Noname manuscript No.
(will be inserted by the editor)

What does not get observed can be used to make age curves stronger: estimating player age curves using regression and imputation

Michael Schuckers · Michael Lopez · Brian Macdonald

Received: date / Accepted: date

Abstract The impact of player age on performance has received attention across sport. Most research has focused on the performance of players at each age, ignoring the reality that age likewise influences which players receive opportunities to perform. Our manuscript makes two contributions. First, we highlight how selection bias is linked to both (i) which players receive opportunity to perform in sport, and (ii) at which ages we observe these players perform. This approach is used to generate underlying distributions of how players move in and out of sport organizations. Second, motivated by methods for missing data, we propose novel estimation methods of age curves by using both observed and unobserved (imputed) data. We use simulations to compare several comparative approaches for estimating aging curves. Imputation-based methods, as well as models that account for individual player skill, tend to generate lower RMSE and age curve shapes that better match the truth. We implement our approach using data from the National Hockey League.

Keywords Age curves · Generalized Additive Models · NBA · NHL · Sport · Simulation · Imputation

1 Introduction

How athletes perform as they age is a question that undermines models of player valuation and prediction across sport. The impact of age is typically measured using age curves, which reflect the expected average performance at each age among all players that participate.

This material is based upon work supported by the U.S. National Science Foundation under Grant No. CNS-1919554

Michael Schuckers
St. Lawrence University, Canton, NY
E-mail: schuckers@stlawu.edu

Michael Lopez
National Football League, Skidmore College
E-mail: Michael.Lopez@nfl.com

Brian Macdonald
Yale University
E-mail: brian.macdonald@yale.edu

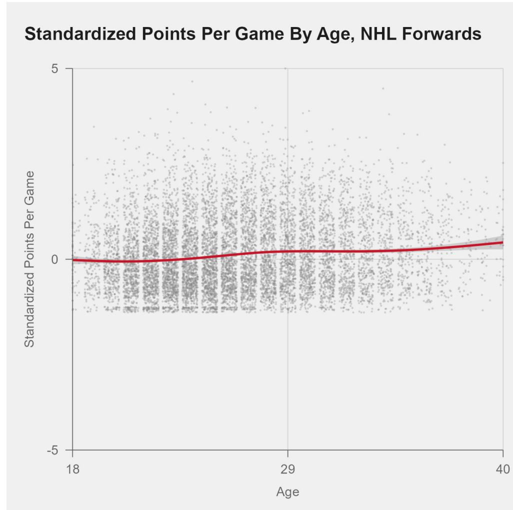

Fig. 1 Standardized points per game by age for NHL forwards between the 1995-96 and 2018-19 seasons, along with a cubic spline model fit to the data. The spline model does not decrease as it should for older ages because of selection bias: the only players that are observed at older ages are very good players.

Because players receive different opportunities at different ages, a selection bias exists, one that is linked to both entry into a league (more talented players typically start younger) and drop out (less talented players stop earlier). More talented players are more likely to be observed at both ends of the age curve, which makes it challenging to extrapolate the impact of age. For example, Figure 1 shows standardized points per game for National Hockey League (NHL) forwards, along with a spline regression model, one that uses age as the only predictor, fit to that observed data. There seems to be a slight bump at 28 or 29, but then the curve starts *increasing*, which is the opposite of what we would expect from an age curve. Because average and below average players tend drop out, the only players competing in their late 30s tend to be very good players.

In Figure 2, we highlight observed data for four example players who were in the league at least until their late 30s (left) and three players who were in the NHL through their late 20s and early 30s (right). These data for individual players initially exhibit the same general trend − increasing through their mid-to-late 20s. Unlike Figure 1, performance for these players decreases in their 30s, which is closer to what we'd expect in an age curve.

Along with selection bias, a second challenge when estimating age curves is lower sample sizes for higher ages. For some methods, a lack of data for mid-to-late 30s can lead to curves with higher variance for those age ranges. In Figure 3, we show 100 age curves derived using the Delta method (left, [10]) and a cubic spline using only observed data and no player effects (right) for estimating the age curves for 100 bootstrapped samples of the NHL forwards data. Both methods are described further in Section 2. Variance is low for some ages (e.g. mid 20s) but variance increases for older ages as fewer observed data are available.

Our goal in this paper is to assess the performance of various methods for age curve estimation under the selection bias of player entry and issues of small samples at younger and older ages. We propose several methods for player age curve estimation, introduce a missing data framework, and compare these new methods to more familiar approaches including both parametric and semi-

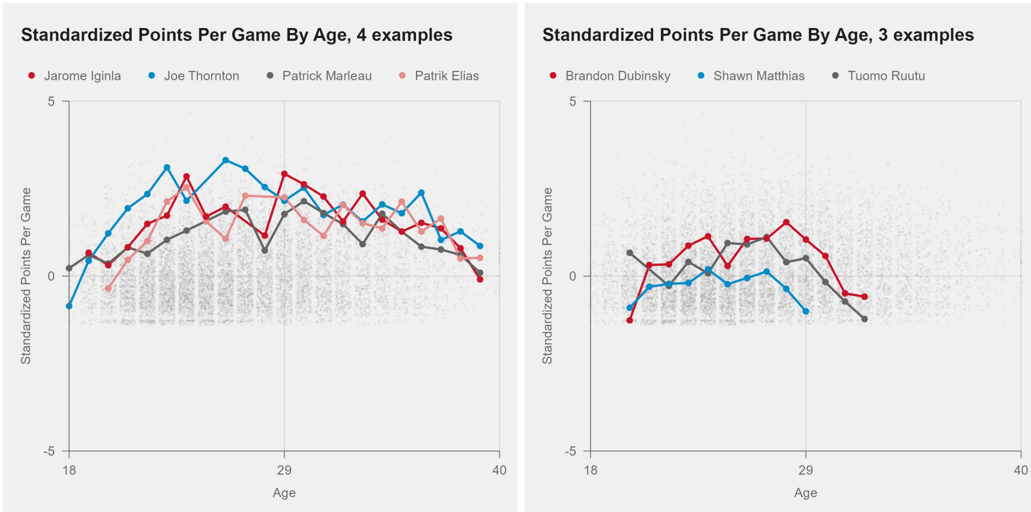

Fig. 2 Standardized points per game for 4 examples of very good forwards who played in the NHL until at least there late 30s (left), and 3 examples of good NHL forwards did not stay in the NHL through their late 30s. Individual player statistics over time tend to follow an inverted U-shaped trend, increasing from the mid to late 20s, and decreasing thereafter. For the players on the right, once their production fell, they exited the league, and have no observed data in their late 30s that would factor into the curve generated in Figure 1.

