Benjamin S. Baumer*, Shane T. Jensen and Gregory J. Matthews

openWAR: An open source system for evaluating overall player performance in major league baseball

Abstract: Within sports analytics, there is substantial interest in comprehensive statistics intended to capture overall player performance. In baseball, one such measure is wins above replacement (WAR), which aggregates the contributions of a player in each facet of the game: hitting, pitching, baserunning, and fielding. However, current versions of WAR depend upon proprietary data, ad hoc methodology, and opaque calculations. We propose a competitive aggregate measure, *openWAR*, that is based on public data, a methodology with greater rigor and transparency, and a principled standard for the nebulous concept of a "replacement" player. Finally, we use simulation-based techniques to provide interval estimates for our *openWAR* measure that are easily portable to other domains.

Keywords: baseball; R; reproducibility; simulation; statistical modeling.

1 Introduction

In sports analytics, researchers apply statistical methods to game data in order to estimate key quantities of interest. In team sports, arguably the most fundamental challenge is to quantify the contributions of individual players towards the collective performance of their team. In all sports the ultimate goal is winning and so the ultimate measure of player performance is that player's overall contribution to the number of games that his team wins. Although we focus on a particular measure of player contribution, wins above replacement (WAR) in major league

baseball, the issues and approaches examined in this paper apply more generally to any endeavor to provide a *comprehensive* measure of individual player performance in sports.

A common comprehensive strategy used in sports such as basketball, hockey, and soccer is the *plus/minus* measure (Kubatko et al. 2007; Macdonald 2011). Although many variations of plus/minus exist, the basic idea is to tabulate changes in team score during each player's appearance on the court, ice, or pitch. If a player's team scores more often than their opponents while he is playing, then that player is considered to have a positive contribution. Whether those contributions are primarily offensive or defensive is not delineated, since the fluid nature of these sports make it extremely difficult to separate player performance into specific aspects of gameplay.

In contrast, baseball is a sport where the majority of the action is discrete and player roles are more clearly defined. This has led to a historical focus on separate measures for each aspect of the game: hitting, baserunning, pitching and fielding. For measuring hitting, the three most-often cited measures are batting average (*BA*), on-base percentage (*OBP*) and slugging percentage (*SLG*) which comprise the conventional "triple-slash line" (*BA/OBP/SLG*). More advanced measures of hitting include runs created (James 1986), and linear weights-based metrics like weighted on-base average (*wOBA*) (Tango, Lichtman, and Dolphin 2007) and extrapolated runs (Furtado 1999). Similar linear weights-based metrics are employed in the evaluation of baserunning (Lichtman 2011).

Classical measures of pitching include walks and hits per innings pitched (*WHIP*) and earned run average (*ERA*). McCracken (2001) introduced defense independent pitching measures (DIPS) under the theory that pitchers do not exert control over the rate of hits on balls put into play. Additional advancements for evaluating pitching include fielding independent pitching (*FIP*) (Tango 2003) and *xFIP* (Studeman 2005). Measures for fielding include ultimate zone rating (*UZR*) (Lichtman 2010), defensive runs saved (*DRS*) (Fangraphs Staff 2013), and spatial aggregate fielding evaluation (*SAFE*) (Jensen, Shirley, and Wyner 2009). For a more thorough review of the measures for different

*Corresponding author: Benjamin S. Baumer, Smith College – Statistical and Data Sciences, Northampton, MA, USA, e-mail: bbaumer@smith.edu
Shane T. Jensen: Statistics, The Wharton School, University of Pennsylvania, Philadelphia, PA, USA
Gregory J. Matthews: Mathematics and Statistics, Loyola University Chicago, Chicago, IL, USA

aspects of player performance in baseball, we refer to the reader to Thorn and Palmer (1984), Lewis (2003), Albert and Bennett (2003), Schwarz (2005), Tango et al. (2007), Baumer and Zimbalist (2014).

Having separate measures for the different aspects of baseball has the benefit of isolating different aspects of player ability. However, there is also a strong desire for a comprehensive measure of overall player performance, especially if that measure is closely connected to the success of the team. The ideal measure of player performance is each player's contribution to the number of games that his team wins. The fundamental question is how to apportion this total number of wins to each player, given the wide variation in the performance and roles among players.

Win shares (James and Henzler 2002) was an early attempt to measure player contributions on the scale of wins. Value over Replacement player (Jacques 2007) measures player contribution on the scale of runs relative to a *baseline* player. An intuitive choice of this baseline comparison is a "league average" player but since average players themselves are quite valuable, it is not reasonable to assume that a team would have the ability to replace the player being evaluated with another player of league average quality. Rather, the team will likely be forced to replace him with a minor league player who is considerably less productive than the average major league player. Thus, a more reasonable choice for this baseline comparison is to define a "replacement" level player as the typical player that is readily accessible in the absence of the player being evaluated.

The desire for a comprehensive summary of an individual baseball player's contribution on the scale of team wins, relative to a replacement level player, has culminated in the popular measure of WAR. The three most popular existing implementations of WAR are: *fWAR* (Slowinski 2010), *rWAR* (sometimes called *bWAR*) (Forman 2010, 2013), and *WARP* (Baseball Prospectus Staff 2013). A thorough comparison of the differences in their methodologies is presented in our supplementary materials.

WAR has two virtues that have fueled its recent popularity. First, having an accurate assessment of each player's contribution allows team management to value each player appropriately, both for the purposes of salary and as a trading chip. Second, the units and scale are easy to understand. To say that Miguel Cabrera is worth about seven wins above replacement means that losing Cabrera to injury should cause his team to drop about seven games in the standings over the course of a full season. Unlike many baseball measures, no advanced statistical knowledge is required to understand this

statement about Miguel Cabrera's performance. Accordingly, WAR is now cited in mainstream media outlets like *ESPN*, *Sports Illustrated*, *The New York Times*, and the *Wall Street Journal*.

In recent years, this concept has generated significant interest among baseball statisticians, writers, and fans (Schoenfield 2012). WAR values have been used as quantitative evidence to buttress arguments for decisions upon which millions of dollars will change hands (Rosenberg 2012). Recently, WAR has achieved two additional hallmarks of mainstream acceptance: 1) the 2012 American League MVP debate seemed to hinge upon a disagreement about the value of WAR (Rosenberg 2012); and 2) it was announced that the Topps baseball card company will include WAR on the back of their next card set (Axisa (2013). Testifying to the static nature of baseball card statistics, WAR is only the second statistic (after OPS) to be added by Topps since 1981.

1.1 Problems with WAR

While WAR is comprehensive and easily-interpretable as described above, the use of WAR as a statistical measure of player performance has two fundamental problems: a lack of uncertainty estimation and a lack of reproducibility. Although we focus on WAR in particular, these two problems are prevalent for many measures for player performance in sports as well as statistical estimators in other fields of interest.

WAR is usually misrepresented in the media as a known quantity without any evaluation of the uncertainty in its value. While it was reported in the media that Miguel Cabrera's WAR was 6.9 in 2012, it would be more accurate to say that his WAR was *estimated to be* 6.9 in 2012, since WAR has no single definition. The existing WAR implementations mentioned above (*fWAR*, *rWAR* and *WARP*) do not publish uncertainty estimates for their WAR values. As Nate Silver articulated in this 2013 ASA presidential address, journalists struggle to correctly interpret probability, but it is the duty of statisticians to communciate uncertainty (Rickert 2013).

Even more important than the lack of uncertainty estimates is the lack of reproducibility in current WAR implementations (*fWAR*, *rWAR* and *WARP*). The notion of reproducible research began with Knuth's introduction of *literate programming* (Knuth 1984). The term *reproducible research* first appeared about a decade later (Claerbout 1994), but quickly attracted attention. Buckheit and Donoho (1995) asserted that a scientific publication in a computing field represented only an advertisement

for the scholarly work – not the work itself. Rather, "the actual scholarship is the complete software development environment and complete set of instructions which generated the figures" (Buckheit and Donoho (1995). Thus, the burden of proof for reproducibility is on the scholar, and the publication of computer code is a necessary, but not sufficient condition. Advancements in computing like the `knitr` package for R (Xie 2014) made reproducible research relatively painless. It is in this spirit that we understand "reproducibility."

Interest in reproducible research has exploded in recent years, amid an increasing realization that many scientific findings are not reproducible (Naik 2011; Zimmer 2012; Ioannidis 2013; *Nature* Editorial 2013; The Economist Editorial 2013; Johnson 2014). Transparency in sports analytics is more tenuous than other scientific fields since much of the cutting edge research is being conducted by proprietary businesses or organizations that are not interested in sharing their results with the public.

To the best of our knowledge, no open-source implementations of *rWAR*, *fWAR*, or *WARP* exist in the public domain and the existing implementations do not meet the standard for reproducibility outlined above. Two of the three methods use proprietary data sources, and the third implementation, despite making overtures toward openness, is still not reproducible without needing extra proprietary details about their methods. This is frustrating since these WAR implementations are essentially "black boxes" containing *ad hoc* adjustments and lacking in a unified methodology.[1]

1.2 Contributions of openWAR

We address both the lack of uncertainty estimates and the lack of reproducibility in WAR by presenting a fully transparent statistical model based on our *conservation of runs* framework with uncertainty in our model-based WAR values estimated by resampling methods. In this paper we present *openWAR*, a reproducible and fully open-source reference implementation for estimating the WAR for each player in major league baseball.

