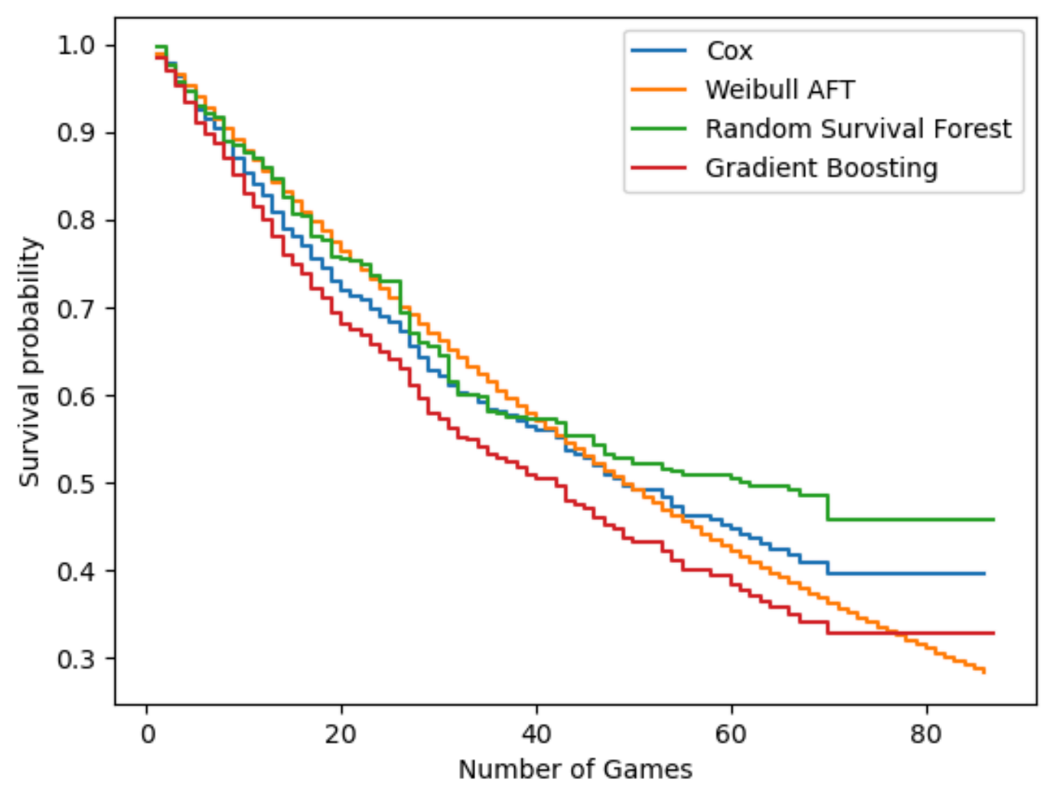
	Fitting/Evaluation Season		
Model	**2021/2022**	**2022/2023**	**2023/2024**
Cox PH	0.17	0.21	0.23
Weibull AFT	0.17	0.22	0.24
Gradient Boosting	0.18	0.22	0.25
Random Survival Forest	0.19	0.23	0.24
SurvivalRNN	0.39	0.53	0.53

Table 1: Integrated Brier Score (IBS) of Game-Level Survival Models

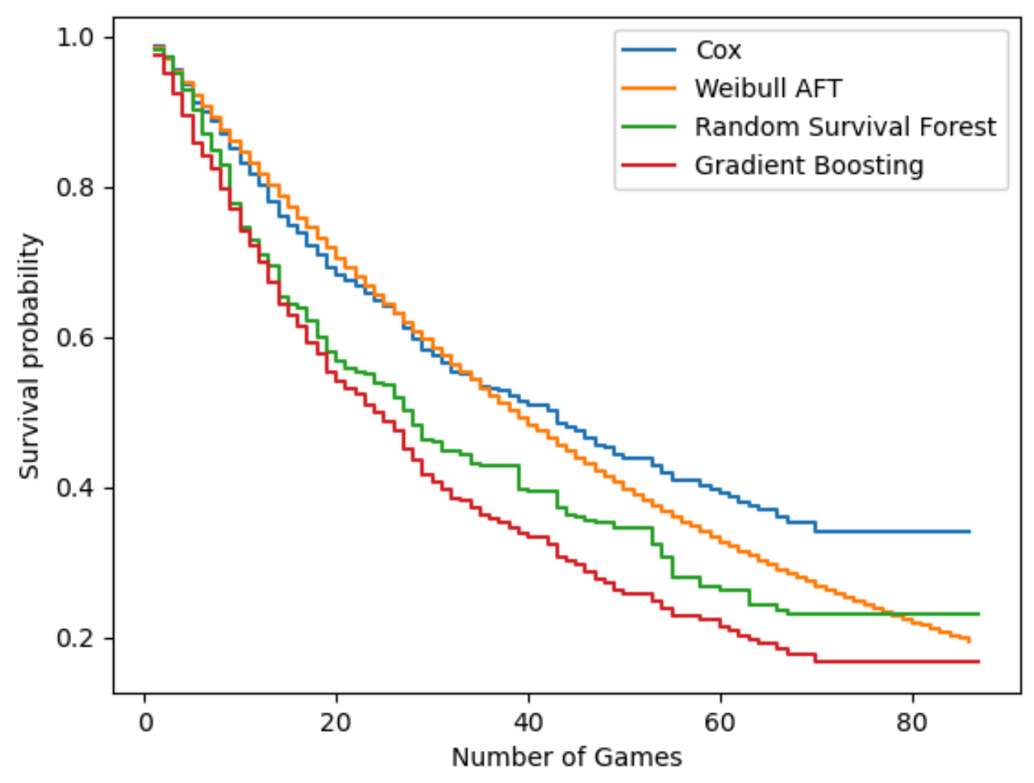
	Fitting/Evaluation Season		
Model	**2021/2022**	**2022/2023**	**2023/2024**
Cox PH	0.59	0.60	0.61
Weibull AFT	0.59	0.59	0.62
Gradient Boosting	0.58	0.57	0.57
Random Survival Forest	0.55	0.53	0.57
SurvivalRNN	0.81	0.97	0.85

Table 2: Concordance Index of Game-Level Survival Models

Predictions for 2025 Season Pitchers



Zack Wheeler



Framber Valdez

Conclusion

In conclusion, we proposed the survival models that predict pitcher injury risk over time and the features most indicative of pitcher injury risk. Experiments with professional Statcast data suggests that the SurvivalRNN model achieved superior performance with regards to baseline statistical and Machine Learning survival models in injury risk prediction.

For future work, we plan to further improve the predictive performance and training efficiency through developing novel Transformer models. We also plan to integrate these models into the Atlantic League Professional Baseball platform for use by coaches and league personnels to make data-driven injury prevention decisions.