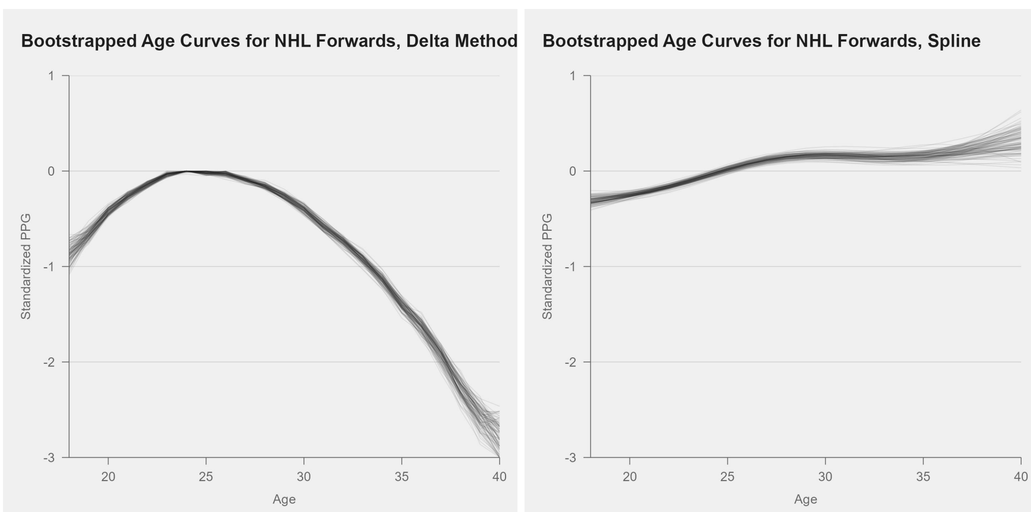

Fig. 3 100 bootstrapped age curves using the Delta Method(left) and Natural Spline regression without player effects (right) with the NHL Forwards data. Variation in the age curves increases for older ages where there are fewer observed data.

parametric modeling. Using simulations derived from NHL data age patterns, we compare the accuracy of these methods with respect to estimating a true, known age curve. We conclude by exploring these methods using actual NHL data.

2 Methodologies

For consistent terminology, we begin with the following notation. Let Y_{it} to represent the performance value of player i at age t. We assume discrete observations at each age (one per year), though it is possible to treat age in a more granular way.

A basic model is:

$$Y_{it} = g(t) + f(i,t) + \epsilon_{it} \tag{1}$$

where $g(t)$ is the average performance at age t for all players, $f(i,t)$ represents a possible performance adjustment at age t for player i, and ϵ_{it} is the model error at age t for player i.

2.1 Related literature

Broadly, current approaches can be split based on assumptions for estimating $g(t)$ (parametric, semi-parametric, or non-parametric) and whether or not to model age specific curves, $f(i,t)$. Table 1 summarizes how several authors have modeled age effects. Columns for author, sport or league, sample of players, and method. Articles in Table 1 are arranged by a rough categorization of approach for modeling the age term and then alphabetically by author.

Table 1 Summary of work on age curves

Paper	Sport/League	Unobserved?	Model
Schulz et al[13]	MLB	No	Average
Lichtman[10]	MLB	No	Fixed effects (Delta Method)
Tulsky[14]	NHL	No	Fixed effects (Delta Method)
Albert[1]	MLB	No	Fixed effects (Quadratic)
Bradbury[3]	MLB	No	Fixed effects (Quadratic)
Fair[5]	MLB, others	No	Fixed effects (Quadratic)
Villareal et al[17]	Triathlon	No	Fixed effects
Brander et al[4]	NHL	Yes	Fixed effects (Quadratic, Cubic)
Tutoro[15]	NHL	No	Semiparametric
Judge[7]	MLB	Yes	Semiparametric
Wakim and Jin[18]	MLB, NBA	No	Semiparametric
Vaci et al[16]	NBA	No	Fixed, random effects
Lailvaux et al[9]	NBA	No	Random effects
Berry et al[2]	MLB, NHL, Golf	No	Random effects
Kovalchik and Stefani[8]	Olympic	No	Random effects

A most naive approach assumes $f(i,t) = 0$ and $\hat{g}(t) = \sum_{i=1}^{n} Y_{it}/n$ for all t, as in [13], such that the average of observed Y_{it} is sufficient for estimating $g(t)$. Such an approach would only be valid if players were chosen to participate in sport completely at random, making it too unrealistic for professional sport. As in Figure 1, we can see that this approach yields results that are not credible.

Player i	Age t	Season	Observable (ψ_{it})	Observed (ϕ_{it})	Note
Jagr	38	2010-11	1	0	Played overseas in KHL
Jagr	39	2011-12	1	1	Returned to NHL
Crosby	40	2027-28	0	0	Future season, not yet observable

Table 2 Examples of player-ages that are observable but not observed, observable and observed, and not observable.

A common parametric assumption ([1], [4], [3], [5], [17]) is that Y_{it} is quadratic in age. If $g(t)$ is quadratic, performance "peaks" at some age, with players improving until this peak and eventually declining after the peak. Quadratic age curves are also symmetric across the peak performance age. [4] and [17] also consider a cubic age effect, while [1] uses a Bayesian model and weighted least squares, with weights proportional to player opportunity. [5] do not assume that quadratic form of $f(i,t)$ is symmetric. The Delta Method ([10]) is a modified fixed effects model using different subsamples at each consecutive pairs of ages, is expanded upon in Section 2.4.1.

A second suite of approaches assumes a semiparametric approach for estimating age curves, including spline regression techniques ([18]) and the broader family of generalized additive models ([7], [15]). Either approach is more flexible in their ability to pick up on non-linear patterns in age effects. A final assumption models age effects via individual age curves $f(i,t)$, as in [2] and [9].

In all but two examples above, authors use observed data only to make inferences on the impact of age. Conditioning on players being observed at each age potentially undersells the impact of age; only players deemed good enough to play will record observations. The two exceptions to this assumption are [4], who extrapolate model fits to players not observed in the data, and [7], who uses a truncated normal distribution to estimate errors. Importantly, [7] confirms that if dropout rate is linked to performance, that effectively shifts the age effect downward, relative to surviving players.

2.2 Notation

Our focus in this paper will be on the estimation of $g(t)$, the average aging curve for all players. Because not all of the players will be observed in a given year of their career (due to injury, lack of talent, etc), we create ψ_{it}, an indicator for if Y_{it} was observed for player i at a given age t, where $t = t_1, \ldots, t_K$. That is,

$$\psi_{it} = \begin{cases} 1 \text{ if } Y_{it} \text{ is observed, and} \\ 0 \text{ otherwise.} \end{cases} \tag{2}$$

we let t_1 be the youngest age considered and t_K will be the oldest age considered.

Similarly we will have players who we would like to include in our analysis but their careers are not yet complete. To that end we let ϕ_{it} is an indicator of if $Y(t)_{it}$ is observable. For our application to NHL data ϕ_{it} is an indicator that the performance could have happened by October 2020, when the data for this project was collected.

$$\phi_{it} = \begin{cases} 1 \text{ if } Y_{it} \text{ is observable, and} \\ 0 \text{ otherwise.} \end{cases} \tag{3}$$

As an example, NHL star Sidney Crosby's age 40 data is $\phi_{it} = 0$ because that is not yet observable as of this writing. See Table 2 for more examples of ψ_{it} and ϕ_{it}.

