David Michael Vock* and Laura Frances Boehm Vock

Estimating the effect of plate discipline using a causal inference framework: an application of the G-computation algorithm

<https://doi.org/10.1515/jqas-2016-0029>

Abstract: Offensive performance in baseball depends on a number of correlated factors: the pitches the batter faces, the batter's choice to swing, and the batter's hitting ability. Recently a renewed focus on the effect of plate discipline on batter performance has emerged. Plate discipline has traditionally been summarized as the proportion of pitches inside and outside of the strike zone a player swings at; however, there have been few metrics proposed to assess the effect of plate discipline directly on batters' outcomes. In this paper, we focus on estimating a batter's performance if he were able to adopt a different plate discipline. Because we wish to assess the effect of a counterfactual plate discipline, we use a potential outcome framework and show how the G-computation algorithm can be used to isolate the effect of plate discipline separately from a batter's hitting ability or the types of pitches the batter faces. As an example, we implement our approach using data collected with the PITCHf/x system over the 2012–2014 seasons to identify the improvement Starlin Castro would expect to see in offensive performance were he able to adopt Andrew McCutchen's plate discipline. We estimate that had Castro adopted McCutchen's discipline his batting average, on-base percentage, and slugging percentage would have increased 0.017 (se = 0.004), 0.040 (se = 0.006), and 0.028 (se = 0.008), respectively.

Keywords: baseball; causal inference; G-computation algorithm; plate discipline; thin plate regression splines.

1 Introduction

Understanding the value of each batter's output in baseball is an important component to analyzing and comparing players. There is no shortage of metrics used to compare players' batting performance, including traditional

statistics such as batting average, on-base percentage, and slugging percentage as well as more refined measures such as weighted runs created (James, 1986), isolated power, weighted on-base percentage (Tango, Lichtman, and Dolphin, 2007), value over replacement player (Miller and Wojciechowski, 2015), and wins above replacement player (Baumer, Jensen, and Matthews, 2015) (although the latter metrics also contain information on fielding and pitching).

There has been substantial effort to understand the factors related to differences in batting production among players to allow for more nuanced comparisons. For example, Berry, Reese, and Larkey (1999), Schell (2005), Albert (2006), and Kvam (2011), among others, examine the effect of era on hitting statistics to directly compare players from different historical times. Several authors including Schmotzer, Switchenko, and Kilgo (2008) and Nieswiadomy, Strazicich, and Clayton (2012) examine the effect of steroids on batting production. Additionally, many researchers attempt to quantify the effect of various factors unrelated to individual batting ability which affect batting statistics. For example, Acharya et al. (2008) quantify park effects on batting. Kaplan (2006) quantifies the variation in batting due to team and individual effects, and Phillips (2011) assesses the effect of the following hitter in the batting order on batter production.

Another factor related to differences in batting production among players which has received renewed focus is plate discipline (Carleton, 2007; Paine, 2014; Lentzer, 2011). We define plate discipline as the likelihood of a player to swing at pitches with different characteristics in different scenarios. Although we will later define plate discipline precisely, a player with poor plate discipline swings indiscriminately at pitches regardless of location or characteristics while a player with good discipline is more selective. Typically, the plate discipline of a particular player has been summarized as the proportion of pitches inside and outside the strike zone that a player swings at. Unfortunately, these metrics only consider the binary condition of the pitch, inside or outside the strike zone. Although it seems obvious that players should swing at pitches in the middle of the strike zone and take pitches far outside the strike zone, it is less clear what players should do on more borderline pitches. Importantly, to our knowledge no analysis has attempted to quantify how

*Corresponding author: David Michael Vock, Division of Biostatistics, University of Minnesota Twin Cities, 420 Delaware St SE, MMC 303, Minneapolis, MN 55455, USA, e-mail: vock@umn.edu

Laura Frances Boehm Vock: Department of Mathematics, Computer Science, and Statistics, Gustavus Adolphus College, Saint Peter, MN, USA, e-mail: lboehmvo@gustavus.edu

different plate disciplines affect the expected outcomes of at-bats.

