Modeling Career Longevity for Position Players in Major League Baseball Using Survival Analysis

Undergraduate Statistics Research Project Competition (USRESP)

Abstract

This study develops a survival analysis framework for modeling career longevity and exit risk among Major League Baseball (MLB) position players using historical performance and injury data. Drawing on 63,956 player-season observations from the Lahman Baseball Database (1871–2024), supplemented by external injury and retirement records, we construct a time-to-event model centered around the Cox proportional hazards framework with time-varying covariates. Offensive performance metrics (lagged wOBA and plate appearances), aging effects via natural cubic splines, categorical injury severity, and missed-season indicators are evaluated across thirteen candidate model specifications using AIC, BIC, and the concordance index (C-index). Results indicate that playing opportunity, age, prior offensive production, injury accumulation, and season continuity all significantly influence career survival. Plate appearances emerge as the strongest predictor of career stability, while interrupted playing continuity substantially amplifies exit risk. Player-specific trajectory projections illustrate how the framework adapts across diverse career archetypes.

1. Introduction

Major League Baseball (MLB) careers are inherently uncertain. Few players are granted the opportunity to sustain long, productive tenures in such a competitive environment. Understanding the mechanisms that govern career duration has significant implications for player evaluation, contract structuring, and long term organizational planning.

Most existing approaches to career analysis rely on age-performance curves, which describe how metrics such as Wins Above Replacement (WAR) or On Base Plus Slugging (OPS) evolve over time. While informative, these curves primarily describe performance trajectories rather than addressing when and why careers actually end. This study reframes career analysis as a time-to-event problem, where the event of interest is a player's exit from MLB. Focusing exclusively on position players, whose career sustainability depends on a complex mix of offensive output, durability, positional flexibility, and organizational role, we construct a survival modeling framework capable of capturing these contributing influences.

Research Question: How do age, recent offensive performance, playing opportunity, injury history, and season continuity jointly influence the instantaneous risk of career exit for MLB position players?

2. Background and Significance

Career longevity in professional baseball has been studied primarily through the lens of aging curves, statistical models that describe how player performance evolves as a function of age. The delta method, developed and refined through work at Baseball Prospectus, estimates year over year performance changes to infer aging effects (Judge, 2020). Extensions of this approach, including joint modeling of aging curves using value based metrics such as WAR, have provided increasingly nuanced descriptions of how offensive and overall player contribution declines across career stages (FanGraphs Community Research, 2025). Additional research published through sabermetric outlets including FanGraphs, Monster Baseball, and independent analysts has documented that MLB position players typically peak offensively in their late 20s and experience accelerating decline thereafter (Monster Baseball, 2024; Hartzell, n.d.).

A consistent limitation of existing aging curve research is survivorship bias: players observed at older ages are disproportionately those who remained healthy and productive enough to retain roster spots, creating misleading impressions of how aging affects the broader population of players. Additionally, traditional aging curve models focus on describing the trajectory of performance rather than estimating the probability or timing of career exit itself. Nguyen (2022) demonstrated the value of longitudinal statistical approaches for studying MLB player trajectories, reinforcing the potential of time to event frameworks for career modeling.

Survival analysis, widely applied in medicine, engineering, and economics to model time to event outcomes, has seen limited application in professional sports career research. The Cox proportional hazards model (Cox, 1972), the most widely used semi-parametric survival model, offers particular advantages for career modeling: it accommodates censored observations (active players), incorporates time-varying predictors, and requires no assumptions about the underlying shape of career exit risk over time. This study applies and extends the Cox framework to MLB position player careers, integrating performance, workload, and injury data to produce season-by-season estimates of career exit risk.

Methodologically, this work demonstrates how survival analysis can be adapted to sports career data with multi-source longitudinal structure and heterogeneous historical records. Empirically, it produces the first systematic survival model for MLB position player careers integrating lagged performance metrics, injury severity classifications, and missed-season interaction effects across the full recorded history of the sport. Practically, the framework provides actionable tools for organizational decision-making, including season-by-season career projections for individual players.

3. Data

3.1 Sources

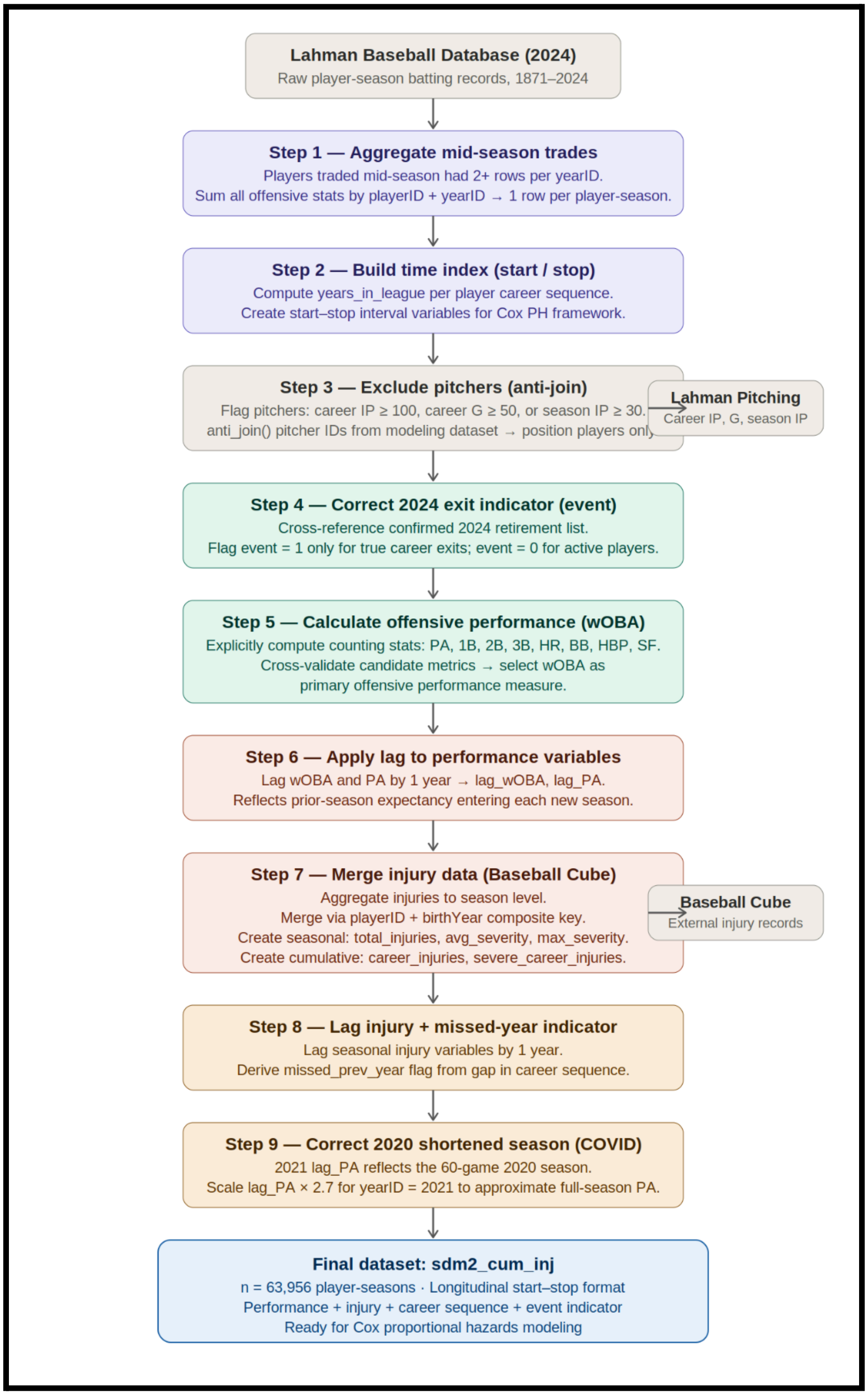
The primary dataset for this research is the Lahman Baseball Database (2024), a comprehensive relational database of season-level MLB statistics dating to 1871. Player demographic information from the People table was merged with batting statistics to construct the analytical panel. All pitchers were excluded via an anti-join against the Lahman Pitching table using thresholds on career innings pitched, games pitched, and maximum single-season innings pitched, preventing misclassification of position players who occasionally pitch in blowout affairs.