For the methods below, we also add the following definitions and notation. The data for a given player will be represented by the vector of values $\mathbf{Y}_i = (Y_{it_1}, \ldots, Y_{it_K})^T$ and the observed subset of those values will be denoted by the vector of values $\mathbf{Y}_i^{obs} = (Y_{it} \mid \phi_{it} = 1)^T$. Let $\mathbf{Y}^{obs} = ((Y_1^{obs})^T, (Y_2^{obs})^T, \ldots, (Y_N^{obs})^T)^T$ be the vector of all observed values, while $\mathbf{Y} = (Y_1^T, Y_2^T, \ldots, Y_N^T)^T$.

2.3 Estimation Methods

In this section we describe several current and novel approaches to estimation of the mean aging curve, $g(t)$. We begin by describing the de facto standard methodology in the sport analytics literature, the Delta method, [10], and an extension that we call Delta-plus. Below we outline some facets of our proposed approaches to the problem of estimation of $g(t)$. Roughly these methods breakdown into the general approach to estimation of $g(t)$, the data to be used for model fitting and the additional fixed or random effects terms. To help the reader we develop a notational shorthand for combining these facets that is *method:data:effects*.

2.4 Mean aging curve estimation

For this paper we consider four approaches to estimation of the mean player aging curve. The first is the non-parametric Delta method and a simple extension of this approach which we call *delta-plus*. The Delta Method which been discussed by [6], [15] and [10], is commonly used in practice. The second approach that we consider is a natural spline regression approach which we denote by *spline* and the third is a quadratic model *quad*. Finally, we propose a novel *quantile* methodology that utilizes information about the ratio of observed players to observable players at a given age.

2.4.1 Delta Method

The Delta Method is an approach to estimation that focuses on the maxima of $g(t)$ over t. The basic idea is to estimate the year of year change in the average player response curve by averaging among only the players that we observed in both years. A benefit of this approach is that it implicitly adjusts for a player effect by only using those players who have appeared in both years t_k and $t_k + 1$. Further, by simply calculating year over year averages, this approach is hyper-localized. However, the requirement that Y_{t_k} and Y_{t_k+1} are both observed (i.e. that $\psi_{it_k} = \psi_{it_k+1} = 1$) can be limiting in sports where there is significant drop-in/drop-out across years or seasons.

To develop the delta method, let

$$\delta_{t_k} = \overline{Y}_{i*_k, t_k+1} - \overline{Y}_{i*_k, t_k} \tag{4}$$

where

$$\overline{Y}_{i*_k, t_k} = \frac{1}{n*_{t_k}} \sum_i Y_{i \in i*_k, t_k},$$

$i*_k = \{i \mid \psi_{it_k} = \psi_{it_k+1} = 1\}$, and $n*_{t_k} = |i*_k|$ is the number of elements in $i*_k$. Thus, $i*_k$ is the set of players whose performance value was observed at age t_k and $t_k + 1$. Traditionally, the delta method is standardardized

$$\hat{g}(t) = \delta_t - \max_k \delta_{t_k} \tag{5}$$

for $t = t_1, \ldots, t_K$ so that the largest value of $\hat{g}(t)$ is zero. To date the Delta Method has proved effective at estimation of the maxima of $g(t)$. See [15] and [6] for details on some evaluation of the performance of the Delta method. This performance relative to other methods that use only observed data is likely due to this focus on year over year differences, which are susceptible to small sample size issues for older ages that can cause non-smooth age curves. Older ages are not (as big of) an issue for regression techniques (e.g. spline regression), which use information from nearby ages and result in a smooth age curve even when sample sizes are small.

2.4.2 Delta Plus Method

One drawback of the Delta Method as described above is that the maximal value for $\hat{g}(t)$ is forced to take the value zero. In part, this has been the case because the focus of estimation was on the age of maximal performance not on the estimation of the full aging curve; though the application of the Delta Method has taken on the latter function. Using the notation from above we define the estimation of μ_t as the following:

$$\hat{g}(t) = \delta_t - \max_k \delta_{t_k} + \max_k \overline{Y}^{obs}_{\cdot t_k}, \tag{6}$$

where $\overline{Y}^{obs}_{\cdot t_k} = \frac{1}{n_{t_k}} \sum_{i:\psi_{it_k}=1} Y^{obs}_{it_k}$ is the average of the n_{t_k} observed values at age t_k. Below we will refer to this method as *delta-plus*. We only use the Delta Plus Method in our evaluation below since it offers a more flexible approach than the Delta Method.

2.4.3 Spline approach

For this spline approach (spline), we utilize flexible natural basis spline regression and apply them to the Y_{it}'s, the player performance data. Thus, our spline regression use age, t, as a predictor and the performance metric Y_{ik} as the response. In particular, we use the `s()` option with 6 degrees of freedom in the `mgcv` package from `R`. In our shorthand for these methods we call this approach *spline* and use $s(t)$ to denote a spline model for the mean aging curve.

2.4.4 Quadratic approach

As the name implies, this methodology assumes that the mean aging curve is quadratic in terms of a player's age. In our shorthand we will use 'quad' and the model will be written as $g(t) = \gamma_0 + \gamma_1 t + \gamma_2 t^2$. Some authors including [5], [1] and [4] have assumed that $g(t)$ has this particular functional form. So we include this approach for comparison.

2.4.5 Quantile approach

In this method we aim to estimate $g(t)$ by utilizing the distribution and number of the observed values at each age t. We observe at each age t some fraction of the players which is reasonably assumed to be a truncated sample from a larger population. Let n_t be the number of players whose performance is observed at age t from the population of N_t players whose performance was observable at any age. If we know the fraction of players relative to the larger population (and the form of the distribution) then we can map percentiles in our sample to percentiles in the larger population.

For example, suppose we had a corpus of $N_{32} = 1000$ players and that at age 32 we observed $n_{32} = 400$ of them. The 75th percentile of observed metrics among the age 32 players might be reasonably used as an estimate for the 90th percentile of the population of 1000. Explicitly this is because the 100^{th} best Y_{it} among the combined observed can be assumed to be the 100^{th} best Y_{it} among the $n_{32} = 400$ observed and the $N = 1000$ unobserved $Y_{it}'s$. Similarly the 25th percentile of the above example would be the 70th percentile in the population. Below we assume a Normal distribution and use the additivity of the Normal distribution to obtain estimates for $g(t)$ at each age t. More generally, we can calculate the $q \cdot 100^{th}$ percentile from among the n_t observed values. Call that value ν_t. The value of ν_t is approximately the $G_t = \left(1 - \frac{n_t}{N_t}(1-q)\right) \cdot 100^{th}$ percentile from the population of values at age t. From our example above the $0.90 = 1 - \frac{400}{1000}(1 - 0.75)$.