We are motivated by trying to assess the effect of Starlin Castro's proclivity to swing at pitches outside of the strike zone on his offensive production between 2012 and 2014. After a promising 2011 season with a slash line of 0.307/0.341/0.432 (batting average/on-base percentage/slugging percentage), his performance declined over the next two seasons to a 2013 slash line of 0.245/0.284/0.347. Many blame his declining performance on his poor plate discipline, specifically, his tendency to swing at breaking pitches down and outside the strike zone and fastballs up and inside the strike zone (Figure 1). In contrast, Andrew McCutchen celebrated a breakout year in 2013 and was named the National League most valuable player. Both players are right-handed batters, listed at the same height and weight, and have similar experience in the major leagues, so it is natural to compare them. An important question is, if Castro had the same plate discipline and faced a similar set of pitches as McCutchen, how would his offensive production have been different in 2013? Or, put another way, to what extent is the difference in batting outcomes between McCutchen and Castro due to differences in their choices to swing at certain pitches, due to differences in the pitches that they face, and due to differences in their hitting ability?

Because we wish to estimate the effect of various counterfactual plate disciplines, we adopt a potential outcomes framework (Rubin, 1974; Holland, 1986; Rubin, 1990; Pearl, 2009). The advantage of this approach is that we can be precise about the assumptions and models needed to estimate the causal effect of swinging/not swinging at a pitch (i.e. plate discipline). Understanding the effect of the choice to swing or not swing at a pitch separately from the effects of batting ability and the pitcher's choice of what type of pitch to throw on the overall outcome of a plate appearance (e.g. walk, single, strikeout, etc.) is challenging because standard regression-based approaches do not consistently estimate the causal effect of these choices or actions.

For example, we could model the result of each pitch (e.g. whether or not the player got a hit, the pitch was called a ball, etc.) given the pitch's location and other pitch and game characteristics (e.g. speed, location, break, count, inning, etc.) to assess the effect of swinging or not swinging at pitches at various locations. However, there are two main problems with that analysis. First, it is unclear what information the immediate outcome of a pitch, such as a foul ball or whiff, gives us about the effect of a player's plate discipline on the ultimate outcome of the plate appearance. We could disregard pitches that do not end the at-bat, but this approach may leave out important

information about the effect of plate discipline early on in the at-bat. Second, adjusting for pitch and game characteristics in the regression model does not lead to causal estimates of plate discipline because including these "pitch-varying" confounders in the model may "block" the causal pathway under investigation. For example, poor plate discipline leads to unfavorable counts for the batter, but this effect of poor plate discipline is "adjusted" away in the analysis. However, exclusion of those factors may lead to bias due to lack of confounding adjustment.

The challenge encountered in modeling the causal effect of plate discipline on the outcome of plate appearances is similar to the situation encountered in many biomedical applications when researchers are interested in determining the effect of time-varying treatments, actions, or interventions (what type of pitch to throw and whether or not to swing in this application) on long-term outcomes. In particular, Robins and colleagues have identified three general techniques for identifying the causal effect of time-varying interventions in the presence of time-varying confounding: G-computation (Robins, 1986), G-estimation of structural nested mean models (Robins and Greenland, 1994; Almirall, Ten Have, and Murphy, 2010), and inverse probability weighting of marginal structural models (Robins, Hernan, and Brumback, 2000). To date, G-computation to assess the causal effect of treatments has been infrequently used outside the biomedical/epidemiological literature (see Taubman et al., 2009; Young et al., 2011; Westreich et al., 2012, for examples with biomedical applications).

The purpose of this article is to outline a framework for comparing the effect of plate discipline between players and illustrate the use of the parametric G-computation algorithm to assess causal effects. In Section 2, we outline the notation used throughout, define the causal quantities of interest and the assumptions needed to identify those quantities, and describe G-computation. Section 3 describes the models used as part of the parametric G-computation algorithm and the data used to estimate those models. Section 4 gives the results of the comparison of the effect of plate discipline between Starlin Castro and Andrew McCutchen. We conclude in Section 5.

2 Methodological framework

2.1 Observed data

To estimate the effect of pitch selection and plate discipline on outcomes of plate appearances, we observe a sample of n plate appearances. We let O_i be the eventual random outcome of the i th plate appearance (e.g.