In Section 3, we introduce the notion of *conservation of runs*, which forms the backbone of our WAR calculations. The central concept of our model is that the positive and negative consequences of all runs scored in the game

of baseball must be allocated across four types of baseball performance: 1) batting; 2) baserunning; 3) fielding; and 4) pitching. While there are four components of *openWAR*, each is viewed as a component of our unified conservation of runs model.

In contrast, the four components of WAR are estimated separately in each previous WAR implementation (*rWAR*, *fWAR*, or *WARP*) and these implementations only provide point estimates of WAR. We employ resampling techniques to derive uncertainty estimates for *openWAR*, and report those alongside our point estimates. While the apportionment scheme that we outline here is specific to baseball, the resampling-based uncertainty measures presented in Section 4 are generalizable to any sport.

Our goal in this effort is to provide a coherent and principled fully open-source estimate of player performance in baseball that may serve as a reference implementation for the statistical community and the media. Our hope is that in time, we can solidify WAR's important role in baseball by rallying the community around an open implementation. In addition to the full model specification provided in this paper, our claim of reproducibility is supported by the simultaneous release of a software package for the open-source statistical computing environment R, which contains all of the code necessary to download the data and compute *openWAR*.

1.3 OpenWAR vs. previous WAR implementations

In our approach, WAR for a player is defined as the sum of all of their marginal contributions in each of the four aspects of the game, relative to a hypothetical replacement level player after controlling for potential confounders (e.g., ballpark, handedness, position, etc.). Previous WAR estimates, such as *rWAR*, *fWAR*, and *WARP*, serve as an inspiration for our approach but we make several key assumptions that differentiate our WAR philosophy from these previous efforts. In addition to using higher resolution ball-in-play data than previous methods, we also have several differences in perspective.

First, *openWAR* is a retrospective measure of player performance – it is not a measure of player ability to be used for forecasting. It is not context-independent, because we feel that context is important for accurate accounting of what actually happened. Second, we control for defensive position in both our batting and fielding estimates. We do this at the plate appearance level, which allows for more refined comparisons of players to their appropriate peer group. Third, we believe that credit or

---

1 For example, *rWAR* and *fWAR* are constrained to sum to 1000 in a season for no apparent substantive reason. See Section 3.8 for a fuller discussion.

blame for hits on balls in play should be shared between the pitcher and fielders. We use the location of the batted ball to inform the extent to which they should be shared. Fourth, we propose a new definition of replacement level based on distribution of performance beyond the 750 active major league players that play each day, which is different from existing implementations. Thus, *openWAR* is not an attempt to reverse-engineer any of the existing implementations of WAR. Rather, it is a new, fully open-source attempt to estimate player performance on the scale of wins above replacement.

2 Preliminaries: Expected runs

A major hurdle in producing a reproducible version of WAR is the data source. *openWAR* uses data published by Major League Baseball Advanced Media for use in their GameDay web application (Bowman 2013). A thorough description of the MLBAM data set obtainable using the *openWAR* package is presented in our supplementary materials.

Our *openWAR* implementation is based upon a *conservation of runs* framework, which tracks the changes in the number of expected runs scored and actual runs scored resulting from each in-game hitting event. The starting point for these calculations is establishing the number of runs that would be expected to score as a function of the current state of the game. Here, we illustrate that the *expected run matrix* – a common sabermetric construction dating back to the work of Lindsey (1959, 1961) – can be used to model these quantities.[2]

There are 24 different states in which a baseball game can be at the beginning of a plate appearance: three states corresponding to the number of outs (0, 1, or 2) and eight states corresponding to the base configuration (bases empty, man on first, man on second, man on third, man on first and second, man on first and third, man on second and third, bases loaded). A 25th state occurs when three outs are achieved by the defensive team and the half-inning ends.

We define expected runs at the start of a plate appearance given the current state of an inning,

$$\rho(o, b) = \mathbb{E}[R | startOuts = o, startBases = b],$$

where R is a random variable counting the number of runs that will be scored from the current plate appearance to the end of the half-inning when three outs are achieved.

startOuts is the number of outs at the beginning of the plate appearance, and *startBases* is the base configuration at the beginning of the plate appearance. The value of $\rho(o, b)$ is estimated as the empirical average of the number of runs scored (until the end of the half-inning) whenever a game was in state (o, b). Note that the value of the three out state is defined to be zero [i.e., $\rho(3, 0) \equiv 0$].

We can then define the *change* in expected runs due to a particular plate appearance as

$$\Delta\rho = \rho_{endState} - \rho_{startState},$$

where $\rho_{startState}$ and $\rho_{endState}$ are the values of the expected runs in the state at the beginning of the plate appearance and the state at the end of the plate appearance, respectively. However, we must also account for the actual number of runs scored r in that plate appearance, which gives us

$$\delta = \Delta\rho + r.$$

For each plate appearance i, we can calculate δ_i from the observed start and end states for that plate appearance as well as the observed number of runs scored. This quantity δ_i can be interpreted as the total run value that the particular plate appearance i is worth. Sabermetricians often refer to this quantity as RE24 (Appelman 2008).[3]

3 openWAR model

The central idea of our approach to valuing individual player contributions is the assumption that every run value δ_i gained by the offense as a result of a plate appearance i is accompanied by a corresponding $-\delta_i$ gained by the defense. We call this principle our *conservation of runs* framework. The remainder of this section will outline a principled methodology for apportioning δ_i among the offensive players and apportioning $-\delta_i$ among the defensive players involved in plate appearance i.

3.1 Adjusting offensive run values

As outlined above, δ_i is the run value for the offensive team as a result of plate appearance i. We begin our modeling of offensive run value by adjusting δ_i for several factors beyond the control of the hitter or baserunners that make it difficult to compare run values across contexts. Specifically, we want to first adjust for the ballpark of the event and any platoon advantage the batter may have over the pitcher (i.e., a left-handed batter against a right-handed

---

2 The expected run matrix is also the basis for Markov Chain models, which have been used to, among other things, optimize batting order (Freeze 1974; Pankin 1978; Bukiet, Harold, and Palacios 1997; Sokol 2003).

3 RE for "run expectancy" and 24 for the 24 distinct states.

pitcher). We control for these factors by fitting a linear regression model to the offensive run values,

$$\delta_i = \boldsymbol{B}_i \cdot \boldsymbol{\alpha} + \epsilon_i, \tag{1}$$

where the covariate vector $\boldsymbol{B}_i$ contains a set of indicator variables for the specific ballpark for plate appearance i and an indicator variable for whether or not the batter has a platoon advantage over the pitcher. The coefficient vector α contains the effects of each ballpark and the effect of a platoon advantage on the offensive run values. Regression-based ballpark factors have been previously estimated by Acharya et al. (2008). Estimated coefficients $\hat{\alpha}$ are calculated by ordinary least squares using every plate appearance in our dataset.

The estimated residuals from the regression model (1),

$$\hat{\epsilon}_i = \delta_i - \boldsymbol{B}_i \cdot \hat{\alpha} \tag{2}$$

represent the portion of the offensive run value δ_i that is not attributable to the ballpark or platoon advantage, and so we refer to them as *adjusted* offensive run values.

3.2 Baserunner run values

The next task is determining the portion of $\hat{\epsilon}_i$ that is attributable to the baserunners for each plate appearance i based on the following principle: *baserunners should only get credit for advancement beyond what would be expected given their starting locations, the number of outs, and the hitting event that occurred*. We can estimate this expected baserunner advancement by fitting a second regression model to our adjusted offensive run values,

$$\hat{\epsilon}_i = \boldsymbol{S}_i \cdot \boldsymbol{\beta} + \eta_i, \tag{3}$$

where the covariate vector $\boldsymbol{S}_i$ consists of: 1) a set of indicator variables that indicate the specific game state (number of outs, locations of baserunners) at the start of plate appearance i and; 2) the hitting event (e.g., single, double, etc.) that occurred during plate appearance i. The 31 event types in the MLBAM data set that describe the outcome of a plate appearance are listed in our supplementary materials. Estimated coefficients $\hat{\beta}$ are calculated by ordinary least squares using every plate appearance in our dataset. The estimated residuals from the regression model (3),

$$\hat{\eta}_i = \hat{\epsilon}_i - \boldsymbol{S}_i \cdot \hat{\boldsymbol{\beta}}, \tag{4}$$

represent the portion of the adjusted offensive run value that is attributable to the baserunners. If the baserunners take extra bases beyond what is expected, then $\hat{\eta}_i$ will be positive, whereas if they take fewer bases or get

thrown out then $\hat{\eta}_i$ will be negative. Note that $\hat{\eta}_i$ also contains the baserunning contribution of the hitter for plate appearance i.