Injury records were obtained from Baseball Cube through sponsor access, providing documentation of major and minor league injury list placements with associated text descriptions. No shared primary key existed between datasets, a composite key matching player first name, last name, birth year, and season year was used to link injury records to the performance panel after standardizing name formatting to lowercase. A 2024 retirement list was incorporated to correct the career exit indicator for the final observation year.

3.2 Dataset Construction

The final analytical dataset contains 63,956 player-season observations spanning 1871–2024, structured in counting-process (start-stop) format. Displayed in Figure 1, the nine-step preprocessing pipeline included: (1) aggregating mid-season trade statistics to one row per player-season; (2) constructing the years_in_league time index and start-stop intervals; (3) excluding pitchers via anti-join; (4) correcting the 2024 exit indicator using the retirement list; (5) computing plate appearances and wOBA using 2013 linear weights; (6) applying one-year lags to wOBA and plate appearances; (7) merging injury data and constructing seasonal and cumulative injury variables; (8) lagging injury variables one season and deriving the missed_prev_year indicator; and (9) scaling 2021 lagged plate appearances by 2.7 to correct for the COVID-19 shortened 2020 season (60 games that were played being projected onto a full 162 game season).

Figure 1



Nine-step preprocessing pipeline transforming raw Lahman batting records (1871–2024) into a Cox proportional hazards-ready survival dataset.

Injury severity was classified via keyword-based regex detection applied to text injury notes, producing three ordinal categories (Table 1). Because return-date fields contained substantial inconsistencies across historical eras, particularly for injuries spanning offseason periods, continuous time-missed severity was not feasible; categorical classification using available text descriptions was adopted to ensure consistency.

Table 1

Severity	Score	Keyword Trigger Examples (from regex classification)	Example Injury Notes
Minor	1	sore, soreness, bruise, bruised, back issues, pain, stiffness, irritation, fatigue	"DL-15 Sore right wrist" / "DL-15 Right shoulder soreness" / "DL-15 Bruised right heel" / "DL-15 Back issues"
Moderate	2	strain, sprain, groin, hamstring, ankle, concussion, inflammation, tightness, back strain	"DL-15 Groin strain" / "DL-15 Strained left hamstring" / "DL-7 Lower back strain" / "DL-15 Sprained ankle *Originally injured during spring training, tried to play with injury*" / "DL-15 Concussion *Related to 2011 concussion*"
Severe	3	torn, tear, acl, achilles, surgery, fracture, rupture, ligament, tendon surgery	"DL-60 Torn ACL in left knee" / "DL-15 Torn left oblique" / "DL-60 Right shoulder surgery" / "DL-60 Back surgery *Recovery*" / "DL-15 Left achilles tendon surgery *Surgery*"

Injury severity classification scheme examples used in the thesis, mapping keyword-based regex triggers from injury notes to categorical severity scores (1–3).

3.3 Key Variables

The outcome variable (event = 1) identifies a player's final observed MLB season. Right-censored observations (event = 0) include active players in 2024 not confirmed as retired. The primary time-varying predictors are lagged plate appearances (lag_PA), lagged wOBA (lag_wOBA), age (updated each season), cumulative career injuries (career_injuries), lagged maximum prior-season injury severity (lag_max_severity, categorical: 0 = none, 1 = minor, 2 = moderate, 3 = severe), and a binary indicator for missing the prior season (missed_prev_year). Cumulative career plate appearances (career_PA) and lagged games played (lag_G) were also considered, but lag_G was excluded from the final model due to high collinearity with lag_PA, while career_PA was not found to be statistically significant after accounting for the remaining predictors already included in the model.

4. Methods

4.1 Cox Proportional Hazards Model

The Cox proportional hazards model expresses the hazard of career exit for player i at season t as:

$$h_i(t) = h_0(t) \cdot \exp(X_i(t) \cdot \beta)$$

where $h_0(t)$ is an unspecified baseline hazard, $X_i(t)$ is the vector of covariates at season t, and β is the coefficient vector. Positive coefficients indicate elevated exit risk for higher values of the predictor; negative coefficients indicate reduced risk in the same fashion (lower values of the predictor). The semi-parametric structure allows relative covariate effects to be estimated without assuming a parametric form for the baseline hazard. Coefficients are estimated via partial likelihood. The counting process structure contributes multiple rows per player, allowing time-varying covariates to be updated at each season interval.

4.2 Nonlinear Effects via Natural Cubic Splines

Age-related career risk is documented to be nonlinear: consistently increasing from ages 20-35, a rate of change that plateaus as careers stretch beyond that value. Similarly, the marginal protective effect of plate appearances and wOBA varies substantially across the observed range, with the largest risk reductions occurring at lower values where players transition from marginal to established roles. A linear form would misrepresent these relationships.

Natural cubic spline transformations with three degrees of freedom were applied to age, lag_wOBA, and lag_PA: ns(age, 3), ns(lag_wOBA, 3), ns(lag_PA, 3). Natural splines fit piecewise cubic polynomials joined at interior knots, constrained to linearity beyond the boundary knots to reduce extrapolation instability. The basis expansion replaces each continuous predictor with K basis functions, allowing the log-hazard to follow a flexible data-driven curve rather than a fixed slope.

4.3 Interaction Effects

Two interaction terms allow the relationship between offensive performance and exit risk to differ depending on whether a player was active in the prior season:

- missed_prev_year × ns(lag_wOBA, 3): the wOBA-to-hazard relationship is estimated separately for players who missed the prior season versus those with continuous participation.
- missed_prev_year × ns(lag_PA, 3): the plate-appearance-to-hazard relationship is similarly stratified by prior-season availability.