Figure 1: Observed probability of swinging at four-seam fastballs (A) and sliders (B) from right-handed pitchers for Starlin Castro and Andrew McCutchen during the 2012–2014 seasons. The probability of swinging has been smoothed using a generalized additive model with thin plate regression splines. The perspective here is from the batter/catcher and the green box indicates the strike zone.

walk, single, fly out, etc.). For each pitch of the plate appearance, two decisions/actions must be made. First, the pitcher must decide what type of pitch to throw (by type of pitch, we mean the speed, break, movement, location, and other characteristics of the pitch). Second the batter must decide whether or not to swing. Let D_{ij} be the vector of pitch characteristics of the j th pitch and S_{ij} the indicator for whether or not the batter swung at the j th pitch of the i th plate appearance. In this baseball application, the pitch number (j) is analogous to time and the decision of the pitch type (D_{ij}) and whether or not to swing (S_{ij}) are analogous to treatment decisions in biomedical applications assessing the effect of time-varying treatments on outcomes.

Additionally, let O_{ij} be the immediate outcome of the j th pitch (e.g. ball, called strike, single, etc.). We define $O_{ij} = S_{ij} = D_{ij} =$ “plate appearance complete” with probability 1 if the at-bat has ended before the j th pitch (this ensures that O_{ij} is well-defined for all $j = 1, 2, \dots$). We note that $I(O_i = \text{hit}) = \sum_j I(O_{ij} = \text{hit})$, where the indicator function $I(x) = 1$ if the condition x is true, and similarly for other outcomes of a plate appearance. Other variables specific to the at-bat, including game situation characteristics or the arsenal of the particular pitcher are denoted by U_i . Finally, let A_i denote the batter for the i th plate appearance. A directed acyclic graph describing the causal relationship between these variables is given in Figure 2.

Figure 2: Directed acyclic graph (DAG) denoting causal pathways for batter A with pitch characteristics D_j and swing choice S_j (the “treatments”) and intermediate outcomes O_j (“time-varying confounders”) of pitch j . Game situation variables (such as current score or outs) are indicated by U . In this DAG, the unit being indexed is a single plate appearance.

2.2 Potential outcomes and dynamic treatment regimes

Conceptually, we could think about what would have been the result of a particular plate appearance had a batter, contrary to fact, chosen to swing at certain pitches and the pitcher, contrary to fact, chosen certain pitch types.

Let $O_j^{(\bar{d}_j, \bar{s}_j)}$ be the immediate outcome of the j th pitch of a random plate appearance if, possibly contrary to fact, the characteristics of the j pitches were $d_1, d_2, \dots, d_j$ and the indicator for swinging at the j pitches were $s_1, s_2, \dots, s_j$. $O_j^{(\bar{d}_j, \bar{s}_j)}$ is referred to as a potential or counterfactual outcome as we might not observe a random plate appearance with that sequence of pitch types and swinging through the j th pitch. Throughout, we will use the superscript notation to distinguish between different potential outcomes and use the overbar notation to denote history so that $\bar{d}_j = (d_1, d_2, \dots, d_j)$.

This notation assumes that there is no need to index the counterfactual outcomes of a plate appearance by the actions of other plate appearances and that the outcome does not depend on how the action is selected. This is known as the consistency assumption (Robins, 1997), which we discuss further in Section 2.6. This ensures that the random potential outcomes for the j th pitch can be written unambiguously as $O_j^{(\bar{d}_j, \bar{s}_j)}$ and that $O_{ij}^{(\bar{d}_j, \bar{s}_j)} = O_{ij}$ if $\bar{d}_{ij} = \bar{d}_j$ and $\bar{s}_{ij} = \bar{s}_j$.

A (nonrandom) dynamic treatment regime is a sequence of functions $(g_1, g_2, \dots)$ for each decision point

which maps from the covariate and action/treatment history available at the decision point to an action to take. In this application, a (nonrandom) dynamic regime is a sequence of functions for the pitch type decision and swing decisions which, for the j th pitch, map from the intermediate outcomes of the plate appearance and the history of prior actions to the pitch type to throw and whether or not to swing. One example of a (nonrandom) dynamic swing regime would be to take (i.e. do not swing at) all pitches until there are two strikes, then swing at all pitches in the strike zone.