We apportion baserunner run value, $\hat{\eta}_i$ amongst the individual baserunners involved in plate appearance i based upon their expected base advancement compared to their actual base advancement. If we denote k_{ij} as the number of bases advanced by the jth baserunner after hitting event m_i, then we can use all plate appearances in our dataset to calculate the empirical probability

$$\hat{\kappa}_{ij} = \Pr(K \le k_{ij} \mid m_i)$$

that a typical baserunner would have advanced at least the k_{ij} bases that baserunner j did advance in plate appearance i. If baserunner j does worse than expected (e.g., not advancing from second on a single) then $\hat{\kappa}_{ij}$ will be small whereas if baserunner j takes an extra base (e.g., scoring from second on a single), then $\hat{\kappa}_{ij}$ will be large. These advancement probabilities $\hat{\kappa}_{ij}$ are used as weights for apportioning the baserunner run value, $\hat{\eta}_i$, to each individual baserunner,

$$\mathrm{RAA}_{ij}^{br} = \frac{\hat{\kappa}_{ij}}{\sum_j \hat{\kappa}_{ij}} \cdot \hat{\eta}_i \tag{5}$$

The value RAA_{ij}^{br} is the runs above average attributable to the jth baserunner on the ith plate appearance.

3.3 Hitter run values

As calculated in (4) above, $\hat{\eta}_i$ represents the portion of the adjusted offensive run value $\hat{\epsilon}_i$, that is attributable to the baserunners during plate appearance i. The remaining portion of the adjusted offensive run value,

$$\hat{\mu}_i = \hat{\epsilon}_i - \hat{\eta}_i \tag{6}$$

is the adjusted offensive run value attributable to the hitter during plate appearance i. Our remaining task for hitters is to calibrate their hitting performance relative to the expected hitting performance based on all players at the same fielding position.[4] We fit another linear regression model to adjust the hitter run value by the hitter's fielding position,

---

4 This is necessary because players who play more difficult fielding positions tend to be weaker hitters. In the extreme case, pitchers as a group are far worse hitters than those who play any other position. Thus, to evaluate the batting performance of pitchers without correcting for their defensive position would result in almost every pitcher being assigned a huge negative value for their batting performance. This would result in a dramatic undervaluation of pitchers (in the National League, at least) since they are obligated to hit while they are pitching.

$$\hat{\mu}_i = \boldsymbol{H}_i \cdot \boldsymbol{\gamma} + \nu_i \tag{7}$$

where the covariate vector $\boldsymbol{H}_i$ consists of a set of indicator variables for the fielding position of the hitter in plate appearance i. Note that pinch-hitter (PH) and designated hitter (DH) are also valid values for hitter position. Estimated coefficients $\hat{\boldsymbol{\gamma}}$ are calculated by ordinary least squares using every plate appearance in our dataset. The estimated residuals from this regression model,

$$\mathrm{RAA}_i^{hit} = \hat{\nu}_i = \hat{\mu}_i - \boldsymbol{H}_i \cdot \hat{\boldsymbol{\gamma}} \tag{8}$$

represent the run values (above the average for the hitter's position) for the hitter in each plate appearance i.

3.4 Apportioning defensive run values

As we discussed in Section 2, each plate appearance i is associated with a particular run value δ_i, and we apportioned the offensive run value δ_i between the hitters and various baserunners in Sections 3.1–3.3. Now, we must apportion the defensive run value $-\delta_i$ between the pitcher and various fielders involved in plate appearance i.

The degree to which the pitcher (versus the fielders) is responsible for the run value of a ball in play depends on how difficult that batted ball was to field. Surely, if the pitcher allows a batter to hit a home run, the fielders share no responsibility for that run value. Conversely, if a routine groundball is muffed by the fielder, the pitcher should bear very little responsibility.

We assign the entire defensive run value $-\delta_i$ to the pitcher for any plate appearance that does not result in a ball in play (e.g., strikeout, home run, hit by pitch, etc.). For balls hit into play, we must estimate the probability p that each ball-in-play would result in an out given the location that ball in play was hit.

The MLBAM data set contains (x, y)-coordinates that give the location of each batted ball, and we use a two-dimensional kernel density smoother (Wand 1994) to estimate the probability of an out at each coordinate in the field,

$$\hat{p}_i = f(x_i, y_i)$$

Figure 1 gives the contour plot of our estimated probability of an out, $\hat{p}_i$, for a ball in play i hit to coordinate (x_i, y_i) in the field. For that ball in play i, we use $\hat{p}_i$ to divide the responsibility between the pitcher and the fielders. Specifically, we apportion

$$\begin{aligned} \delta_i^p &= -\delta_i \cdot (1 - p_i) \quad \text{to the pitcher} \\ \delta_i^f &= -\delta_i \cdot p_i \quad\quad\;\; \text{to the fielders} \end{aligned}$$
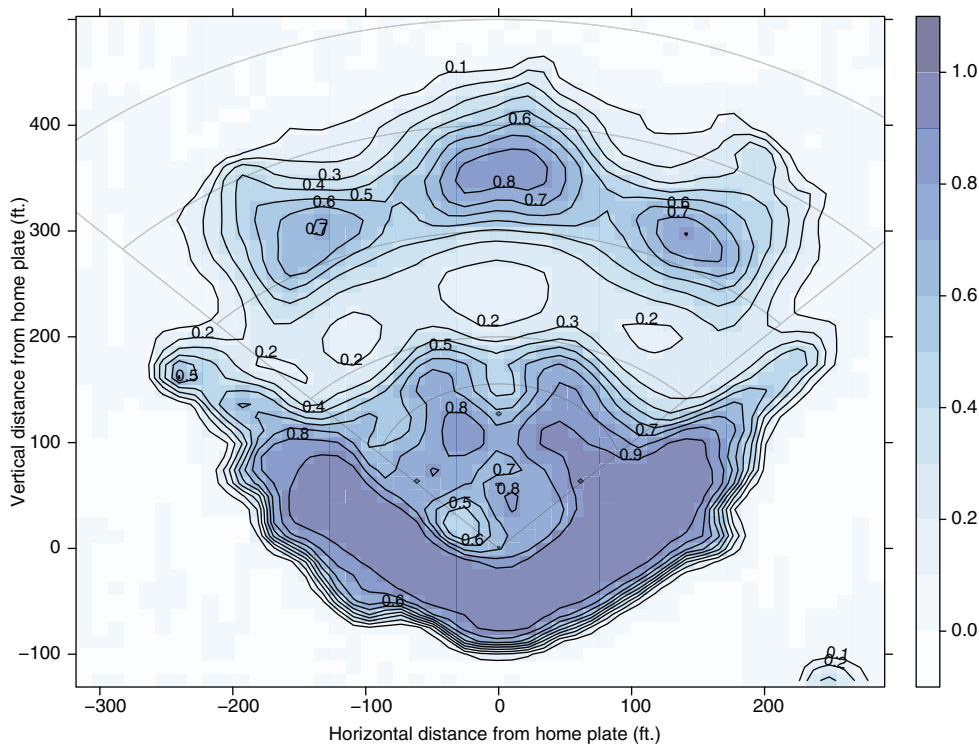

Figure 1: Contour plot of our estimated probability of an out $\hat{p}_i$ for a ball in play i as a function of the coordinates (x_i, y_i) for that ball in play. Numerical labels give the estimated probability of an out at that contour line.
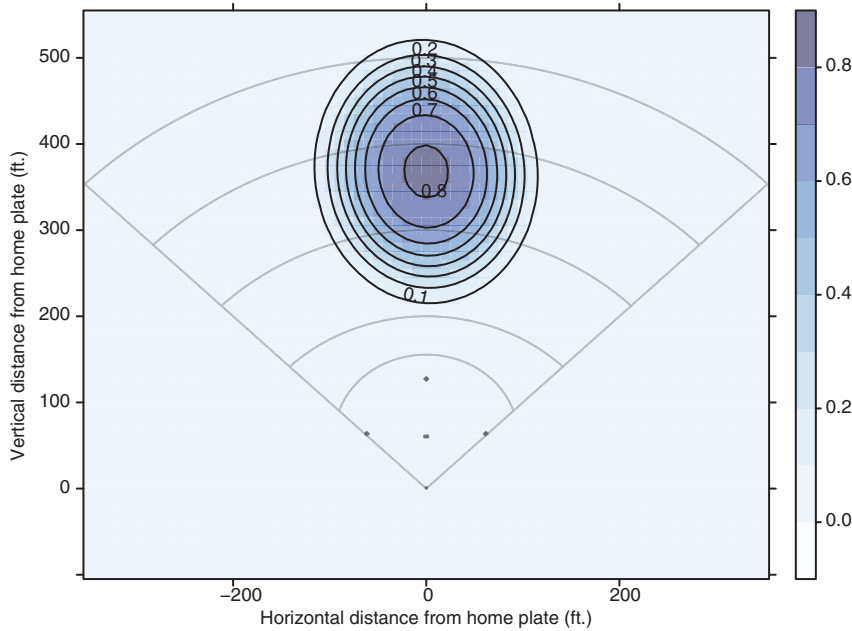
Figure 2: Contour plot of fielding model for centerfielders. The contours indicate the expected probability that any given centerfielder will catch a fly ball hit to the corresponding location on the field.

The fielders bear more responsibility for a ball in play that is relatively easy to field ($\hat{p}_i$ near 1) whereas a pitcher bears more responsibility for a ball in play that is relatively hard to field ($\hat{p}_i$ near 0).