Without these terms, the model would assume that the protective effect of strong performance is identical regardless of prior-season absence, an assumption inconsistent with how organizations evaluate returning players, whose recent statistics may reflect a smaller, selected sample. When a player missed the prior season, lag_wOBA and lag_PA reference performance from two seasons prior rather than one, making the interaction terms essential for correctly contextualizing offensive metrics that originate from a non-adjacent season.

4.4 Model Selection

Thirteen candidate specifications were evaluated in order of increasing complexity, from a fully linear benchmark to specifications incorporating splines, categorical injury terms, cumulative injury burden, and interaction effects. Model selection balanced three criteria: AIC (penalizes complexity proportionally to parameter count), BIC (applies a stronger complexity penalty, particularly appropriate given n = 63,956), and the concordance index. The C-index measures predictive discrimination via the probability that the model correctly ranks two players by exit timing, originally introduced by Harrell et al. (1982). Cross-validation procedures were incorporated alongside the C-index to assess out-of-sample generalization. The final model was selected to minimize BIC while maintaining a C-index competitive with the best-performing specifications.

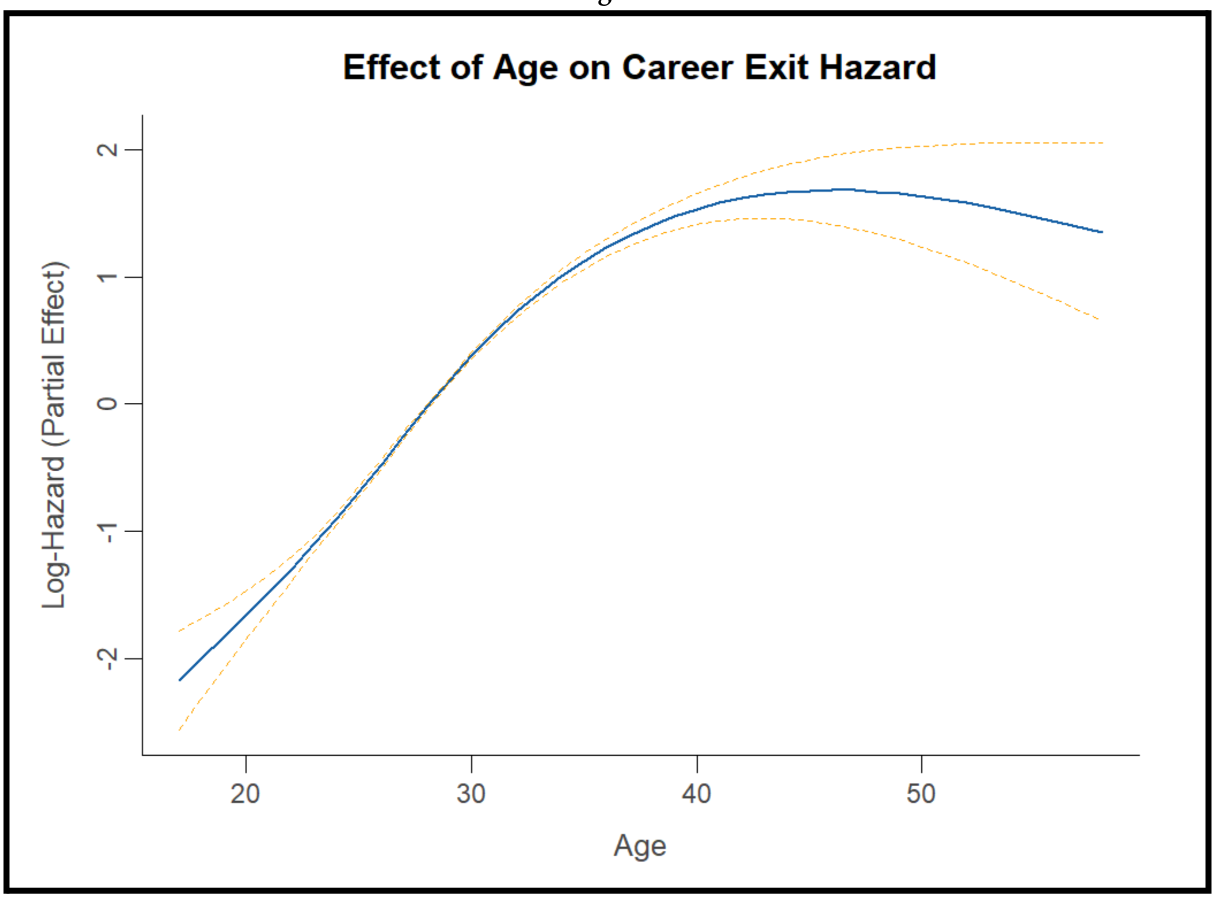
5. Results

5.1 Descriptive Overview

The final dataset spans 63,956 player-season observations across more than 150 years of MLB history. Career lengths range from one season to 26 seasons (mean ≈ 4.0 seasons; median = 3 seasons). Lagged plate appearances range from 0 to 778 (median = 288); lagged wOBA ranges from 0 to 2.1 with a median of 0.308. Approximately 6.8% of player-seasons follow a missed prior season. Career injury counts range from 0 to 16 (mean ≈ 0.67); about 69% of observations record no prior-season injury.

In order to accommodate potential nonlinear relationships between covariates and career exit risk, this work employs natural spline transformations within the Cox proportional hazards framework. Rather than modeling covariate effects as strictly linear, spline-based methods expand each continuous predictor into a set of basis functions that allow for flexible, data-driven curvature in the estimated effect. For example, Figure 2 displays the partial effect of age - capture via a natural cubic spline - on the log of the hazard function.

Figure 2



Partial effect of splined age on log-hazard of career exit (ages 17–58). Dashed lines show a 95% confidence band.

5.2 Model Comparison

Table 2 summarizes performance across thirteen representative model specifications. The career_count model produced the lowest BIC (121,604.2) and a competitive AIC (121,464.5), with C-index = 0.7677.

Nonlinear specifications consistently outperformed the linear benchmark across all three criteria, confirming that career exit risk is governed by nonlinear relationships. Models incorporating injury severity and interaction terms outperformed injury-free specifications. The injury_rate model achieved the highest C-index (0.7683) but at a BIC cost, pointing to a worse tradeoff between fit quality and complexity.