In some applications, it is unrealistic to consider what would happen if treatment, action, or intervention decisions were a deterministic function of prior covariate and treatment/action history. For example, Cain et al. (2010) considered dynamic treatment regimes in which patients with HIV initiate treatment after their CD4 count falls below a certain threshold. However, the authors acknowledge that it would be unrealistic to consider the possibility that all subjects initiate treatment immediately, and a more clinically relevant regime would allow for stochastic uptake of treatment. Similarly, nonrandom dynamic treatment regimes do not specifically allow us to answer how the typical outcomes of a plate appearance would change if a player were to adopt a different tendency to swing at pitches because whether or not a player swings at a pitch is never a deterministic function.

Cain et al. (2010) and Murphy, van der Laan, and Robins (2001) introduce the idea of a random dynamic treatment regime. A random dynamic treatment regime

is a sequence of conditional probability distributions at each decision point which gives the probability of taking a certain treatment or action given past covariate history and prior actions taken. For example, in this application, we consider a stochastic regime, i.e. a sequence of conditional probability distributions, for pitch characteristics $\bar{P}_D^{(g)} = (P_D^{(g_1)}, P_D^{(g_2)}, \dots)$ and swing choices $\bar{P}_S^{(g)} = (P_S^{(g_1)}, P_S^{(g_2)}, \dots)$ where, for example,

$$P_S^{(g_j)} [S_j = s_j \mid \bar{D}_j = \bar{d}_j, \bar{S}_{j-1} = \bar{s}_{j-1}, \bar{O}_{j-1} = \bar{o}_{j-1}, \\ U = u, A = a]$$

gives the probability of swinging at the j th pitch under this stochastic regime. For short, the probability of a plate appearance resulting in a particular outcome, say a hit, for a batter, a_H , given he follows the stochastic regime $\bar{P}_D^{(g)}$ and $\bar{P}_S^{(g)}$, that is $P [O_i^{(\bar{P}_D^{(g)}, \bar{P}_S^{(g)})} = \text{hit} \mid A_i = a_H]$, can be expressed as a weighted average across the distribution of potential outcomes $O_i^{(\bar{d}, \bar{s})}$ for all possible $\bar{d}$ and $\bar{s}$. More precisely,

$$\begin{aligned} & P [O_i^{(\bar{P}_D^{(g)}, \bar{P}_S^{(g)})} = \text{hit} \mid A_i = a_H] \\ &= \sum_j P [O_{ij}^{(\bar{P}_D^{(g)}, \bar{P}_S^{(g)})} = \text{hit} \mid A_i = a_H] \\ &= \sum_j \int \dots \int \left\{ P [U_i = u \mid A_i = a_H] \right. \\ &\quad \times \prod_{k=1}^j P_D^{(g_k)} [D_{ik} = d_k \mid \bar{D}_{ik-1} = \bar{d}_{k-1}, \bar{S}_{ik-1} = \bar{s}_{k-1}, \\ &\quad \quad \bar{O}_{ik-1}^{(\bar{d}_{k-1}, \bar{s}_{k-1})} = \bar{o}_{k-1}, U_i = u, A_i = a_H] \\ &\quad \times \prod_{k=1}^j P_S^{(g_k)} [S_{ik} = s_k \mid \bar{D}_{ik} = \bar{d}_k, \bar{S}_{ik-1} = \bar{s}_{k-1}, \\ &\quad \quad \bar{O}_{ik-1}^{(\bar{d}_{k-1}, \bar{s}_{k-1})} = \bar{o}_{k-1}, U_i = u, A_i = a_H] \\ &\quad \times \prod_{k=1}^{j-1} P [O_{ik}^{(\bar{d}_k, \bar{s}_k)} = o_k \mid \bar{O}_{ik-1}^{(\bar{d}_{k-1}, \bar{s}_{k-1})} = \bar{o}_{k-1}, \\ &\quad \quad U_i = u, A_i = a_H] \\ &\quad \times P [O_{ij}^{(\bar{d}_j, \bar{s}_j)} = \text{hit} \mid \bar{O}_{ij-1}^{(\bar{d}_{j-1}, \bar{s}_{j-1})} = \bar{o}_{j-1}, U_i = u, \\ &\quad \quad A_i = a_H] \left. \right\} do_1 \dots do_{j-1} ds_1 \dots ds_j dd_1 \dots dd_j du, \end{aligned} \quad (1)$$