From G_t we can estimate the mean of the full population, $g(t)$ at age t via: $\hat{\zeta}_t = \nu_t - \Phi^{-1}(G_t)\hat{\sigma}_t$ where $\Phi(\ )$ is the cumulative density of a standard Normal distribution. From $\hat{\zeta}_t$ we can estimate other percentiles as long as we can assume Normality and we have a reasonable estimate of the standard deviation, $\hat{\sigma}_t$. Our justification for assuming Normality notes that most performance metrics are averages or other linear combinations of in-game measurements and that the Central Limit Theorem applies to these linear combinations.

For estimation of σ_t, the standard deviation of the population at age t, we can estimate the standard deviation at age t, denote that by s_t, and adjust based upon the proportion of truncation. Note that the variance of a truncated Normal is always less than the variance of an untruncated one with similar underlying variance. So then we use $\hat{\sigma}_t = \frac{s_t}{\theta_t}$ where θ_t is the ratio of truncation from a standard Normal for these data at age t. To obtain θ_t we use the `vtruncnorm` function in the `truncnorm` library in R [12]. For our methodology shorthand we will use *quant* for this approach and denote the quantile based aging curve approach as $\zeta(t)$.

2.5 Data Imputation Methods

Another modeling choice to be made when working with missing data is whether or not to impute the missing values. We consider three possible options for imputation of responses, Y_{it}'s. The first option is simply to use the observed data and only the observed data. When we do this, we will use *obs* as part of our shorthand. The second and third options involve imputation of the unobserved but observable data, that is the set of values $\{Y_{it} \mid \psi_{it} = 0, \phi_{it} = 1\}$. In the second option *trunc*, we impute values for Y_{it} with truncation. In the third option *notrunc* we impute values without truncation. Below we describe our algorithm for imputation.

In this general imputation algorithm, we first estimate a naive aging curve using a regression approach on only the observed data. Then we use that estimated curve to generate imputed values for the unobserved values, those with $\psi_{it} = 0$, ensuring that these unobserved values are below a smoothed boundary upper threshold at each age which is defined by a lower percentile of the observed values. Our default percentile is the 75^{th} percentile. The choice of the 75^{th} percentile is arbitrary, though we considered several choices, in particular, the 20^{th} and 50^{th} percentiles and found that using the 75^{th} percentile produced better estimation performance. It is sensible that there is an upper bound on the performance a player who is not in a top league would have; otherwise they would be in the top league. We then refit our estimated model using both the observed and imputed values and use this new estimated curve to generate a second set of imputed values. It is this second set of imputed values that we report as our estimate of the the mean aging curve $g(t)$.

The impetus behind the second imputation and curve estimation is to wash out any initial impact on estimated curve which is based solely on the observed values.

For the algorithm below:

1. Fit the model $Y_{it}^{obs} = \eta_{it}^{'} + \epsilon_{ij}$ for just the fully observed data, i.e. Y_{it} where $\psi_{it} = 1$. Calculate the estimated standard deviation of the $\hat{\epsilon}_{ij}$, call that σ_0.
2. Estimate a smoothed boundary via splines of performance for players in the NHL based upon the q^{th} percentile of players at each age, say the 75^{th} percentile. Call the boundary value $\tilde{\beta}_t$. Note that for imputation without truncation, *notrunc*, $\tilde{\beta}_t = \infty$.
3. Simulate values for the Y_{it} with $\psi_{it} = 0$ from a Normal distribution with mean $\eta_{it}^{'}$ and standard deviation σ_0 but truncated so they are not larger than $\tilde{\beta}_t$. That is the range of possible values for these Y_{it} is $(-\infty, \tilde{\beta}_t)$, we will denote this by $Y_{it}^{imp} \sim TN(\hat{\eta}_{it}^{'}, \hat{\sigma}_0^2, -\infty, \tilde{\beta}_t)$ where $Y \sim TN(\mu, \sigma^2, a, b)$ denotes a random variable Y with a Truncated Normal distribution with mean μ, variance σ^2 and takes values such that $a \leq Y \leq b$. Call these simulated values $Y_{it_k}^{imp}$.
4. Fit $Y_{it}^{full} = \eta_{it} + \epsilon_{it}$ using
$$\mathbf{Y}^{full} = \left\{ Y_{it_k}^{obs}, Y_{it_k}^{imp} \mid i = 1, \ldots, N, k = 1, \ldots, K \right\}.$$
5. We next impute the original missing values again. We do this because the first set of imputed values, the Y_{it}^{imp}'s, was based upon predicted values from a model that used only fully observed data. So then we imputed again, this time using $Y_{it}^{imp} \sim TN(\hat{\eta}_{it}, \hat{\sigma}_0^2, -\infty, \hat{\beta}_t)$.
6. We then refit the model above using $Y_{it}^{full} = g(t) + f(i, t) + \epsilon_{it}$ based upon the imputed data from the previous step.

Figure 4 depicts Steps 1 and 3, model fitting and imputation without truncation for one example player in one simulation (left) and the upper bounds used for truncation at each age for one of the simulations that did use truncation (right).

Our reasoning for steps five and six is that there is possibly some initial effect in the model $\eta_{it} = \mu_{0t} + \gamma_{0i}$ which comes from fitting to *only* fully observed data. By adding the second iteration we hope to wash out some of that initial impact. Our use of the boundary for imputation is based upon the idea that there are players who may be good enough to have their performance observed, that is play in whatever top league, but do have that opportunity, while at the same time there are players whose performance puts them well below the performance of other in the same league. Additionally, the estimation of η_{it} for generation of the imputed values was done two ways: using the quantile approach described above as well as the spline approach.

2.6 Player Effects

In addition to the basic model methodology and the data that we fit to these models we considered some additional fixed and random effects for players. The most common approach we used was a fixed constant player effect denoted as *fixed*. Our notation for a this fixed effect is γ_{0i}. We considered two random effects models one with quadratic random effects and the other with spline random effects. The former we denote by *random-quad* and it has the following functional form: $g_{0i} + g_{1i}t + g_{2i}t^2$. For the spline random effects component of our model our shorthand is *random-spline* and our function form is $g_{0i} + \xi_i(t)$. If the model did not have a fixed or random effect for player we use *none* in our shorthand.