3.5 Fielding run values

In Section 3.4 above, we allocated the run value δ_i^f to the fielders. We must now divide that run value amongst the nine fielders who are potentially responsible for ball in play i. For each fielding position ℓ, we use all balls in play to fit a logistic regression model,

$$logit(p_{i\ell}) = \boldsymbol{X}_i \cdot \boldsymbol{\theta}_\ell$$

where $p_{i\ell}$ is the probability that fielder ℓ makes an out[5] on ball in play i hit to coordinate (x_i, y_i) in the field. The covariate vector $\boldsymbol{X}_i$ consists of linear, quadratic and interaction terms of x_i and y_i. The quadratic terms are necessary to incorporate the idea that a player is most likely to field a ball hit directly at him, and the interaction term captures the notion that it may be easier to make plays moving to one side (e.g., shortstops have better range moving to their

left since they are moving towards first base). Estimates of the coefficients $\hat{\boldsymbol{\theta}}_\ell$ are calculated from all balls in play. As an example, the surface of our fielding model for centerfielders is illustrated in Figure 2.

For ball in play i, we use the coordinates (x_i, y_i) and the estimated coefficients $\hat{\boldsymbol{\theta}}_\ell$ for each fielding position ℓ to estimate the probability $\hat{p}_{i\ell}$ that fielder ℓ makes an out on ball in play i. We normalize these probabilities across positions to estimate the responsibility $s_{i\ell}$

$$\hat{s}_{i\ell} = \frac{\hat{p}_{i\ell}}{\Sigma_\ell \hat{p}_{i\ell}},$$

of the ℓth fielder on the ith play, which gives us the run value $\delta_i^f \cdot \hat{s}_{i\ell}$ for each fielder ℓ. Finally, we fit a regression model to adjust the fielding run values for the ballpark in which ball in play i occurred,

$$\delta_i^f \cdot \hat{s}_{i\ell} = \boldsymbol{D}_i \cdot \boldsymbol{\phi} + \tau_{i\ell} \tag{9}$$

where the covariate vector $\boldsymbol{D}_i$ contains a set of indicator variables for the specific ballpark for plate appearance i. The coefficient vector ϕ contains the effects of each ballpark which is estimated across all balls in play. The estimated residuals of this model,

$$RAA_{i\ell}^{field} = \hat{\tau}_{i\ell} = \delta_i^f \cdot \hat{s}_{i\ell} - \boldsymbol{D}_i \cdot \hat{\boldsymbol{\phi}} \tag{10}$$

represent the run value above average for fielder ℓ on ball in play i.

---

5 Here we interpret "making an out" as successfully converting a ball in play into at least one out.

3.6 Pitching run values

In Section 3.4 above, we allocated run value δ_i^p to the pitcher for plate appearance i. We need to adjust these run values to account for ballpark and platoon advantage, since both factors affect our expectation of pitching performance. We fit the following regression model,

$$\delta_i^p = \boldsymbol{B}_i \cdot \psi + \xi_i, \tag{11}$$

where the covariate vector $\boldsymbol{B}_i$ contains a set of indicator variables for the specific ballpark for plate appearance i and an indicator variable for whether or not the batter has a platoon advantage over the pitcher (same as in equation 1). The coefficient vector ψ contains the effects of each ballpark and the effect of a platoon advantage on the pitching run values. We estimate the coefficients $\hat{\psi}$ using the pitching run values for all plate appearances i. The estimated residuals of this model,

$$\text{RAA}_i^{pitch} = \hat{\xi}_i = \delta_i^p - \boldsymbol{B}_i \cdot \hat{\psi} \tag{12}$$

represent the run value above average for the pitcher on plate appearance i.

3.7 Tabulating runs above average

As outlined in Sections 2–3.6, we can calculate the run value for the hitter (RAA_i^{hit}), the run values for each baserunner (RAA_{ij}^{br}), the run values for each fielder ($\text{RAA}_{i\ell}^{field}$) and the run value for the pitcher (RAA_i^{pitch}) in each plate appearance i.

The overall run value for a particular player q is calculated by tabulating these run values across all plate appearances involving that player as a hitter, pitcher, baserunner or fielder,

$$
\begin{aligned}
RAA_q = &\sum_i \text{RAA}_i^{hit} \cdot \text{I}\,(\,\text{hitter} = q\,) + \\
&\sum_j \sum_i \text{RAA}_{ij}^{br} \cdot \text{I}\,(\,\text{runner } j = q\,) + \\
&\sum_\ell \sum_i \text{RAA}_{i\ell}^{field} \cdot \text{I}\,(\,\text{fielder } \ell = q\,) + \\
&\sum_i \text{RAA}_i^{pitch} \cdot \text{I}\,(\,\text{pitcher} = q\,)
\end{aligned}
$$

We present a logical summary of our WAR calculation in Figure 3.

3.8 Replacement level

As noted in our introduction, it is desirable to calibrate our comprehensive measure of player performance relative to a baseline "replacement level" player. However, the definition of a replacement level player remains controversial. The procedure used by both the *fWAR* and *rWAR* implementations is to set replacement level "at 1000 WAR per 2430 Major League games, which is the number of wins available in a 162 game season played by 30 teams. Or, an easier way to put it is that our new replacement level is now equal to a 0.294 winning percentage, which works out to 47.7 wins over a full season" (MacAree 2013). This definition is *ad hoc*, with the primary motivation for the definition seems to be the use of a convenient round number. In contrast, we derive a natural definition for replacement level from first principles.

The purpose of the replacement-level player is the need to *replace* a full-time major league player. There are only so many major league players, and all other players who participate in major league baseball are necessarily replacement players. Since there are 30 major league teams, each of which carries 25 players on its active roster
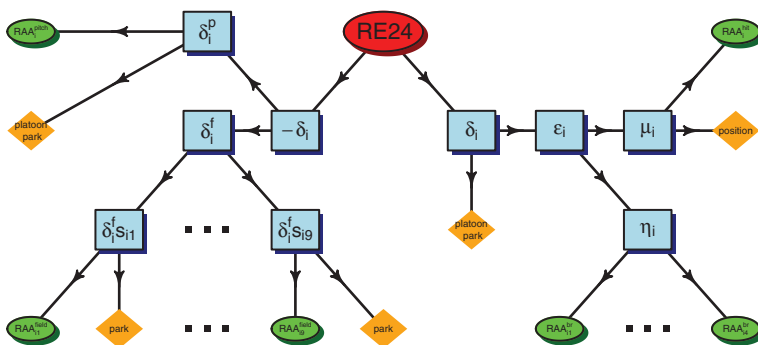

Figure 3: Schematic diagram of *openWAR*: The red ellipse labeled RE24 represents the estimated change in run value of a plate appearance. This value is then split into many parts and atrributed to the appropriate source. The diamonds represent fractions of RE24 that are not attributable to the player, whereas the ellipses on the outside correspond the four components of *openWAR* (hitting, baserunning, pitching, and fielding) that are attributable to a player.

during the season,[6] there are exactly 750 active major league players on any given day. We use this natural limitation to demarcate the set of major league players, and deem all others to be replacement-level players. Since most teams carry 13 position players and 12 pitchers, we designate the 30·13–390 position players with the most plate appearances and the 30·12=360 pitchers with the most batters faced as major league players. We submit that this naturally-motivated definition of replacement level is preferable to the *ad hoc* definition currently in use.

We can associate a replacement-level *shadow* with an actual player by multiplying the average performance across all replacement-level players by the number of events for that actual player. The WAR accumulated by each player's replacement-level shadow provides a meaningful baseline for comparison that is specific to that player. Using the convention that approximately 10 wins are equivalent to one win,[7] our *openWAR* value is computed as

$$WAR_q = \frac{RAA_q - RAA_q^{repl}}{10},$$

where RAA_q^{repl} is the runs above average figure for player q's replacement-level shadow.

4 Sources of variability

Existing implementations of WAR discuss uncertainty vaguely or not at all. We can delineate three sources of variability in the WAR values for each player in a given season: model estimation variability, situational variability, and player outcome variability. Model estimation variability comes from the errors that are made in estimating the parameters of our models for batting, pitching, fielding and baserunning in Section 3 as well as the expected runs model in Section 2. These models are trained on up to hundreds of thousands of observations and so this source of variability is small relative to the player outcome variability described below.

Situational variability comes from the differences in game situations across occurrences of the same batting event. For example, some home runs are hit when the bases are loaded whereas other home runs are hit when the bases are empty. These two situations have very

different run consequences despite the fact that they are driven by the same batting event (a home run). In traditional *WAR* calculations, a linear weights estimator is used for the batting component that assigns a run value to players based on aggregate batting statistics (such as *wOBA*) regardless of the game state. In other words, all home runs are given the same value in traditional *WAR* implementations, which introduces error into a player's *WAR* value in the sense that an equal weighting of all home runs is a less accurate description of *what actually happened*. [See Wyers (2013) for a discussion of quantifying situational error associated with *WARP*.] In contrast, our *openWAR* system is not subject to this type of error as we compute *WAR* based on each plate appearance rather than using aggregate statistics, which is a key distinction from the three previous implementations of WAR (*fWAR*, *rWAR* and *WARP*).

Player outcome variability is the uncertainty inherent in the outcomes of all events involving a particular player for a particular season. Imagine a particular player with a fixed ability, but repeat the same season for that player many times. In each of these seasons, the events involving that player would have variation in their outcomes, which would aggregate to a different *WAR* value for that particular player. The random variation in individual events dominates the variability in a player's *WAR* value, which is why our variance estimation targets this source of uncertainty.