Table 2

Model	AIC	BIC	C-Index	Selected
Selected: Career Count	121464.5	121604.2	0.7677	Yes
Career Severity Score	121466.2	121605.8	0.7677	
Factored Max Severity + Interactions	121490	121622.7	0.768	
Injury Rate	121486.4	121626.1	0.7683	
Nonlinear + Missed Interactions	121521	121632.8	0.7671	
Factored Total Injuries + Interactions	121505.2	121651.9	0.7679	
Factored Avg Severity + Interactions	121502	121697.5	0.7681	
Factored Severity + Factored Missed	121715.1	121805.9	0.766	
Nonlinear + Missed Season	121748.7	121818.5	0.7651	
Nonlinear Base	121892.6	121955.4	0.7619	
Nonlinear + Career Injuries	121892.8	121962.6	0.7617	
Linear Benchmark	122182.6	122203.6	0.7584	
Linear + Career PA	122184.6	122212.5	0.7584	

Model comparison table summarizing Cox proportional hazards specifications evaluated using AIC, BIC, and C-index metrics, highlighting the trade-off between model fit, complexity, and predictive discrimination across linear, nonlinear, interaction, and injury formulations

5.3 Main Predictor Effects

Table 3 summarizes the direction, significance, and interpretation of each predictor in the selected model. The full coefficient output is presented below the table. All spline terms for age, lag_PA, and lag_wOBA, as well as the interaction terms and career_injuries, are statistically significant at $p < 0.001$. The overall model likelihood ratio test yields $\chi^2(20) = 6{,}618$, $p < 0.001$.

Table 3

Term	Coef.	exp(β)	SE	z	p-value	exp(−β)	95% CI Lower	95% CI Upper	Significance
Age (Natural Spline, df=3)									
ns(age, 3)$_1$	4.048	57.28	0.1299	31.172	< 0.001	0.01746	44.41	73.88	***
ns(age, 3)$_2$	5.175	176.8	0.5150	10.048	< 0.001	0.005655	64.44	485.3	***
ns(age, 3)$_3$	3.120	22.65	0.3857	8.089	< 0.001	0.04415	10.63	48.24	***
Previous-Year Injury Severity (ref: 0)									
Lag Max Severity = 1	−0.081	0.922	0.1116	−0.728	0.467	1.085	0.741	1.148	
Lag Max Severity = 2	−0.207	0.813	0.0550	−3.760	< 0.001	1.230	0.730	0.906	***
Lag Max Severity = 3	−0.487	0.614	0.0694	−7.017	< 0.001	1.628	0.536	0.704	***
Career Injury Burden									
Career Injuries	0.040	1.041	0.00747	5.355	< 0.001	0.9608	1.026	1.056	***
Missed Previous Year									
Missed Previous Year	0.052	1.054	0.0844	0.618	0.537	0.9492	0.893	1.243	
Previous-Year wOBA (Natural Spline, df=3)									
ns(lag wOBA, 3)$_1$	−2.574	0.0762	0.1935	−13.305	< 0.001	13.12	0.0522	0.1114	***
ns(lag wOBA, 3)$_2$	−0.008	0.992	0.2950	−0.027	0.979	1.008	0.557	1.769	
ns(lag wOBA, 3)$_3$	2.642	14.05	0.5466	4.834	< 0.001	0.07119	4.812	41.01	***
Previous-Year PA (Natural Spline, df=3)									
ns(lag PA, 3)$_1$	−0.741	0.477	0.2179	−3.401	< 0.001	2.098	0.311	0.731	***
ns(lag PA, 3)$_2$	−13.15	1.95×10^{-6}	0.9670	−13.596	< 0.001	5.13×10^{5}	2.93×10^{-7}	1.30×10^{-5}	***
ns(lag PA, 3)$_3$	−21.29	5.70×10^{-10}	1.8270	−11.653	< 0.001	1.76×10^{9}	1.59×10^{-11}	2.04×10^{-8}	***
Interaction: Missed Year × wOBA Spline									
Missed Prev Year × ns(lag wOBA, 3)$_1$	2.572	13.09	0.4304	5.975	< 0.001	0.07642	5.629	30.42	***
Missed Prev Year × ns(lag wOBA, 3)$_2$	−1.024	0.359	0.9306	−1.100	0.271	2.783	0.0580	2.226	
Missed Prev Year × ns(lag wOBA, 3)$_3$	−3.509	0.0299	1.9490	−1.801	0.072	33.42	0.000657	1.364	.
Interaction: Missed Year × PA Spline									
Missed Prev Year × ns(lag PA, 3)$_1$	−0.221	0.802	0.4152	−0.532	0.595	1.247	0.355	1.809	
Missed Prev Year × ns(lag PA, 3)$_2$	12.01	1.65×10^{5}	1.3810	8.695	< 0.001	6.08×10^{-6}	1.10×10^{4}	2.47×10^{6}	***

Cox Proportional Hazards Model Results: Career Survival of MLB Position Players, Note: Spline basis terms (subscripts 1–3) represent internal knot contrasts and do not have direct interpretations; effects should be visualized jointly. *Significance codes: *** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$, . $p < 0.10$. Model fit — Likelihood ratio test: $\chi^2(20) = 6618$, $p < 0.001$*