where the first equality follows from the law of total probability. We note that we use $P(\cdot)$ to denote either a probability mass function or probability density function (and if $P(X)$ is a probability mass function then the integral with respect to X is simply the sum over all values in the support of the random variable X). To be explicit, $P [O_{ij}^{(\bar{d}_j, \bar{s}_j)} = \text{hit} \mid \bar{O}_{ij-1}^{(\bar{d}_{j-1}, \bar{s}_{j-1})} = \bar{o}_{j-1}, U_i = u, A_i = a] = 0$ if $O_{ik}^{(\bar{d}_k, \bar{s}_k)}$ is a terminal outcome of the plate appearance for any $k < j$.

2.3 Adopting the pitch selection and plate discipline of another batter

We can now precisely define, in terms of random dynamic treatment regimes, what we mean by adopting the plate discipline of another batter or adopting the pitch selection as if another batter were in the batter's box. In particular, we are interested in estimating the distribution of the outcome of the plate appearance under the stochastic regime

$$\begin{aligned} & P_S^{(g_j)} [S_j \mid \bar{d}_j, \bar{s}_{j-1}, \bar{o}_{j-1}, u, a_H] \\ &= P [S_j = s_j \mid \bar{D}_j = \bar{d}_j, \bar{S}_{j-1} = \bar{s}_{j-1}, \bar{O}_{j-1} = \bar{o}_{j-1}, \\ &\quad U_i = u, A_i = a_S] \end{aligned} \quad (2)$$

and

$$\begin{aligned} & P_D^{(g_j)} [d_j \mid \bar{d}_{j-1}, \bar{s}_{j-1}, \bar{o}_{j-1}, u, a_H] \\ &= P [D_j = d_j \mid \bar{D}_{j-1} = \bar{d}_{j-1}, \bar{S}_{j-1} = \bar{s}_{j-1}, \bar{O}_{j-1} = \bar{o}_{j-1}, \\ &\quad U_i = u, A_i = a_D] , \end{aligned} \quad (3)$$

where a_S and a_D need not equal a_H . That is, we examine the effect of random dynamic regimes for swinging and pitch type for player a_H , which equal the observed distribution of swinging and pitch types of player a_S and a_D respectively. We refer to random dynamic regimes in which Equation 2 holds as “adopting” the plate discipline of player a_S and similarly for Equation 3, “adopting” the pitch selection, or distribution of typical pitch characteristics, thrown to batter a_D . For notational brevity, we write $O_i^{(\bar{P}_D^{(g)}, \bar{P}_S^{(g)})}$ in which Equations 2 and 3 hold as $O_i^{(a_D, a_S)}$. That is, $O_i^{(a_D, a_S)}$ indicates the potential outcome of plate appearance i under the random dynamic regime in which a batter sees the pitch selection typical of player a_D and adopts the plate discipline of player a_S . In the following, any remaining differences in the distribution of outcomes for different players we attribute to what we call hitting

ability though this also includes strike zone/framing and park effects.

2.4 Functionals of the distribution of potential outcomes

The goal of this analysis is to compare the distribution of potential outcomes of the plate appearance under different combinations of pitch selection (a_D) and plate discipline (a_S) for different batters (a_H). That is, we wish to estimate the conditional distribution of $O_i^{(a_D, a_S)}$ given $A_i = a_H$ for different combinations of a_D , a_S , and a_H . Once we know (or estimate) the distribution of the potential outcomes under a particular combination of a_D and a_S for a given batter a_H , we can then compute common batting statistics such as batting average, on-base percentage, or slugging. For example, the counterfactual batting average is the conditional probability that the ultimate, counterfactual outcome $O_i^{(a_D, a_S)}$ is a hit given that the plate appearance was an at-bat. That is, the batting average under a particular combination of a_D and a_S for batter a_H is given by

$$BA^{(a_D, a_S, a_H)} = \frac{P \left[O_i^{(a_D, a_S)} = \text{hit} \mid A_i = a_H \right]}{P \left[O_i^{(a_D, a_S)} = \text{hit} \mid A_i = a_H \right] + P \left[O_i^{(a_D, a_S)} = \text{out} \mid A_i = a_H \right]}. \quad (4)$$