Specifically, we estimate player outcome variability using a resampling strategy. In a particular season, we resample (with replacement) the *RAA* values for *individual plate appearances*, and re-aggregate them into a new *WAR* value for each player. A single resampling (a theoretical simulated season) will result in a second set of point estimates for the *WAR* of each player for which the models have not changed but the number of different individual events (e.g., the number of home runs hit by a player) could have changed. By performing this resampling procedure many times, we quantify the outcome variability for each player while preserving any inherent correlation within the individual events.[8] Although we have discussed uncertainty specifically for *WAR*, we believe that the above delineation of variability sources is generalizable to most aggregate measures of player performance across sports.

---

6 In September, active rosters may expand to as many as 40 players, although in practice, few teams carry more than 35.

7 Justification for this convention is provided in our supplementary materials.

8 We could use a similar resampling strategy to evaluate the model estimation variability as well, by re-fitting all of our *openWAR* models on each resampled season. However, the computational burden for re-fitting each model on each resampled season is very high and hard to justify given the relatively small size of the model estimation variability compared to the player outcome variability.

5 Results

In the 2012 MLB season, 534 of the $N=1284$ players were designated as replacement-level. *openWAR* was distributed approximately normally among these replacement-level players, with a mean of 0.01 and a standard deviation of 0.41. Conversely, the distribution of *openWAR* across all players was skewed heavily to the right, reflecting the disproportionate amount of *openWAR* accumulated by relatively few players. While the median *openWAR* was close to zero (0.36), the mean was a bit larger (0.91). *openWAR* values for all players fell between –2.6 and 8.6 wins above replacement, giving a range of 10 wins between the best player (Mike Trout) and the worst (Nick Blackburn). In Figure 4, we depict *openWAR* values for 2012, illustrating each player's replacement-level shadow and differentiating the major league players from the replacement-level players. There are $2N$ dots in Figure 4: N non-gray dots representing the RAA values for actual players and N gray dots representing the RAA values for the replacement-level shadows of those players.

We note that the variability associated with player performance is not constant. Figure 5 shows density estimates for the distribution of *openWAR* values under the resampling scheme described in Section 4 for three prominent

players: Miguel Cabrera, Robinson Cano, and Mike Trout. Trout's point estimate for WAR is higher than that of Cabrera or Cano, but the 95% confidence interval for his true *openWAR* is narrower, which suggests that Trout's performance is more consistent on a play-by-play basis than the others. Table 1 shows various quantiles of the distribution of *openWAR* for the top 20 performers in 2012.

Figure 6 depicts the width of 95% confidence intervals for *openWAR* based on resampling all plays that occurred in the 2012 season. As expected, the width of the confidence interval for a particular player widens as that player is exposed to more playing time. In general, the confidence intervals for pitchers tend to be smaller than those for position players with comparable playing time. This may suggest that pitchers perform more consistently across plate appearances, or merely reflect the fact that the replacement level for pitchers is higher (closer to 0) than it is for position players.

As noted in the introduction, WAR was at the core of the debate about the 2012 American League MVP Award. Miguel Cabrera of the Detroit Tigers had become the first player since 1967 to win the Triple Crown, leading the AL in the conventional statistics of batting average, home runs, and runs batted in. However, sabermetricians advocated strongly for Mike Trout, a rookie centerfielder
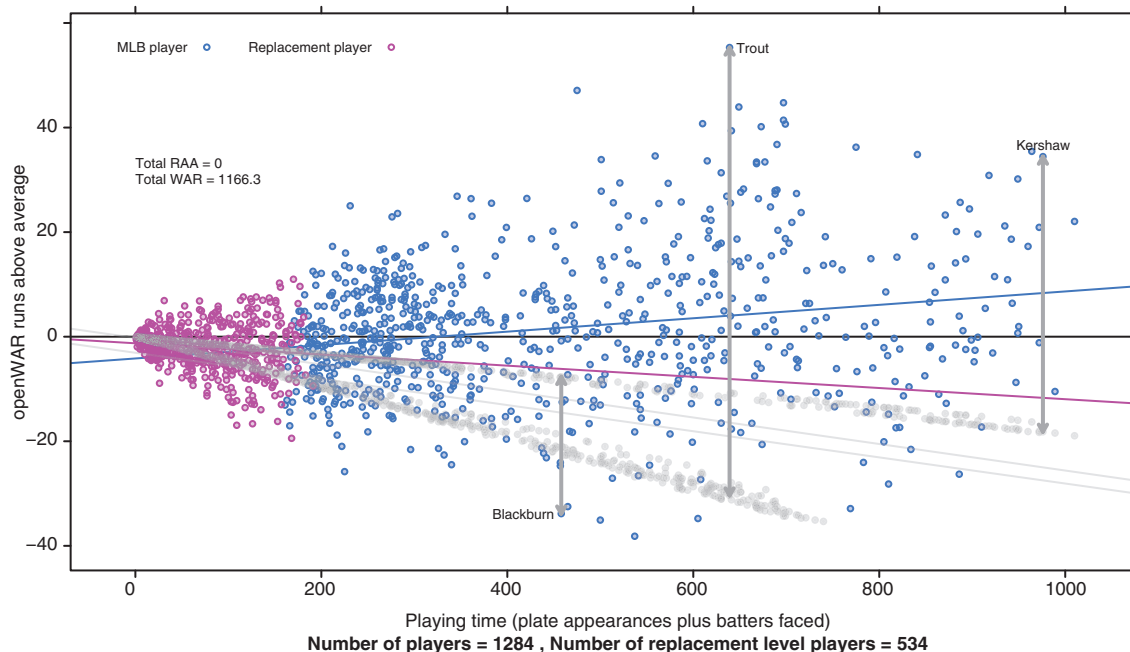

Figure 4: *openWAR* RAA values for the 2012 MLB season. Each blue dot is a major league player, while each pink dot is a replacement-level player. For each player, we also plot a gray dot that represents their replacement-level shadow with the same amount of playing time. For three specific players, we show the vertical distance between their *RAA* and the *RAA* for their replacement-level shadow. Playing time is calculated as "plate appearances + batters faced" to provide an equivalent scale for both pitchers and batters: playing time for pitchers is the number of batters faced, whereas playing time for batters is the number of plate appearances. For pitchers, we also add any plate appearances they had as a batter.
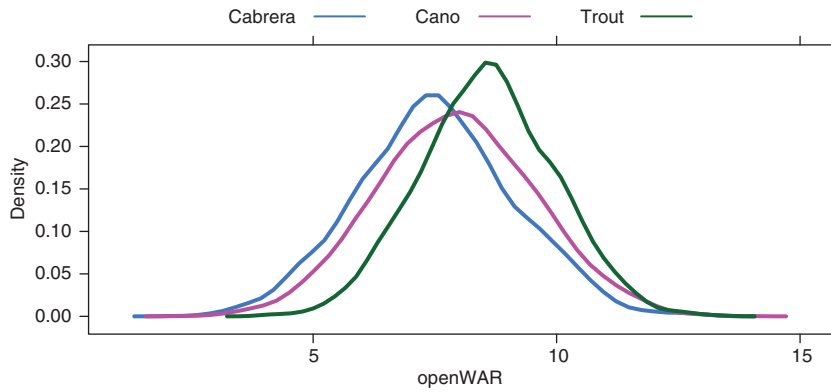
Figure 5: *openWAR* density estimates for Miguel Cabrera (blue), Robinson Cano (pink), and Mike Trout (green). Note that while Trout's density curve is further to the right, it is narrower than the others.

Table 1: Distribution of *openWAR* for 2012. Quantiles reported are based on 3500 simulated seasons.

Name	q0	q2.5	q25	q50	q75	q97.5	q100
Mike Trout	3.52	5.81	7.58	8.53	9.48	11.27	13.91
Robinson Cano	2.56	4.88	6.85	7.92	8.96	11.11	13.74
Chase Headley	1.97	4.47	6.42	7.47	8.50	10.53	13.12
Miguel Cabrera	1.99	4.31	6.43	7.46	8.49	10.49	12.84
Edwin Encarnacion	2.67	4.54	6.36	7.32	8.29	10.29	13.17
Andrew McCutchen	2.02	4.29	6.18	7.19	8.17	10.17	12.07
Joey Votto	2.82	4.76	6.21	7.00	7.77	9.34	10.80
Prince Fielder	2.57	4.16	5.98	6.96	7.90	9.89	12.18
Joe Mauer	2.51	4.32	5.89	6.78	7.64	9.27	11.01
Buster Posey	2.49	4.08	5.79	6.73	7.62	9.46	11.98
Aaron Hill	1.29	3.72	5.64	6.62	7.59	9.52	12.67
Ryan Braun	1.97	3.68	5.55	6.60	7.62	9.56	11.44
Ben Zobrist	1.67	3.94	5.52	6.44	7.33	9.06	11.65
Josh Willingham	0.83	3.33	5.25	6.29	7.27	9.44	11.75
Martin Prado	1.59	3.69	5.27	6.16	7.05	8.65	10.97
Aramis Ramirez	0.52	3.19	5.18	6.15	7.09	9.05	11.73
Elvis Andrus	0.83	3.52	5.25	6.14	7.03	8.89	11.10
Matt Holliday	0.54	3.22	5.07	6.09	7.09	9.02	11.20
Adrian Gonzalez	1.40	3.04	4.91	5.93	6.96	8.84	10.89
David Wright	1.20	3.22	4.95	5.88	6.81	8.64	10.63

who excelled in all aspects of the game. While it was acknowledged on both sides that Cabrera was likely the better hitter, sabermetricians argued that Trout's superior skill at baserunning and fielding more than made up for Cabrera's relatively small edge in batting. In fact, for adherents of sabermetrics, the decision was clear – point estimates showed Trout leading Cabrera by 3.2 *fWAR* and 3.6 *rWAR*.