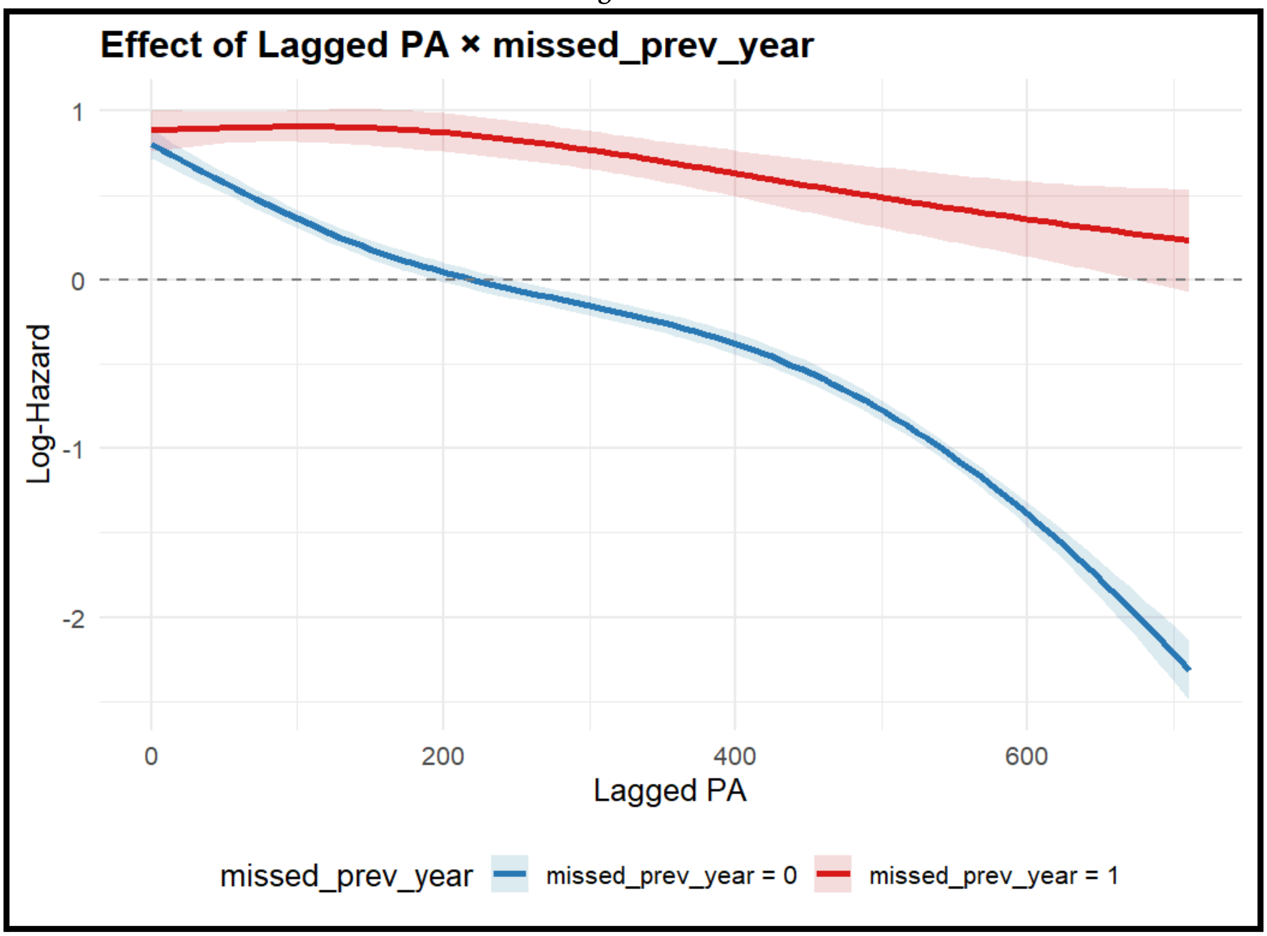
5.4 Age Effects

All three age spline terms are highly significant ($p < 0.001$). The partial effect plot (Figure 1) confirms a strongly nonlinear pattern: exit risk consistently increases from ages 20-35, and then its effect begins to plateau as a significant contributor beyond said interval. This is consistent with established baseball aging literature and confirms that a linear age term would substantially understate exit risk acceleration among player career spans.

5.5 Playing Opportunity (Lagged Plate Appearances)

Prior-season plate appearances are among the strongest predictors of career continuation in the model, producing hazard reductions comparable in magnitude to the age spline terms. The effect is highly nonlinear: players transitioning from limited to moderate roles (roughly 0–300 PA) experience the sharpest reductions in exit risk, while additional plate appearances beyond an established-starter threshold yield progressively smaller incremental benefits. This finding suggests that organizational trust, reflected in consistent lineup utilization, is nearly as influential as age in determining short-term career survival.

Figure 3



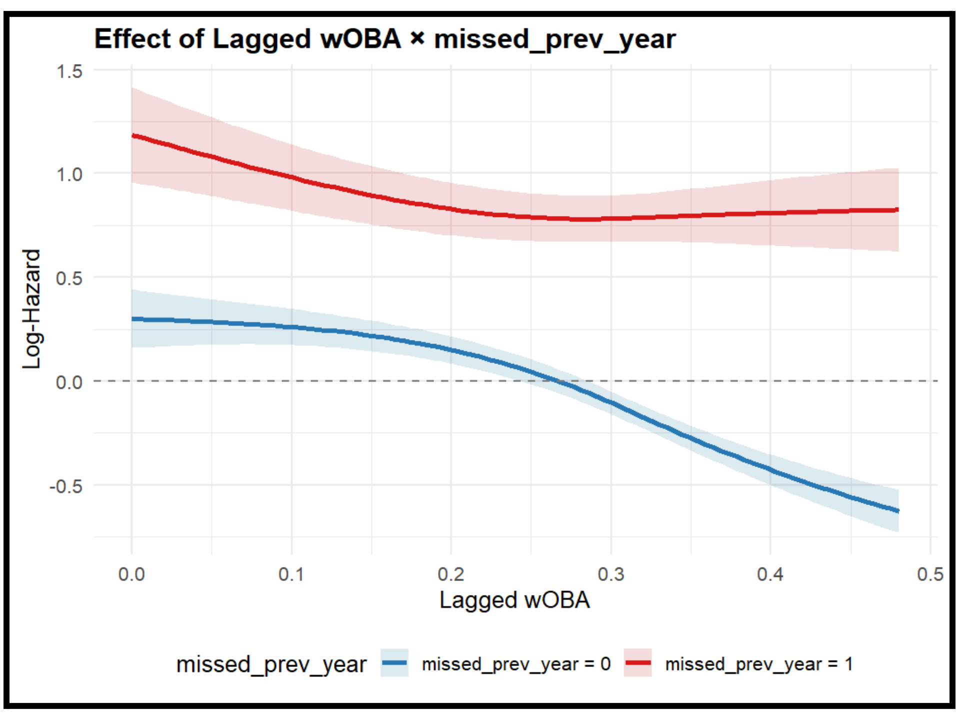
Nonlinear (splined) interaction between prior-season plate appearances (lag_PA) and missed-year status on log-hazard of career exit. Blue = active prior season; Red = missed prior season. Players returning from absence require substantially more plate appearances to reach equivalent hazard reductions.

For players with an active prior season, the plate-appearance-to-hazard relationship is relatively steady across the 0 to 400 plate appearance range, with exit risk declining at a moderate and consistent pace before accelerating more sharply beyond 400 plate appearances, where the protective benefit of established high-volume playing time becomes most pronounced. For players returning from a missed season, the curve behaves markedly differently: there is almost no distinguishable reduction in exit risk between zero and approximately 200 prior plate appearances, reflecting the fact that these statistics originate from two seasons prior rather than one, and therefore carry considerably less signal about a player's current organizational standing. Beyond 200 plate appearances, risk does begin to decline steadily, but the hazard remains substantially elevated relative to continuously active players with equivalent plate appearance totals. Both of these patterns are intuitively consistent with the structure of the interaction term. The missed season curve is higher overall because absent players face genuinely elevated exit risk regardless of their prior offensive output, as organizational continuity and roster security are themselves disrupted by the absence. The curve is also flatter because performance statistics from two seasons ago are inherently less informative about a player's current role and value than statistics from the immediately preceding season, naturally dampening the protective signal that plate appearances would otherwise provide.

5.6 Offensive Performance (Lagged wOBA)

Lagged wOBA displays a nonlinear, threshold-driven relationship with career hazard. Below a wOBA of approximately 0.20, increases in offensive quality produce only modest reductions in exit risk, suggesting that performance below this threshold does little to meaningfully alter a player's career outlook. Beyond this point, corresponding roughly to the transition from below-replacement to league-average offensive production, there is a considerably more drastic and sustained decrease in exit risk as wOBA increases from 0.20 toward 0.50, reflecting the substantial career security that consistent above-average offensive output provides to an established player.

Figure 4



Nonlinear (splined) interaction between prior-season wOBA (lag_wOBA) and missed-year status on log-hazard of career exit. Blue = active prior season; Red = missed prior season. The protective effect of wOBA is substantially muted for players returning from absence.