Similarly, on-base percentage ($OBP^{(a_D, a_S, a_H)}$) and slugging percentage ($SLG^{(a_D, a_S, a_H)}$) may be defined in terms of the probability of counterfactual outcomes:

$$OBP^{(a_D, a_S, a_H)} = \frac{P \left[O_i^{(a_D, a_S)} = \text{hit} \mid A_i = a_H \right] + P \left[O_i^{(a_D, a_S)} = \text{walk} \mid A_i = a_H \right]}{P \left[O_i^{(a_D, a_S)} = \text{hit} \mid A_i = a_H \right] + P \left[O_i^{(a_D, a_S)} = \text{out} \mid A_i = a_H \right] + P \left[O_i^{(a_D, a_S)} = \text{walk} \mid A_i = a_H \right]} \quad (5)$$

$$SLG^{(a_D, a_S, a_H)} = \frac{1 \times P \left[O_i^{(a_D, a_S)} = 1B \mid A_i = a_H \right] + 2 \times P \left[O_i^{(a_D, a_S)} = 2B \mid A_i = a_H \right] + 3 \times P \left[O_i^{(a_D, a_S)} = 3B \mid A_i = a_H \right] + 4 \times P \left[O_i^{(a_D, a_S)} = \text{HR} \mid A_i = a_H \right]}{P \left[O_i^{(a_D, a_S)} = \text{hit} \mid A_i = a_H \right] + P \left[O_i^{(a_D, a_S)} = \text{out} \mid A_i = a_H \right]} \\ = \frac{E \left[TB(O_i^{(a_D, a_S)}) \mid O_i^{(a_D, a_S)} = \text{in-play}, A_i = a_H \right] P \left[O_i^{(a_D, a_S)} = \text{in-play} \mid A_i = a_H \right]}{P \left[O_i^{(a_D, a_S)} = \text{hit} \mid A_i = a_H \right] + P \left[O_i^{(a_D, a_S)} = \text{out} \mid A_i = a_H \right]}, \quad (6)$$

where the preceding definition of on-base percentage has ignored plate appearances that ended in a sacrifice or hit-by-pitch and $TB(O_i^{(a_D, a_S)})$ is the total bases of the at-bat. We show that the equality in Equation 6 holds in the Supplementary Material. These statistics could then be compared to the batting average, on-base percentage, and slugging percentage that we would have observed had the pitch selection and plate discipline been a'_D and a'_S , respectively.

2.5 G-computation algorithm

Equation 1 suggests that rather than estimate the distribution of $O_i^{(a_D, a_S)}$ directly, we can estimate the distribution of $O_i^{(a_D, a_S)}$ by estimating the conditional distributions of the counterfactual, intermediate outcomes of the at-bat given the past events of the at-bat. The idea behind the G-computation algorithm is that, under appropriate assumptions, the conditional distribution for the counterfactual outcome of the pitch (i.e. $O_{ij}^{(\bar{a}_j, \bar{s}_j)}$) given the prior counterfactual events of the plate appearance, game situation variables, and the batter (i.e. $O_{ij-1}^{(\bar{a}_{j-1}, \bar{s}_{j-1})}$, U_i , and A_i) is equal to the conditional distribution of the observed outcome (i.e. O_{ij}) conditioned on the observed intermediate events and actions (i.e. $\bar{D}_{ij}$, $\bar{S}_{ij}$, $\bar{O}_{ij-1}$, U_i , A_i). In mathematical notation,

$$P \left[O_{ij}^{(\bar{a}_j, \bar{s}_j)} \mid \bar{O}_{ij-1}^{(\bar{a}_{j-1}, \bar{s}_{j-1})}, U_i, A_i \right] = P \left(O_{ij} \mid \bar{D}_{ij} = \bar{a}_j, \bar{S}_{ij} = \bar{s}_j, \bar{O}_{ij-1}, U_i, A_i \right). \quad (7)$$

This says that all the conditional distributions given in Equation 1 can be replaced with the conditional distributions given in the right-hand side of Equations 2, 3,

and 7 which can be estimated from observed data. Thus, provided the assumptions in Section 2.6 hold, we can use observed data to estimate the distribution of (unobserved) potential outcomes $O_i^{(a_D, a_S)}$.