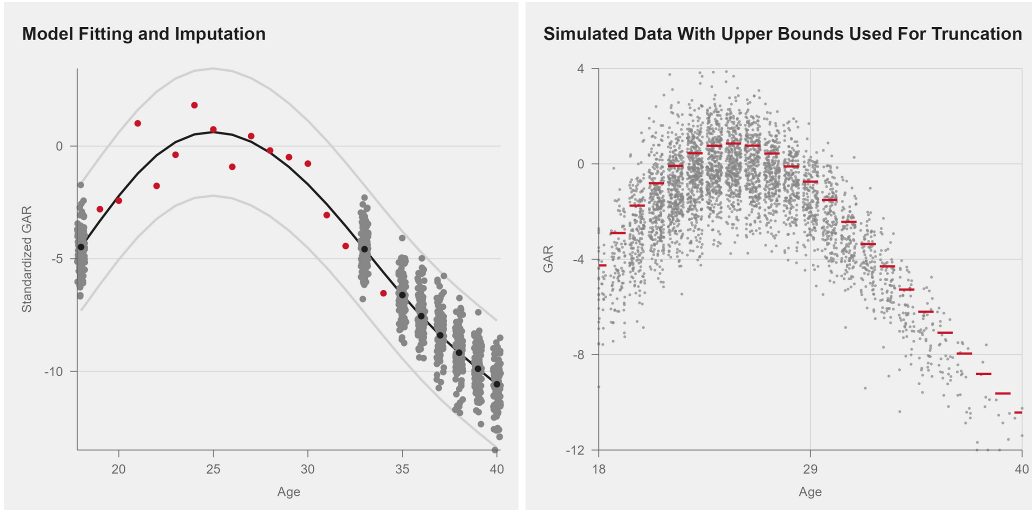

Fig. 4 Modeling fitting and imputation without truncation for one example player in one simulation (left), and all simulated players (gray dots) and upper bounds (red line) used for truncation (right). In the left figure, red dots are observed data for one example player from one simulation, the black line denotes fitted values using the cubic spline model *spline:trunc:fixed* for this player. The black dots, which lie on the line, are the means of the normal distribution that is used for imputing data for the ages for which data is missing, and the gray lines are 3 standard deviations away from the mean for those normal distributions.

Table 3 Summary of Estimation Methods

Method Name	Model formulation	Estimation of $g(t)$	Imputation
delta-plus	Non-parametric	Piecewise	No
spline:obs:none	$s(t)$	Natural Splines	No
spline:obs:fixed	$s(t) + \gamma_{0i}$	Natural Splines	No
spline:trunc:fixed	$s(t) + \gamma_{0i}$	Natural Splines	Truncated
spline:notrunc:fixed	$s(t) + \gamma_{0i}$	Natural Splines	Not Truncated
quant:trunc:fixed	$\zeta(t) + \gamma_{0i}$	Quantile Approx.	Truncated
quant:obs:none	$\zeta(t)$	Quantile Approx.	No
quad:trunc:fixed	$\gamma_0 + \gamma_{0i} + \gamma_1 t + \gamma_2 t^2$	Quad. Linear Model	Truncated
spline:trunc:random-quad	$s(t) + g_{0i} + g_{1i}t + g_{2i}t^2$	Natural Splines	Truncated
spline:trunc:random-spline	$s(t) + g_{0i} + \xi_i(t)$	Natural Splines	Truncated

In the subsections above, we have described three facets of our estimation approaches for $g(t)$. We combine these three facets to consider a range of estimation methods though we do not consider all the possible combinations of these approaches. We chose a specific subset of these combinations to make comparisons between our proposed methodologies and existing methodologies. Table 3 shows the full list of methods that we considered for evaluation along with shorthand and descriptions of the methodology. For example, *spline:trunc:fixed* uses natural splines to estimate $g(t)$ fit to both observed and truncated imputed data via a model that includes a fixed effect for each player. Likewise *quant:obs:none* use the quantile methodology described above to estimate the mean aging curve using only the observed Y_{it}'s. Somewhat unique is *quant:trunc:fixed* which generates the mean of the truncated imputed values via the quantile approach then fits a natural spline with fixed player effects to those truncated values.

3 Simulation Study

3.1 Simulation Design

We created a series of simulations to evaluate how well the various approaches described in the previous section estimate $g(t)$. To generate data we followed a similar approach to that of [15] and created an underlying smooth curve, and generated values for player performance based upon that curve. We then modelled the dropout of players through a missingness process that generated values for ψ_{it}. These simulations focused on the distribution of players and the methodology for their missingness.

3.2 Simulating individual player observations

Our simulated performance values for the i^{th} player in year t are generated in the following manner:

$$Y_{it} = \omega + \gamma_{0i} + a(t - t_{max})^2 + (b + b_i)(t - t_{max})^2 I_{\{t>t_{max}\}} + c(t - t_{max})^3 I_{\{t>t_{max}\}} + \epsilon_{it} \quad (7)$$

where $i = 1, \ldots, N$ and $t = t_1, \ldots, t_K$. The first three terms form a piecewise quadratic curves that serves as the underlying generating curve. The term $\gamma_{0i} \sim N(0, \sigma^2_\gamma)$. Thus this model is a piecewise cubic function with constant and quadratic player random effects.

We can rewrite Equation 7 to highlight the model components as:

$$Y_{it} = g(t) + f(i,t) + \epsilon_{it}, \quad (8)$$
$$g(t) = \omega + a(t - t_{max})^2 + b(t - t_{max})^2 I_{\{t>t_{max}\}} + c(t - t_{max})^3 I_{\{t>t_{max}\}}, \quad (9)$$
$$f(i,t) = \gamma_{0i} + b_i(t - t_{max})^2 I_{\{t>t_{max}\}}. \quad (10)$$

The term $\epsilon_{it} \sim N(0, \sigma^2_\epsilon)$ is random noise that simulates year-to-year randomness in player performance and variation unaccounted for elsewhere in the model. For our simulations we fixed $t_{max} = 25$, $a = -1/9$, $b = -6/1000$, $c = 45/10000$ $t_1 = 18$, $t_K = 40$, and $\sigma^2_\epsilon = 1$. Our choice of these values for a, b, and c was based upon trying to match the pattern of drop-in/drop-out for a subset of National Hockey League Forwards. We used varying values of N, ω, and σ_γ in our simulations. In particular we simulated a full factorial from $N = 300, 600, 1000$, $\omega = 0, 1$ and $\sigma_\gamma = 0.4, 0.8, 1.5$. The values of N that we chose were to illustrate the impact of sample size on estimation performance while also have a reasonable range of values. For ω which represents the maximum value that our simulated $g(t)$ would take, we want to have both a zero and a non-zero option. Finally, we based our choices for σ_γ on the standard deviation of the estimated player effects from fitting a natural spline with player effects model to standardized points per game from National Hockey League forwards. That value was approximately 0.8 and we subsequently chose values for σ_γ that were half and twice as large. For each combination of values for N, ω and σ_γ, we simulated 200 data sets and calculated the estimated $g(t)$ for each method on each of these data sets.