For players returning from a missed season, the wOBA effect is substantially muted. The hazard reduction curve flattens above wOBA ≈ 0.20, and then presents a more drastic decrease in hazard as wOBA ventures between 0.20 and 0.50. This suggests that organizational evaluation weights recent performance in the context of availability: strong wOBA following an absence does not provide the same career-security signal as equivalent performance from an uninterrupted contributor.

5.7 Injury Effects

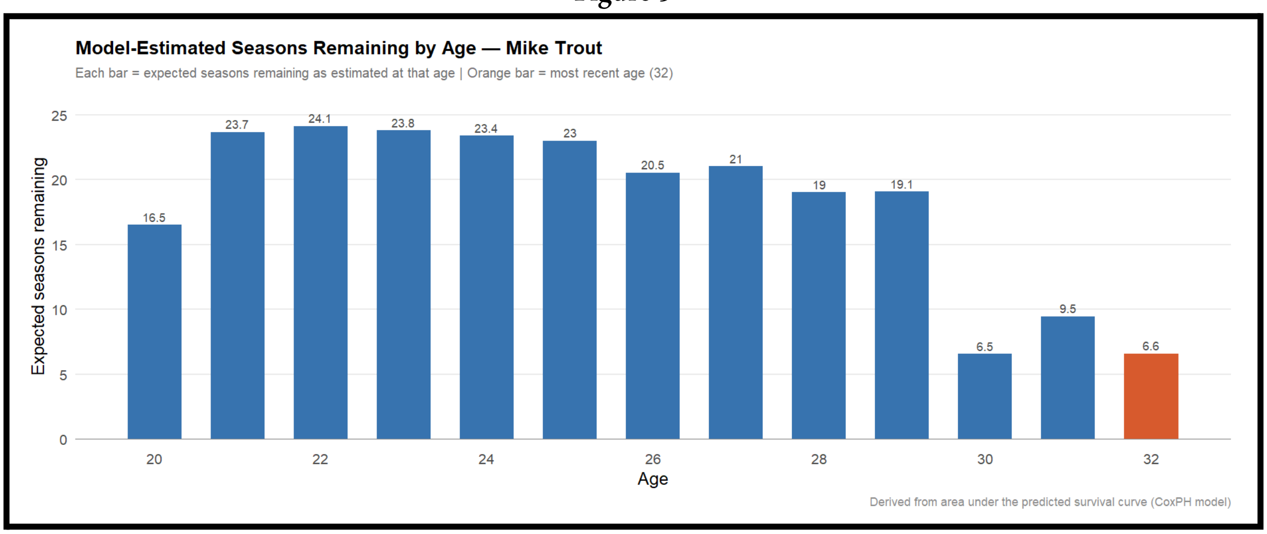
Prior-season maximum injury severity (lag_max_severity) produces hazard ratios of 0.813 for moderate injuries and 0.614 for severe injuries, suggesting lower hazard in the season following a recorded injury compared to injury-free seasons. This pattern reflects survivor selection: players who sustain severe injuries and successfully return are disproportionately those with high roster value, strong organizational investment, and sufficient performance to warrant reinstatement. Players injured at the margins of roster viability, who do not return, do not contribute post-injury observations.

Cumulative career injuries (career_injuries) behave as expected: each additional career injury increases exit risk by approximately 4.1% (HR = 1.041, $p < 0.001$). While modest per event, this risk compounds significantly across careers with extensive injury histories, 10 career injuries correspond to approximately a 50% increase in exit hazard ($1.041^{10} \approx 1.50$).

6. Predicted Career Trajectories

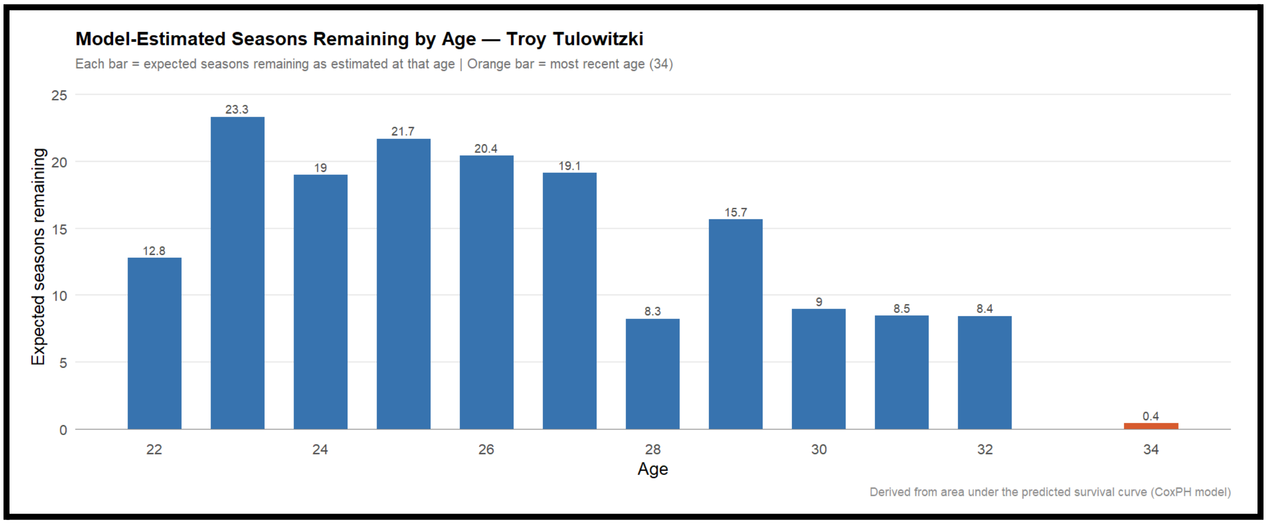
To illustrate framework flexibility, season-by-season career trajectory predictions were generated for four players representing distinct career archetypes: an elite performer (Mike Trout), an injury-disrupted star (Troy Tulowitzki), an average-production long-career player (Eric Young Jr.), and a short-duration career (Willians Astudillo). For each player, the model produces an expected seasons-remaining estimate at each observed age, derived from the area under the predicted survival curve conditional on that season's covariate values.

Figure 5



Model-estimated seasons remaining by age for career archetypes: Mike Trout (elite). Each point represents the model estimate from that season's covariate values.

Figure 6



Model-estimated seasons remaining by age for four career archetypes: Troy Tulowitzki (injury-disrupted). Each point represents the model estimate from that season's covariate values.