Our *openWAR* values provide a more sophisticated perspective on this debate. Trout's point estimate for *openWAR* in 2012 is 1.05 wins larger than Cabrera's, but it is important to note that their interval estimates overlap considerably. In Figure 7, the joint distribution

of *openWAR* values for Cabrera and Trout's 2012 seasons are plotted. In nearly 32% of those simulated seasons, Cabrera's *openWAR* was higher than Trout's. Thus, our results suggest that there is a high probability that Trout had a better season than Cabrera, but there is substantial uncertainty in their comparison. This exercise illustrates two strengths of *openWAR*: 1) distinctions made through point estimates tend to accord with those made via the existing implementations (note that Cabrera was not even the second-best player in any implementation); and 2) the interval estimates provided by *openWAR* allow for more nuanced conclusions to be drawn.

Table 2 shows the top ten best and worst baserunners, according to *openWAR* in 2012. We note many true positives (Mike Trout and Desmond Jennings are considered to be excellent baserunners, while Paul Konerko, David Ortiz and Adrian Gonzalez are plodding) with no eyebrow-raising surprises.

Similarly, Table 3 shows the top ten best and worst fielders according to *openWAR* in 2012. Here again we see some true positives (Brandon Crawford, Darwin Barney, and Adam Jones are reputedly excellent fielders) but also some head-scratchers (Prince Fielder is anecdotally considered a poor fielder). We also note that the magnitudes of the fielding numbers reported by *openWAR* are smaller than those reported by UZR. This may be a result of the fact that *openWAR* currently only measures some defensive skills, or it could reflect weaknesses in the unknown models underlying UZR, which merits further study.

Results for *openWAR* in the 2013 seasons were similar to those of 2012, with an observed range of –2.0 to 10.7. Mike Trout was again the best player, and Clayton Kershaw was again the best pitcher (6.5 *openWAR*). Figure 8 shows the full *openWAR* results for all players in 2013, and quantiles for simulated *openWAR* for 2013 are presented in Table 4.
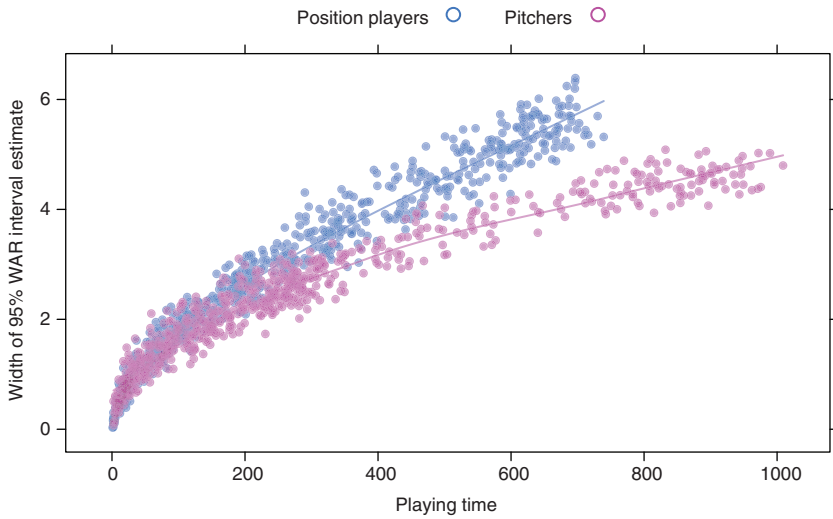
Figure 6: Spread of 95% Confidence Intervals for *openWAR* based on resampling all plays that occurred in the 2012 season. Intervals are narrower for pitchers compared to position players and variation in the length of CIs exists among players.
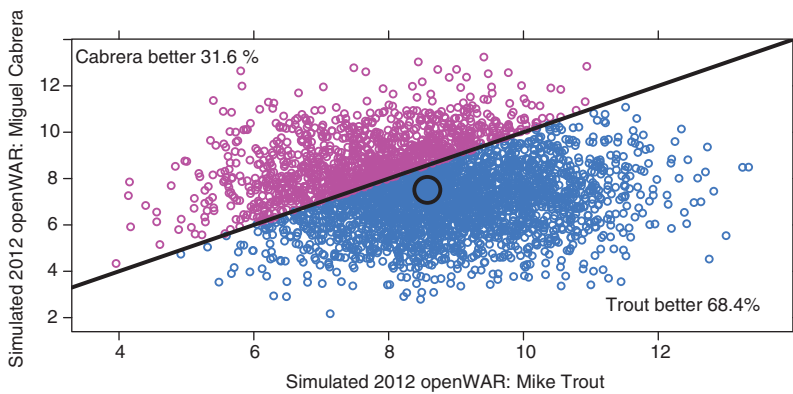

Figure 7: Joint distribution of *openWAR* for Mike Trout vs. Miguel Cabrera, 2012. We note that in about 68% of 3500 simulated seasons, Trout produced a higher WAR than Cabrera.

Table 2: 2012 baserunning RAA leaders.

Best	RAA	Worst	RAA
Mike Trout	14.79	Paul Konerko	−9.28
Martin Prado	9.03	David Ortiz	−8.43
Desmond Jennings	8.84	Jamey Carroll	−8.27
Jarrod Dyson	8.62	Michael Young	−8.07
Evereth Cabrera	8.62	Todd Helton	−7.08
Drew Stubbs	7.86	Prince Fielder	−6.82
Jason Heyward	7.75	Adrian Beltre	−6.34
Darwin Barney	7.67	Justin Morneau	−6.29
Torii Hunter	7.64	Adrian Gonzalez	−6.26
Dustin Ackley	7.51	Howie Kendrick	−5.90

Table 3: 2012 fielding RAA leaders.

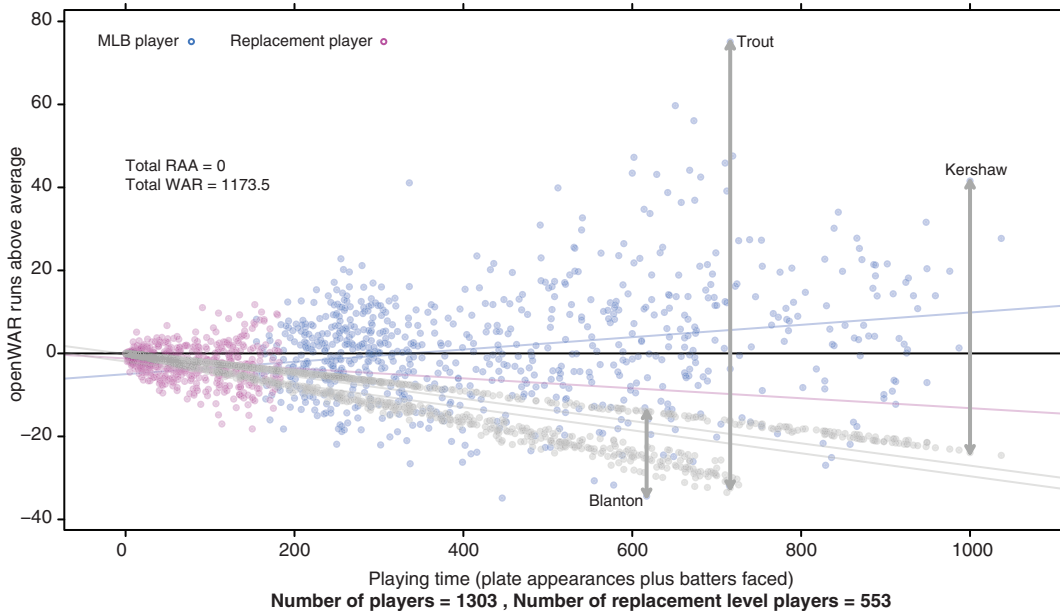| Best	RAA	Worst	RAA
Jason Heyward	10.17	Colby Rasmus	−10.44
Brandon Crawford	9.91	Jose Altuve	−9.03
Yunel Escober	9.19	Tyler Greene	−8.30
Ben Zobrist	8.23	Brian Dozier	−7.82
Darwin Barney	8.05	Lucas Duda	−7.82
Prince Fielder	7.66	Shin-Soo Choo	−7.77
Adrian Gonzalez	7.43	Orlando Cespedes	−7.49
Alejandro De Aza	7.13	Justin Smoak	−7.34
Adam Jones	7.05	Garrett Jones	−6.83
Craig Gentry	6.72	Rickie Weeks	−6.64

Figure 8: *openWAR* RAA values for the 2013 MLB season. Each blue dot is a major league player, while each pink dot is a replacement-level player. For each player, we also plot a gray dot that represents their replacement-level shadow with the same amount of playing time. Playing time is calculated just as in Figure 4. Mike Trout and Clayton Kershaw were the best position player and pitcher, respectively, while Joe Blanton was the worst player.

Table 4: Distribution of *openWAR* for 2013. Quantiles reported are based on 3500 simulated seasons.