More intuitively, the main idea of the G-computation algorithm is that we can model the distribution of a long-term outcome (e.g. the outcome of the plate appearance) under different intervention strategies (e.g. probability of

swinging) by modeling the effect of the intervention and prior outcomes on the intermediate outcomes. That is, we can estimate the causal effect of a time-varying intervention strategy by modeling its effect on the entire longitudinal process. In our context we develop models for the effect of swinging or not swinging and pitch characteristics and prior history of the at-bat on the outcomes of each pitch of the plate appearance. Several articles give detailed tutorials on this algorithm (Ahern, Hubbard, and Galea, 2009; Young et al., 2011; Westreich et al., 2012).

2.6 Causal assumptions

To obtain valid inference using the G-computation algorithm, we must make the following identifying assumptions. Briefly, these assumptions require the intervention to be precisely defined and that we have collected all (possibly time-varying) confounders. First, we make the so-called consistency assumption that observed random variables equal the potential outcomes for the “treatments” received, i.e. $O_{ij}^{(\bar{d}_j, \bar{s}_j)} = O_{ij}$ if $\bar{D}_j = \bar{d}_j$ and $\bar{S}_j = \bar{s}_j$. The consistency assumption and its implication are described at length in Cole and Frangakis (2009). In this application, the assumption requires, for example, that the way in which a batter swings (e.g. bat speed) would not change if forced to swing at a certain pitch as compared to their usual approach.

Second, we assume $O_{ij}^{(\bar{d}_j, \bar{s}_j)} \perp D_{ij} \mid \bar{O}_{ij-1}, \bar{S}_{ij-1}, \bar{D}_{ij-1}, U_i, A_i$ and $O_{ij}^{(\bar{d}_j, \bar{s}_j)} \perp S_{ij} \mid \bar{O}_{ij-1}, \bar{S}_{ij-1}, \bar{D}_{ij}, U_i, A_i$ where $\perp$ indicates independence. This set of assumptions is known in the causal inference literature as the sequential ignorability assumption. Whether or not this assumption, which is unidentifiable, is reasonable depends on the amount and quality of the data collected as part of D_{ij} and O_{ij} which we discuss in Section 3. To evaluate if this assumption is reasonable, it is helpful to consider a scenario in which the sequential ignorability assumption is not tenable. For example, suppose that we did not collect movement, break, and location information and D_{ij} only contained information about the pitch speed. Batters tend to swing at pitches near the middle of the plate and have better outcomes on pitches thrown there. Therefore, knowing that the batter actually swung (S_{ij}) at a pitch would give us information on the distribution of $O_{ij}^{(\bar{d}_j, \bar{s}_j)}$ regardless of the value of s_j . Because we have collected information about movement, break, and location as well as speed, we feel that knowing whether or not the batter actually swung at a pitch would give no further information on the counterfactual outcomes; and, thus, the sequential ignorability assumption is satisfied.

Additionally, we assume that there is no interference between plate appearances by the same batter. That is, the intervention (i.e. whether to swing or not and what pitch type to throw) from one plate appearance does not impact the outcome of another plate appearance. Although intervening on plate discipline might have an effect on the types of pitches that a batter sees over the long-term (as scouting reports adjust to new information on a batter), we fix the distribution of types of pitches that a batter sees in our analysis to avoid violating the no interference assumption. There may still be some interference (for example, a better plate discipline early in the game may result in improved outcomes which could result in facing a different pitcher for a later plate appearance). However, we believe such effects are (vanishingly) small.

Under these assumptions we can show that the equality given in Equation 7 holds. For example, it follows that the conditional distribution of $O_{ij}^{(\bar{d}_j, \bar{s}_j)}$ is given by

$$\begin{aligned} P \left[O_{ij} = o \mid \bar{D}_{ij} = \bar{d}_j, \bar{S}_{ij} = \bar{s}_j, \bar{O}_{ij-1} = \bar{o}_{ij-1}, \right. \\ \left. U_i = u, A_i = a_H \right] \\ = P \left[O_{ij}^{(\bar{d}_j, \bar{s}_j)} = o \mid \bar{D}_{ij} = \bar{d}_j, \bar{S}_{ij} = \bar{s}_j, \bar{O}_{ij-1}^{(\bar{d}_{j-1}, \bar{s}_{j-1})} \right. \\ \left. = \bar{o}_{ij-1}, U_i = u, A_i = a_H \right] \\ = P \left[O_{ij}^{(\bar{d}_j, \bar{s}_j)} = o \mid \bar{O}_{ij-1}^{(\bar{d}_{j-1}, \bar{s}_{j-1})} = \bar{o}_{j-1}, \right. \\ \left. U_i = u, A_i = a_H \right] \end{aligned}$$

where the first equality follows from the consistency assumption and the second follows from the sequential ignorability assumption.