Figure 7

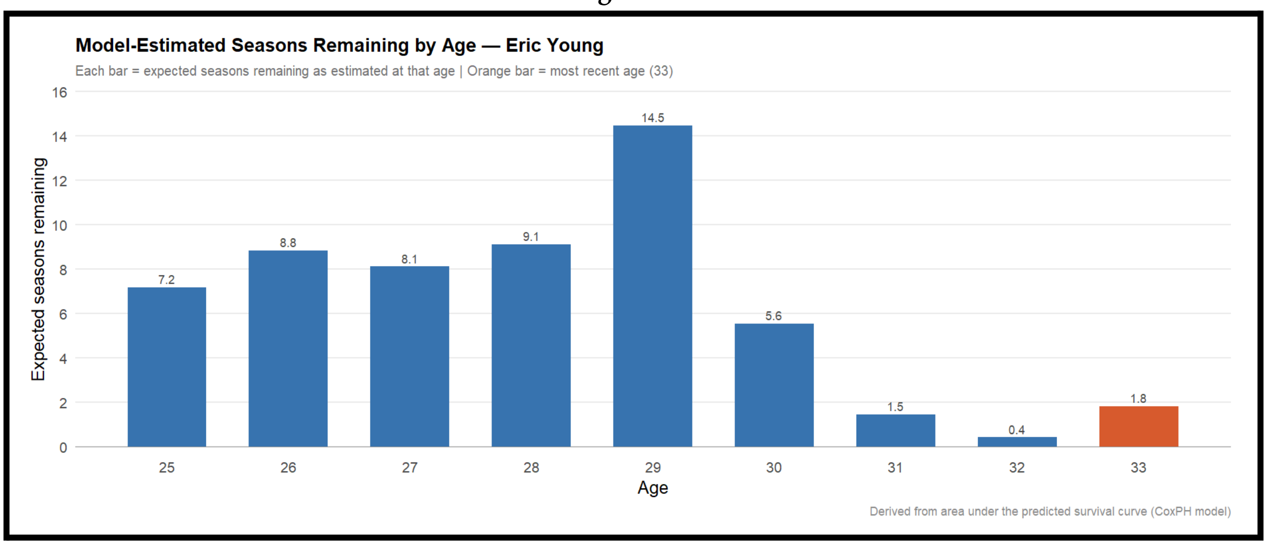
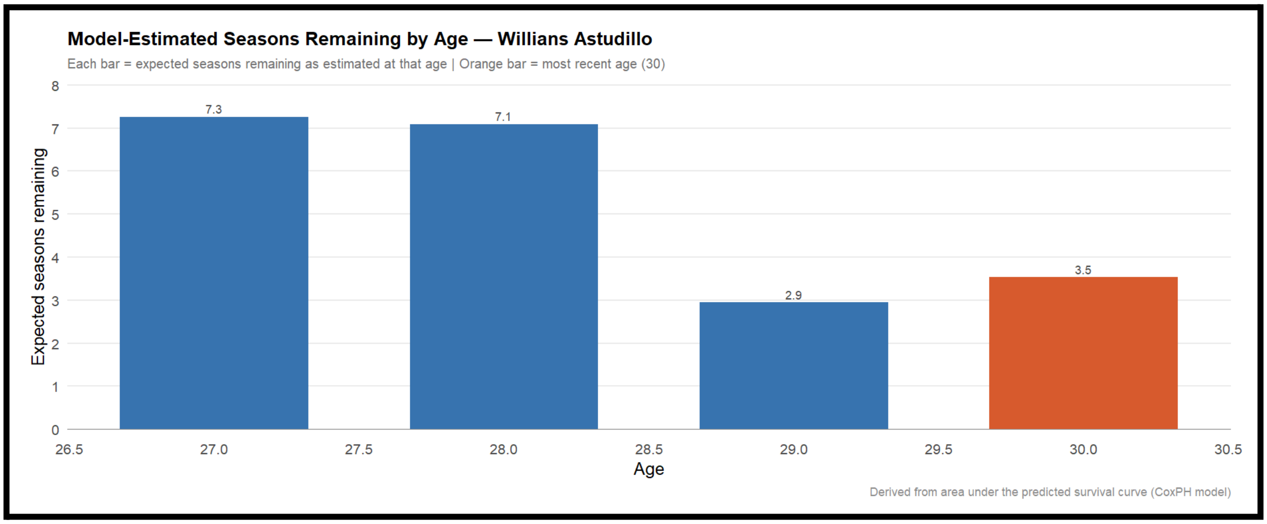


Figure 4. Model-estimated seasons remaining by age for four career archetypes: Eric Young Jr. (average). Each point represents the model estimate from that season's covariate values.

Figure 8



Model-estimated seasons remaining by age for four career archetypes: Willians Astudillo (short career). Each point represents the model estimate from that season's covariate values.

6.1 Elite Performance: Mike Trout

Trout's trajectory demonstrates the model's sensitivity to elite performance and injury-related disruption. Following an initial prediction of approximately 15.5 seasons remaining in his debut year (age 20), projections rise to 23–24 seasons remaining through his mid-20s as he establishes himself as one of the most productive offensive players in recent history. The model's most instructive response occurs at age 30: a prior-season shortened to 36 games and 146 plate appearances due to a right calf strain triggers a 12.6-season reduction in projected seasons remaining. The subsequent season, in which Trout posted a 0.421 wOBA across 499 plate appearances, restores the projection to 9.5 seasons. This trajectory illustrates that the model is appropriately responsive to disrupted playing continuity in otherwise elite career profiles.

6.2 Injury-Disrupted Star: Troy Tulowitzki

Tulowitzki's predictions follow a markedly volatile pattern consistent with a career in which 5 of 13 seasons were shortened to fewer than 100 games by injury. Years-remaining predictions decline sharply in injury-shortened seasons and recover partially in full-performance seasons, but the cumulative trajectory reflects 10 career injuries (3 severe). Following a missed 2018 season due to bilateral heel surgery, his final observed season (age 34) yields a projection of 0.4 seasons remaining, accurately preceding his career exit and demonstrating the model's ability to capture compounding injury-driven erosion of career longevity.

6.3 Average Performance: Eric Young Jr.

Young's 10-season career without a missed year illustrates a conventional hazard progression. Predictions track a gradual increase through his late-20s career year (598 PA, wOBA = 0.289; 14.5 projected seasons remaining at age 29) followed by a smooth decline reflecting age-driven hazard acceleration. The absence of significant injury disruption produces a clean trajectory that closely follows the global career-duration benchmark, providing an interpretable baseline against which more complex career patterns can be compared.

6.4 Short Career: Willians Astudillo

Astudillo's trajectory exemplifies elevated early-career hazard driven by limited playing time and inconsistent performance. Across five seasons, he never exceeded 73 games or 216 plate appearances. Despite a strong debut wOBA of 0.385 in 97 plate appearances, declining performance (final season wOBA = 0.245) combined with limited organizational utilization produced consistently high exit risk throughout. Projected exit age never exceeded 35.1, and final-season predictions converged accurately with his observed exit at age 30. This case illustrates that a strong debut alone is insufficient to ensure career stability; sustained opportunity and consistent performance are required to reduce exit risk over time.