Name	q0	q2.5	q25	q50	q75	q97.5	q100
Mike Trout	5.48	7.60	9.58	10.60	11.60	13.58	15.39
Miguel Cabrera	2.74	5.70	7.62	8.71	9.79	11.83	14.74
Chris Davis	3.14	5.39	7.41	8.53	9.64	11.82	13.87
Matt Carpenter	2.99	5.00	6.77	7.67	8.56	10.34	12.52
Paul Goldschmidt	1.47	4.39	6.46	7.64	8.78	10.93	13.81
Josh Donaldson	1.76	4.35	6.22	7.21	8.17	10.15	12.73
Matt Holliday	2.07	4.34	6.16	7.07	7.97	9.89	12.58
Shin-Soo Choo	2.63	4.33	6.14	7.06	7.98	9.89	12.02
Freddie Freeman	1.13	4.28	6.03	7.05	8.04	9.93	11.59
Robinson Cano	1.23	4.07	5.94	6.98	8.01	10.03	12.10
Andrew McCutchen	1.75	4.00	5.74	6.71	7.70	9.53	11.80
David Ortiz	1.61	3.87	5.60	6.63	7.63	9.59	12.08
Clayton Kershaw	2.12	4.31	5.77	6.54	7.31	8.79	10.60
Carlos Santana	2.35	3.89	5.54	6.42	7.30	8.89	11.04
Jason Kipnis	1.68	3.62	5.35	6.29	7.23	9.04	11.23
Ian Kinsler	1.17	3.33	5.06	5.92	6.79	8.47	10.76
Edwin Encarnacion	1.03	3.12	4.94	5.91	6.84	8.90	11.71
Joey Votto	1.43	3.31	4.96	5.91	6.82	8.63	10.44
Troy Tulowitzki	1.15	3.33	5.02	5.88	6.75	8.47	10.04
Cliff Lee	1.41	3.15	4.60	5.39	6.18	7.70	9.48

5.1 Comparison to previous WAR implementations

Our *openWAR* point estimates are similar to existing implementations of WAR, though as noted above, we also provide uncertainty estimates. In Table 5, we list the top 10 performance in *openWAR* alongside those of *fWAR*. There is considerable (though not universal) agreement with respect to these players and the magnitudes of the WAR values are similar. Comparison to *rWAR* yields similar results.

We can examine the overall correlation between previous WAR implementations and our openWAR point estimates in Table 6. *openWAR* correlates highly with both

Table 5: 2012 WAR Leaders, *fWAR* (left) and *openWAR* (right).

Name	fWAR	Name	openWAR
Mike Trout	10.0	Mike Trout	8.57
Robinson Cano	7.7	Robinson Cano	7.91
Buster Posey	7.7	Miguel Cabrera	7.52
Ryan Braun	7.6	Chase Headley	7.50
David Wright	7.4	Edwin Encarnacion	7.28
Chase Headley	7.2	Andrew McCutchen	7.24
Miguel Cabrera	6.8	Joey Votto	6.96
Andrew McCutchen	6.8	Prince Fielder	6.92
Jason Heyward	6.4	Joe Mauer	6.73
Adrian Beltre	6.3	Buster Posey	6.71

Table 6: Correlation matrix between *openWAR*, *fWAR*, and *rWAR*.

	rWAR	fWAR	openWAR
rWAR	1	0.918	0.881
fWAR	0.918	1	0.875

fWAR and *rWAR*, although not as highly as they correlate with each other.

We also examined the consistency of *openWAR* from season-to-season by calculating the autocorrelation within players between their 2012 and 2013 seasons. As seen in Table 7, the within-player autocorrelation of our *openWAR* values are similar to those of *fWAR* and *rWAR*.

As illustrated in Figure 4, the sum of all *RAA* values in 2012 is exactly 0, and the sum of all *openWAR* values is 1166. Whereas the former figure is guaranteed based on the way we have defined runs *above average*, the latter is sensitive to changes in the definition of replacement-level. However, as noted in Section 3.8, replacement-level is defined in both *fWAR* and *rWAR* so that the sum of all WARs is 1000. In order to compare the magnitudes of *openWAR* to *fWAR* and *rWAR* directly, we can generate more comparable values by increasing the number of replacement-level players. This in turn raises the performance of the replacement-level shadows, and lowers the amount of WAR in the system (see Figure 9). Given the *ad hoc* nature of the previous definition of replacement-level,

we prefer our definition. Moreover, the fact that the total unnormalized *openWAR* was nearly identical in 2012 and 2013 (1166 and 1173), suggest that there may be some intrinsic meaning to this number. Additionally, the total RAA values for *rWAR* did *not* add up to 0 in either 2012 or 2013 – a logical weakness in that system.

6 Summary and further discussion

The concept of WAR has been one of the great success stories in the long history of sabermetrics, and sports analytics in general. However, there are major limitations in previous methodology both in terms of calculating WAR, and in the public's understanding of what WAR values mean. Chiefly, the previous implementations of WAR are not reproducible and do not contain uncertainty estimates. This leads to the unpleasant situation where journalists are forced to take WAR estimates on faith, with no understanding of the accuracy (or construction) of those estimates. In this paper, we have addressed the issues of reproducibility and uncertainty estimation by providing a fully open source, statistical model for Wins Above Replacement based on our *conservation of runs* framework with uncertainty in our model-based WAR values estimated by resampling methods.

There remain several limitations of *openWAR* that offer the opportunity for further research. The first limitation
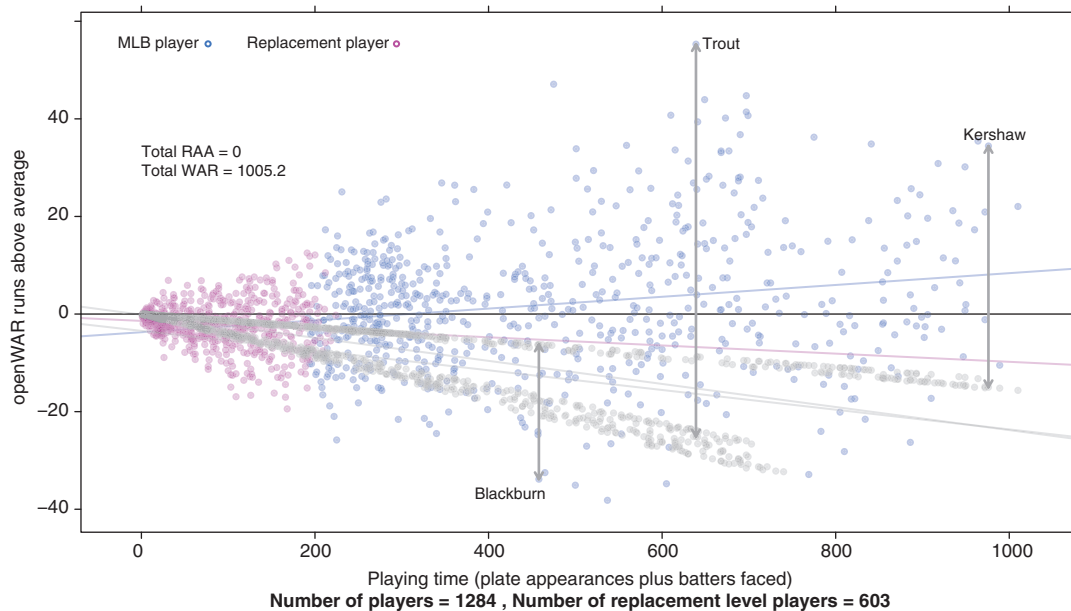
Table 7: Autocorrelation of WAR implementations. Each player's WAR in 2012 and 2013 was calculated, and the correlation between the matched pairs is shown.

	rWAR	*fWAR*	*openWAR*
Autocorrelation	0.522	0.596	0.571

Figure 9: *openWAR* RAA values for 2012, normalized so that the total WAR is about 1000. Compared to Figure 4, here the definition of replacement-level is more inclusive. Playing time is calculated just as in Figure 4.

is data quality. Although the fidelity of the MLBAM data is very high, it is not perfect. There were instances in the data where players were listed with the wrong ID and for most balls in play, there was not a description of whether that ball was hit on the ground or in the air. Furthermore, there is no indication of how long each batted ball took to get to the specified location, making both the *trajectory* and *speed* of each batted ball unknown.

The accounting of baserunner movement for non-batting events like stolen bases, caught stealings, wild pitches, and errors merits further work. All baserunner movement is captured, but it is modeled implicitly. Our approach only takes into account the actual baserunner movement but is indifferent to the various mechanisms by which a baserunner advanced. For example, a runner on first who steals second base and advances to third on a single is rewarded the same amount as a baserunner who advances to second on a wild pitch and then advances to third on a single. The same holds for a runner who simply advances directly from first to third on a single.

The defensive models used in *openWAR* are somewhat rudimentary and could certainly be improved with more resolute data. Since there is no record in the data of where each fielder was standing at the beginning of the play, there is no way to distinguish between fielder range versus fielder positioning. This drawback is also true in most current fielding measures, such as UZR and SAFE. Some fielding measures such as UZR add additional components for throwing and the ability to turn a double play, which we hope to add to *openWAR* in future work.