For example, the components of $g(t)$ from (7) are depicted in Figure 5. Shown in that figure are the underlying generating curve, the player intercept terms for overall player ability, the player quadratic curves for player aging, the random noise terms, and the final simulated player curves. These were taken from one simulation with the parameters $N = 600$, $\omega = 0$, and $\sigma_\gamma = 0.8$. All simulated players are shown in gray, and the curves generated for one particular simulated player

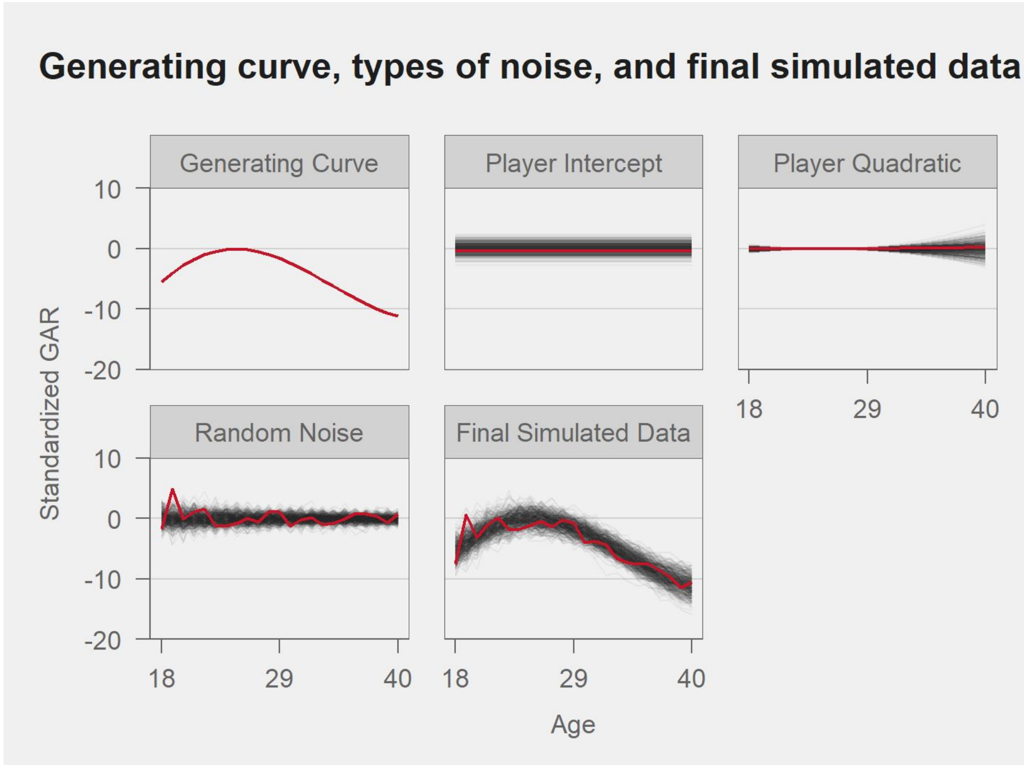

Fig. 5 The underlying generating curve, different types of noise, and the final simulated data used in one simulation. The gray curves represent simulated players. The curve for one simulated player is highlighted in red.

are shown in red. This highlighted player is slightly below average, intercept is slightly below zero, and ages slightly more slowly than average. The player quadratic is slightly concave up, which when added to the generating curve, gives a curve that indicates that this player's decline will not be as steep as an average player.

3.3 Simulating missingness of observations

For generating the ψ_{it} values for each player and age, we chose a methodology that tried to mirror the drop-in/drop-out behavior of forwards in the National Hockey League. Recall that $\psi_{it} = 1$ if observation Y_{it} was observed. Using the methodology from the previous section, we simulated Y_{it} for each player i and each age t between $t_1 = 18$ and $t_K = 40$. Using those values, we obtained ψ_{it} for each player via the following steps:

1. For each age t, round $N\pi_t$ to a whole number and call that n_t^+.
2. For each player i, calculate $\rho_{it} = \exp\{\sum_{k=1}^{t} Y_{ik}\}$.
3. Sample without replacement n_t^+ players from the N players with each having probability of selection $p_{it} = \frac{\rho_{it}}{\sum_j \rho_{jt}}$.

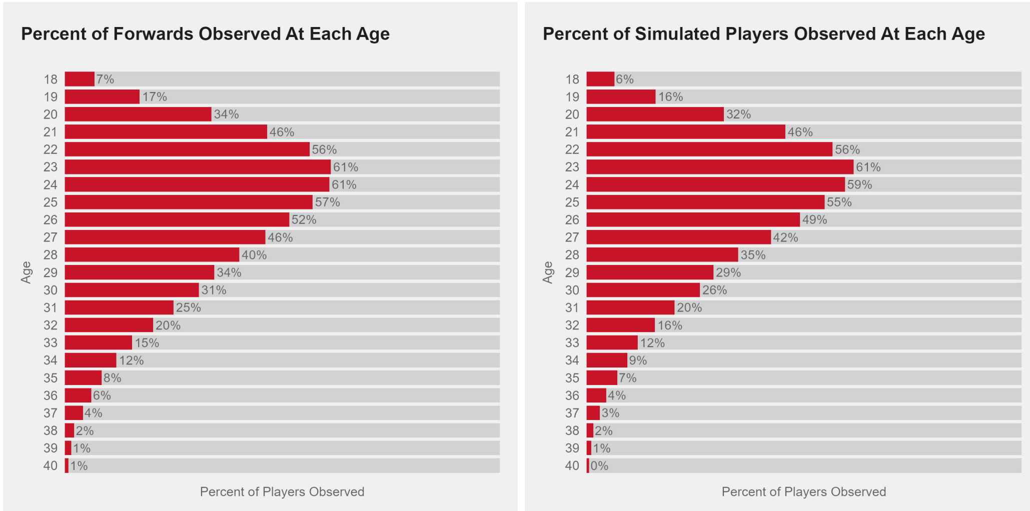

Fig. 6 Percent of players that are observed at each age for NHL forwards and for the simulated data. The distributions follow similar shapes and peak around ages 23-24.

4. Make $\psi_{it} = 1$ for the n_t^+ players selected in the previous step and $\psi_{it} = 0$ for the remaining players.

Simulations that used the cumulative performance missingness generation were paired with a full factorial of values for $\omega = 0, 1$, $N = 600, 800, 1000$, and $\sigma_\gamma = 0.4, 0.8, 1.5$. For simplicity we assumed that all simulated value of Y_{it} were observable, i.e. $\phi_{it} = 1$ for all i and t. To evaluate the methods proposed in the previous section, we generated simulated data following the data generation and missingness approaches described above in this section.

Figure 6 shows the percentage of players that are observed by age for NHL forwards (left) and for simulated data (right) with $N = 600$,$\omega = 0$ and $\sigma_\gamma = 0.8$.

3.4 Simulation Results

Next we consider how each of our methods performed in their estimation of the mean aging curve on their average error and on their overall shape.

3.4.1 Root Mean Squared Error

Estimated curves are first compared to $g(t)$ on their average root mean squared error (RMSE) at each age. Figure 7 shows RMSE by age, averaged across simulations, and split for each of simulations with 300 versus 1000 players. Six curves are shown; ones not presented in Figure 7 showed RMSE's that were larger and required a re-scaling of the y-axis that rendered comparisons of the remaining methods difficult.