7. Discussion

7.1 Summary of Findings

This study demonstrates that MLB career longevity is shaped by the combination of aging, offensive production, playing opportunity, and injury accumulation. Age governs the underlying trajectory of exit risk through an accelerating nonlinear pattern. Playing opportunity, measured by prior-season plate appearances, functions as the strongest observable predictor of career continuation, capturing organizational trust and roster stability in a single metric. Offensive quality provides meaningful additional signal, particularly above the league-average level performance thresholds, but its protective effect is substantially diminished following a missed season. Cumulative injury burden compounds over careers, gradually eroding the advantages of strong performance or established roles.

The interaction between missed-season status and both performance metrics is among the most substantively important findings. Organizational evaluation of returning players appears to differ systematically from evaluation of continuously active players: post-absence performance carries less protective weight, and substantially higher plate appearance totals are required to restore baseline survival prospects. This has direct implications for how teams should interpret performance metrics in the context of injury history and playing continuity.

7.2 Modeling Assumptions

The proportional hazards assumption, that covariate effects on hazard are constant over time, was assessed through Schoenfeld residual diagnostics as well as by introducing additional interaction structures between age and the lagged performance variables. Specifically lagged plate appearances and lagged wOBA, to evaluate whether the effects of offensive performance changed systematically across different ages, which would suggest a potential violation of the assumption over time. The Schoenfeld residuals showed no substantial evidence of time-varying coefficient behavior, and the age-interacted specifications similarly indicated that the proportional hazards assumption remained reasonably satisfied, suggesting that the estimated effects of the included predictors were sufficiently stable across the observed age ranges; consequently, the final modeling framework retained the original interaction structure involving missed previous season indicators rather than the age-interacted alternative, avoiding unnecessary increases in model complexity. The use of lagged predictors rather than concurrent-season values avoids data leakage and reflects the information available to organizations at the start of each season, which is the practically relevant decision point.

7.3 Limitations

Several limitations constrain interpretation. First, the event variable captures MLB exit rather than formal retirement; players departing for international or minor league play are treated identically to those who cease playing entirely, introducing heterogeneity in the event definition, though this remains consistent with the organizational question of interest, when players cease to contribute at the major-league level regardless of reason. Second, injury data quality varies substantially across historical eras, with pre-1990 records considerably less complete and consistently formatted; the keyword-based severity classification reflects available documentation rather than objective medical assessment and may not fully capture long-term functional consequences of individual injuries. Third, the 154-year historical scope introduces era-related heterogeneity in league structure, season length, medical practices, and offensive environments, and survivorship bias, particularly the underrepresentation of injury-driven exits in early eras, limits the precision of era-spanning comparisons, such that estimated effects represent aggregate relationships across fundamentally different baseball contexts. Fourth, defensive value was omitted due to the limited availability and inconsistency of comprehensive defensive metrics across the full historical sample, meaning the plate appearances proxy for organizational value is imperfect.

7.4 Future Research

Several extensions could enhance the framework. Era-stratified modeling would allow investigation of whether career survival dynamics differ across major periods of baseball history, improving applicability to contemporary organizational contexts. Competing risks modeling could formally distinguish between performance-driven, injury-driven, and voluntary exits. Refinements to the injury framework, including injury location, acute versus chronic classification, and quantitative time-missed measures corrected for historical inconsistencies, would improve the precision of injury burden estimates. Integration of defensive value metrics (e.g., Outs Above Average for modern seasons) and Statcast batted-ball quality measures would further enrich the feature set and reduce the reliance on plate appearances as a proxy for overall player value.

8. Conclusion

This study presents a comprehensive survival analysis framework for modeling career longevity among MLB position players, integrating age, offensive performance, playing opportunity, and injury history into a flexible time-to-event structure. The selected Cox proportional hazards model, incorporating natural cubic splines, categorical injury severity, cumulative injury burden, and interaction terms between prior-season absence and offensive metrics, achieves strong predictive discrimination (C-index = 0.7677) while maintaining interpretability and parsimony as indicated by leading AIC and BIC values across thirteen candidate specifications.

The central finding is that career survival in MLB is not solely a function of offensive talent. Consistent playing opportunity is the strongest observable predictor of career continuation, reflecting the organizational confidence and roster stability that sustain careers. Cumulative injury burden compounds gradually but meaningfully, and prior-season absence fundamentally alters the protective relationship between performance and hazard reduction. Player-specific trajectory case studies demonstrate that the model responds appropriately across diverse career archetypes and produces projections that align closely with observed outcomes, contributing both a methodological framework and substantive empirical findings with direct applications to player evaluation, contract decision-making, and workload management.

References

Baseball Almanac. (2024). 2024 MLB retirements. https://www.baseball-almanac.com

Baseball Savant. (2025). Statcast data. https://baseballsavant.com

Baseball-Reference. (2025). Historical player statistics. https://www.baseball-reference.com

Cox, D. R. (1972). Regression models and life-tables. Journal of the Royal Statistical Society: Series B, 34(2), 187–220.

FanGraphs. (2025). wOBA library. https://library.fangraphs.com/offense/woba/

FanGraphs Community Research. (2025). Joint model of the WAR aging curve. https://community.fangraphs.com

Fox Sports. (2021). Angels star Mike Trout out 6 to 8 weeks due to calf strain. https://www.foxsports.com

Hartzell, M. (n.d.). An update to MLB aging curves. Medium. https://medium.com/@matthartzell

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2021). An introduction to statistical learning: With applications in R (2nd ed.). Springer.

Judge, J. (2020). The delta method, revisited: Rethinking aging curves. Baseball Prospectus. https://www.baseballprospectus.com

Lahman Baseball Database. (2024). Version 2024. https://www.seanlahman.com/baseball-archive/statistics/

Major League Baseball. (2025). Standard stats glossary. https://www.mlb.com/glossary/standard-stats

Monster Baseball. (2024). Aging curves in professional baseball. https://monster-baseball.com

Nguyen, Q. (2022). Estimating aging curves: Using multiple imputation to examine career trajectories of MLB offensive players. Carnegie Mellon University Sports Analytics Conference.

Sports Illustrated. (2018). Blue Jays' Troy Tulowitzki suffers bone spurs in both heels. https://www.si.com

The Baseball Cube. (2025). Injured list history pro. https://www.thebaseballcube.com

Therneau, T. M. (2024). A package for survival analysis in R (Version 3.x). https://cran.r-project.org/package=survival

Harrell, F. E., Califf, R. M., Pryor, D. B., Lee, K. L., & Rosati, R. A. (1982). Evaluating the yield of medical tests. *JAMA*, 247(18), 2543–2546.

Generative AI Use Statement

Generative AI tools were used in specific capacities during the preparation of this research and submission. The original thesis on which this submission is based acknowledged the use of OpenAI's ChatGPT for portions of coding workflow assistance and editorial wording support. This competition submission was condensed and restructured from the original thesis with the assistance of Claude (Anthropic), which was used to help reformat and condense the written content into the competition submission format. All statistical analyses, model development, data collection, variable construction, and interpretation of results were conducted by the author. All research design decisions, substantive conclusions, and intellectual contributions are the author's own. Generative AI tools were not used to generate or fabricate any data, statistical results, or analytical outputs presented in this report.