Another interesting idea would be a conservation of wins framework for *openWAR* rather than the conservation of runs. Rather than assigning the value of each plate appearance based on the change in expected runs, the value of a plate appearance could alternatively be based on the change in win probability from the beginning to the end of a plate appearance. The *openWAR* framework could then be altered to take changes in *win probability* as inputs rather than the change in the expected run matrix. One rationale for using win probability is that we may not wish to treat each run scored as contributing equally to a win. For example, extra runs when a team is winning by a large margin are not as valuable as an extra run when the teams are tied.

We suspect that using a framework based on the change in win probability will give similar results in terms of magnitude and ranking of players, since every day hitters will get plate appearances in many different game situations. However, certain types of players (closers, relief specialists, pinch hitters, defensive replacements, pinch runners) may only make appearances in games

in specific situations such that the runs that they create or prevent may be systematically more (or less) valuable than the approximately 10% of a win that is assigned to each run now. It would be particularly interesting to look at relief pitchers as they are often only in a game because of the specific game situation (the game is close and near the end of the game), which would make the runs they prevent more valuable than most runs created over the course of the season.

References

Acharya, R. A., A. J. Ahmed, A. N. D'Amour, H. Lu, C. N. Morris, B. D. Oglevee, A. W. Peterson, and R. N. Swift. 2008. "Improving Major League Baseball Park Factor Estimates." *Journal of Quantitative Analysis in Sports* 4.

Albert, J. and J. Bennett. 2003. *Curve Ball: Baseball, Statistics, and the Role of Chance in the Game.* New York, NY: Copernicus Books.

Appelman, D. 2008. "Get to Know: RE24." http://www.fangraphs.com/blogs/get-to-know-re24/. Accessed on March 6, 2015.

Axisa, M. 2013. "Topps will Feature WAR on the Back of their Baseball Cards Soon." http://www.cbssports.com/mlb/eye-on-baseball/22903637/topps-will-feature-war-on-the-back-of-their-baseball-cards-soon. Accessed on March 6, 2015.

Baseball Prospectus Staff 2013. "WARP." http://www.baseballprospectus.com/glossary/index.php?search=warp. Accessed on March 6, 2015.

Baumer, B. and A. Zimbalist. 2014. *The Sabermetric Revolution: Assessing the Growth of Analytics in Baseball.* Philadelphia, PA: University of Pennsylvania Press.

Bowman, B. 2013. "GameDay." http://mlb.mlb.com/mlb/gameday. Accessed on March 6, 2015.

Buckheit, J. B. and D. L. Donoho. 1995. "Wavelab and Reproducible Research," Tech. Rep. 474, Stanford University, http://statistics.stanford.edu/ckirby/techreports/NSF/EFS%20NSF%20474.pdf. Accessed on March 6, 2015.

Bukiet, B., E. Harold, and J. Palacios. 1997. "A Markov Chain Approach to Baseball." *Operations Research* 45:14–23.

Claerbout, J. 1994. "Hypertext Documents about Reproducible Research." Tech. rep., Stanford University, http://sepwww.stanford.edu. Accessed on March 6, 2015.

Fangraphs Staff 2013. "DRS." http://www.fangraphs.com/library/defense/drs/. Accessed on March 6, 2015.

Forman, S. 2010. "Player Wins Above Replacement." http://www.baseball-reference.com/blog/archives/6063. Accessed on March 6, 2015.

Forman, S. 2013. "Position Player WAR Calculations and Details." http://www.baseball-reference.com/about/war_explained_position.shtml. Accessed on March 6, 2015.

Freeze, R. 1974. "An Analysis of Baseball Batting Order by Monte Carlo Simulation." *Operations Research* 22:728–735.

Furtado, J. 1999. "Introducing XR." http://www.baseballthinkfactory.org/btf/scholars/furtado/articles/IntroducingXR.htm. Accessed on March 6, 2015.

Ioannidis, J. P. 2013. "This I Believe in Genetics: Discovery can be a Nuisance, Replication is Science, Implementation Matters." *Frontiers in Genetics* 4:1–2.

Jacques, D. 2007. "Value Over Replacement Player." http://www.baseballprospectus.com/article.php?articleid=6231. Accessed on March 6, 2015.

James, B. and J. Henzler. 2002. *Win shares*, Chicago, IL: STATS Pub.

James, B. 1986. *The Bill James Historical Baseball Abstract*. New York, NY: Random House Inc.

Jensen, S. T., K. E. Shirley, and A. J. Wyner. 2009. "Bayesball: A Bayesian Hierarchical Model for Evaluating Fielding in Major League Baseball." *The Annals of Applied Statistics* 3:491–520.

Johnson, G. 2014. "New Truths that Only One Can See." *The New York Times*, http://www.nytimes.com/2014/01/21/science/new-truths-that-only-one-can-see.html. Accessed on March 6, 2015.

Knuth, D. E. 1984. "Literate Programming." *The Computer Journal* 27:97–111.

Kubatko, J., D. Oliver, K. Pelton, and D. T. Rosenbaum. 2007. "A Starting Point for Analyzing Basketball Statistics." *Journal of Quantitative Analysis in Sports* 3:1–22.

Lewis, M. 2003. *Moneyball: The Art of Winning an Unfair Game*. New York, NY: WW Norton & Company.

Lichtman, M. 2010. "The Fangraphs UZR Primer." http://www.fangraphs.com/blogs/the-fangraphs-uzr-primer/. Accessed on March 6, 2015.

Lichtman, M. 2011. "Ultimate Base Running Primer." http://www.fangraphs.com/blogs/ultimate-base-running-primer/. Accessed on March 6, 2015.

Lindsey, G. 1959. "Statistical Data Useful for the Operation of a Baseball Team." *Operations Research* 7:197–207.

Lindsey, G. 1961. "The Progress of the Score During a Baseball Game." *Journal of the American Statistical Association* 56:703–728.

MacAree, G. 2013. "Replacement Level." http://www.fangraphs.com/library/misc/war/replacement-level/. Accessed on March 6, 2015.

Macdonald, B. 2011. "A Regression-Based Adjusted Plus-Minus Statistic for NHL Players." *Journal of Quantitative Analysis in Sports* 7:4.

McCracken, V. 2001. "Pitching and Defense: How Much Control Do Hurlers Have?" http://baseballprospectus.com/article.php?articleid=878. Accessed on March 6, 2015.

Naik, G. 2011. "Scientists' Elusive goal: Reproducing Study Results." *The Wall Street Journal* 258, A1, http://online.wsj.com/news/articles/SB10001424052970203764804577059841672541590. Accessed on March 6, 2015.

Nature Editorial. 2013. "Announcement: Reducing our irreproducibility," *Nature* 496. http://www.nature.com/news/announcement-reducing-our-irreproducibility-1.12852. Accessed on March 6, 2015.

Pankin, M. 1978. "Evaluating Offensive Performance in Baseball." *Operations Research* 26:610–619.

Rickert, J. 2013. "Nate Silver Addresses Assembled Statisticians at this Year's JSM." http://blog.revolutionanalytics.com/2013/08/nate-silver-jsm.html. Accessed on March 6, 2015.

Rosenberg, M. 2012. "The case for Miguel Cabrera over Mike Trout for AL MVP." http://sportsillustrated.cnn.com/2012/writers/michael_rosenberg/11/15/miguel-cabrara-mvp/index.html. Accessed on March 6, 2015.

Schoenfield, D. 2012. "What we Talk about when we Talk about WAR." http://espn.go.com/blog/sweetspot/post/_/id/27050/what-we-talk-about-when-we-talk-about-war. Accessed on March 6, 2015.

Schwarz, A. 2005. *The Numbers Game: Baseball's Lifelong Fascination With Statistics*. Thomas Dunne Books.

Slowinski, S. 2010. "What is WAR?" http://www.fangraphs.com/library/index.php/misc/war/. Accessed on March 6, 2015.

Sokol, J. 2003. "A Robust Heuristic for Batting Order Optimization Under Uncertainty." *Journal of Heuristics* 9:353–370.

Studeman, D. 2005. "I'm Batty for Baseball Stats." *The Hardball Times*, May 10, 2005, http://www.fangraphs.com/library/pitching/xfip/. Accessed on March 6, 2015.

Tango, T. 2003. "Defensive Responsibility Spectrum (DRS)." http://www.tangotiger.net/drspectrum.html. Accessed on March 6, 2015.

Tango, T., M. Lichtman, and A. Dolphin. 2007. *The Book: Playing the Percentages in Baseball*. Washington, DC: Potomac Books.

The Economist Editorial. 2013. "Trouble at the Lab. (Cover story)." http://www.economist.com/node/21588057/. Accessed on March 6, 2015.

Thorn, J. and P. Palmer. 1984. *The Hidden Game of Baseball: A Revolutionary Approach to Baseball and its Statistics*. Doubleday, Garden City, NY.

Wand, M. 1994. "Fast Computation of Multivariate Kernel Estimators." *Journal of Computational and Graphical Statistics* 3:433–445.

Wyers, C. 2013. "Reworking WARP." http://www.baseballprospectus.com/article.php?articleid=21586. Accessed on March 6, 2015.

Xie, Y. 2014. *Dynamic Documents with R and knitr*. Chapman & Hall/CRC.

Zimmer, C. 2012. "A Sharp Rise in Retractions Prompts Calls for Reform." *The New York Times*, http://www.nytimes.com/2012/04/17/science/rise-in-scientific-journal-retractions-prompts-calls-for-reform.html. Accessed on March 6, 2015.