RMSE at each age is lowest at around age 24, which corresponds to when the majority of NHL players in Figure 7 are observed. At entry (age 18) and at the end of careers (age 30 onwards), RMSE's tend to be higher.

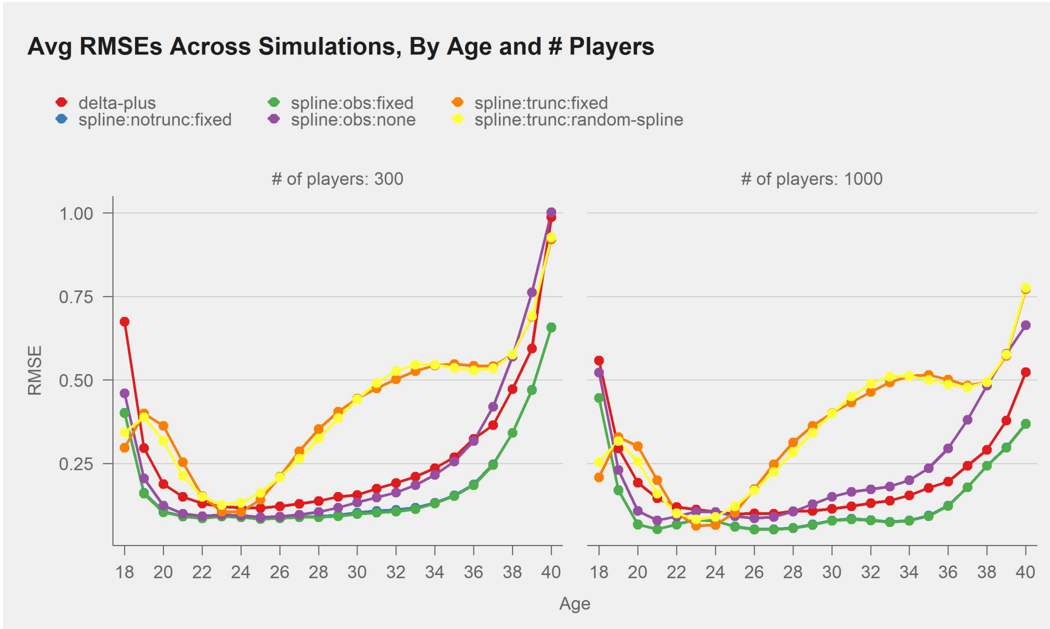

Fig. 7 Root Mean Squared Error by Age and Number of Players.

Overall, two methods – *spline:obs:fixed* and *spline:notrunc:fixed* — boast the lowest average RMSEs at each age. This suggests that for estimating age curves, either imputation or player specific intercept models are preferred. These two methods yield nearly identical average RMSEs at each age, which is why, the *spline:notrunc:fixed* (blue) method is not visible in Figure 7.

All models better estimate age curves with the larger sample size (the right facet of Figure 7). The impact of sample size is largest for the *delta-plus* method. For example, at age 40, *delta-plus* shows the 3rd lowest RMSE with 1,000 players, but the 5th lowest RMSE with 300 players. Additionally, *delta-plus* is worse with the lowest age group – for both 300 and 1000 simulated players, it shows the largest RMSE in Figure 7 among players at age 18. In general, simulations with higher standard deviations (1.5, versus 0.8 and 0.4) averaged higher RMSEs, although the impact of increasing standard deviation appeared uniform across age curve estimating method.

3.4.2 Shape Based Distance

To supplement simulation results based on RMSE, we use shape based distance (SBD, [11]), a metric designed to approximate the similarity of time series curves. SBD normalizes the cross-correlation between time series curves, which in our example is the estimated age curve and the truth. SBD scores ranges from 0 to 1, where lower scores reflect more similar curves.

Figure 8 shows a boxplot with shape based distances at each simulation for each method. As in Figure 7, *spline:obs:fixed* and *spline:notrunc:fixed* boast curve shapes that are closest to the true age curve in Figure 8. The basic spline model without a player intercept (*spline:obs:fixed*) has the third lowest median SBD; however, this method also boasts several outlying SBD observations. Similarly, the *delta-plus* shows SBD outliers. The method that assumes a quadratic response surface has the

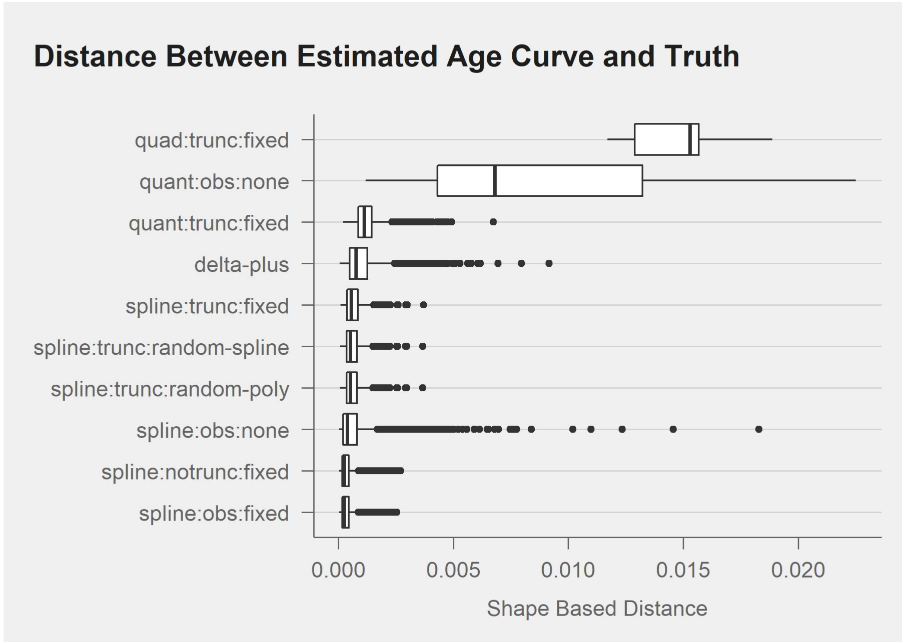

Fig. 8 Shape Based Distance between true and estimated age curves, across simulations

highest median SBD, and is in the top row of Figure 8, suggesting that quadratic-based approaches may not identify the shape of age curves.

4 Application to National Hockey League Data

To illustrate the impact of the methods proposed in this paper, we apply some of the approaches in this paper to data on 1079 NHL forwards who were born on or after January 1, 1970. The data were obtained from `www.eliteprospects.com` and only players who played at least one season in the National Hockey League were included. For our measure of player performance, Y_{it}, we chose a standardized points (goals plus assists) per game where the z-score was calculated relative to the mean and standard deviation for the NHL in that particular season. Our selection of this metric was based upon its availability for all players across the range of seasons (1988-89 season to 2018-19 season).

In Figure 9 we show estimated age curves for NHL player points per game using different methods. The differences in our two best methods, *spline:notrunc:fixed* (red) and *spline:obs:fixed* (lightred), neither of which use truncation, are not perceptible as their lines overlap. The methods with truncation, *spline:trunc:fixed* (blue), *spline:trunc:random-quad* (gray) and *spline:trunc:random-spline* (black) are very similar and have overlapping curves as well. Those curves are lower than the curves from our best methods, which is expected since imputed values were taken from a truncated normal. The overall pattern for all of these curves does not seem quadratic.

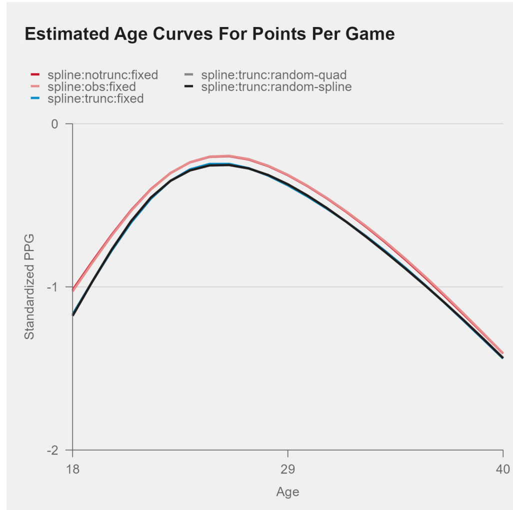

Fig. 9 Estimated age curves for NHL points per game using different methods.

5 Discussion

Understanding how player performance changes as players age is important in sports, particularly for team management who need to sign players to contracts. In this paper we have proposed and evaluated several new methodologies for estimation of mean player aging curves. In particular, we have presented formal methods for adding imputed data to augment the missingness that regularly appears in data from some professional sports leagues. The models that performed best in our simulation study had either a flexible form or incorporated player effects (through imputation, or directly). This paper has also presented a framework for incorporating imputed data into estimation of player aging effects. With player effects, the age curve term contains information about the relative changes in performance from age to age and overall performance is part of the player term. When estimation is done with only observed data, the relative changes in the mean aging curve are only utilized where a particular player is present. Without player effects as a model component, the individual player and the overall are entangled. Specifically, we find that a spline methodology with fixed player effects allow perform better than other methods and these methods have the flexibility to appropriately and efficiently estimate player aging curves.

There are some additional considerations that might improve the performance of the methods we have considered. One possibility would be to consider a fully Bayesian approach that treats all of the unknown aspects of the model both the $g(t) + f(i,t)$ and unobserved Y_{it}'s as random variables. This approach could consider a complete posterior inference given the uncertainty in estimation. Along similar lines, an approach that does multiple imputation for each unobserved Y_{it} could improve performance. Another assumption that has been made in the literature is to treat ages as whole numbers. It certainly seems possible for a regression based approaches for estimation of $g(t)$ to deal in fractional ages t for a given season.

Our simulation study was focused on mean aging curves that yielded similar player drop-in/drop-out patterns to those observed in a particular professional league, the National Hockey League. It would certainly be reasonable to consider other functional forms for the underlying $g(t)$ and $f(i,t)$, though we believe that the results from such a study would be in line with those found in this paper. Another avenue of possible future work would be to consider additional mechanisms for generation of the ψ_{it} values.

Overall the novel methods proposed and evaluated in this paper via simulation study have improved our understanding of how to estimate player aging curves. It is clear from the results in this paper that the best methods for estimation of player aging are those that have model flexibility and that include player effects. The use of imputation also has potential to impact this methodology and, thus, being aware of what we don't observe can make our estimation stronger.

Acknowledgements We would like to thank CJ Turturo for making his code available.

Conflict of interest

The authors declare that they have no conflict of interest.

References

1. Albert, J.: Smoothing career trajectories of baseball hitters. Unpublished manuscript, Bowling Green State University, at bayes. bgsu. edu/papers/career_trajectory. pdf (2002)
2. Berry, S.M., Reese, C.S., Larkey, P.D.: Bridging different eras in sports. Journal of the American Statistical Association **94**(447), 661–676 (1999)
3. Bradbury, J.C.: Peak athletic performance and ageing: evidence from baseball. Journal of Sports Sciences **27**(6), 599–610 (2009)
4. Brander, J.A., Egan, E.J., Yeung, L.: Estimating the effects of age on nhl player performance. Journal of Quantitative Analysis in Sports **10**(2), 241–259 (2014)
5. Fair, R.C.: Estimated age effects in baseball. Journal of Quantitative Analysis in Sports **4**(1) (2008)
6. Judge, J.: An approach to survivor bias in baseball. Baseball Prospectus (https://www.baseballprospectus.com/news/article/59491/an-approach-to-survivor-bias-in-baseball/) (2020)
7. Judge, J.: The delta method, revisited: Rethinking aging curves. Baseball Prospectus (https://www.baseballprospectus.com/news/article/59972/the-delta-method-revisited/) (2020)
8. Kovalchik, S.A., Stefani, R.: Longitudinal analyses of olympic athletics and swimming events find no gender gap in performance improvement. Journal of Quantitative Analysis in Sports **9**(1), 15–24 (2013)
9. Lailvaux, S.P., Wilson, R., Kasumovic, M.M.: Trait compensation and sex-specific aging of performance in male and female professional basketball players. Evolution **68**(5), 1523–1532 (2014)
10. Lichtman, M.: How do baseball players age. Fan Graphs (https://tht.fangraphs.com/how-do-baseball-players-age-part-2/) (2009)
11. Paparrizos, J., Gravano, L.: k-shape: Efficient and accurate clustering of time series. In: Proceedings of the 2015 ACM SIGMOD International Conference on Management of Data, pp. 1855–1870 (2015)
12. R Development Core Team: R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria (2007). URL `http://www.R-project.org`. ISBN 3-900051-07-0
13. Schulz, R., Musa, D., Staszewski, J., Siegler, R.S.: The relationship between age and major league baseball performance: Implications for development. Psychology and aging **9**(2), 274 (1994)
14. Tulsky, E.: How shot attempt differential changes with age. `https://www.sbnation.com/nhl/2014/3/20/5528472/nhl-stats-corsi-vs-age` (2014)
15. Turtoro, C.: Flexible aging in the nhl using gam. `https://rpubs.com/cjtdevil/nhl_aging` (2019)
16. Vaci, N., Cocić, D., Gula, B., Bilalić, M.: Large data and bayesian modeling—aging curves of nba players. Behavior research methods **51**(4), 1544–1564 (2019)
17. Villaroel, C., Mora, R., Gonzalez-Parra, G.C.: Elite triathlete performance related to age. Journal of Human Sport and Exercise **6**(2), 363–373 (2011)
18. Wakim, A., Jin, J.: Functional data analysis of aging curves in sports. arXiv preprint arXiv:1403.7548 (2014)