3 Statistical models to compare Starlin Castro and Andrew McCutchen

To illustrate our methodological approach, we compare the batting average, on-base percentage, and slugging percentage under different counterfactual combinations of pitch selection, plate discipline, and batting ability during the 2012–2014 seasons, where a_D , a_S , and a_H are either Starlin Castro or Andrew McCutchen. Because the movement of pitches for right-handed pitchers is opposite that of left-handed pitchers, the analysis presented here only covers the effect of pitch selection, plate discipline, and batting ability when facing right-handed pitchers. Equations 4–6 indicate that to estimate the counterfactual batting average, on-base percentage,

and slugging percentage we must estimate $P(O_i^{(a_D, a_S)} = o | A_i)$, where $o = \text{hit, out, in-play, or walk as well as}$ $E[TB(O_i^{(a_D, a_S)}) | O_i^{(a_D, a_S)} = \text{in-play}, A_i]$.

To estimate those probabilities of counterfactual outcomes, as we discussed in Section 2.5, we must develop models for $P(D_{ij} | \bar{D}_{ij-1}, \bar{S}_{ij-1}, \bar{O}_{ij-1}, U_i, A_i)$, $P(S_{ij} | \bar{D}_{ij}, \bar{S}_{ij-1}, \bar{O}_{ij-1}, U_i, A_i)$ and $P(O_{ij} | \bar{D}_{ij}, \bar{S}_{ij}, \bar{O}_{ij-1}, U_i, A_i)$. We describe our modeling approach for each of these conditional distributions of the observed data in turn. We note that we estimate each of these conditional distributions separately for each value of A_i (i.e. for each batter, Starlin Castro and Andrew McCutchen). That is, the effect of each covariate in the models was allowed to interact with A_i .

3.1 Available data

Since 2006, each pitch in Major League Baseball (MLB) has been tracked using the PITCHf/x system. The system records the velocity, movement, break, spin, release point, pitch location, and outcome of each pitch. We downloaded the PITCHf/x data from the 2012–2014 regular seasons for Starlin Castro and Andrew McCutchen against right-handed pitching using the `pitchRx` package in R (Siev-ert, 2014a,b). We do not utilize data sources (e.g. MLB Advanced Media) which describe the location and trajectory of all batted balls. Nonetheless, we believe we have enough data to justify the assumptions in Section 2.6.

Plate appearances in which the location and movement of the pitch were not recorded were eliminated from the dataset. We also removed any plate appearances resulting in a sacrifice bunt, sacrifice fly, or batter hit by the pitch because those are rare outcomes which are difficult to model. This resulted in the removal of 7 plate appearances for Castro and 11 for McCutchen. Over the three seasons, this yields 1463 plate appearances for Castro and 1560 for McCutchen against right-handed pitching, comprised of 5290 and 6193 pitches, respectively.

Code to reproduce the analytic dataset and all the analysis presented here are available as a GitHub repository (<https://github.com/docvock>).

3.2 Model for pitch characteristics

As is common in baseball analytics, we make the assumption that the distribution of pitch characteristics D_{ij} , conditioned on the current ball strike count and the arsenal of the pitcher, is independent of the full outcome and pitch characteristic history. By arsenal, we mean the set of pitch types that a pitcher (typically) throws. Conditioning on arsenal is important because there is

more heterogeneity in pitch characteristics across different pitchers than within a plate appearance with the same pitcher. That is, throwing a slider, for example, makes the next pitch more likely to be a slider because we know that a slider is part of the pitcher's arsenal. To define a pitcher's arsenal, we note that the pitch type classification used by PITCHf/x is fairly detailed (e.g. a cutter is a separate pitch type from a slider). If we used this granular level pitch type, the number of pitchers with each arsenal type would be small. Therefore, we categorized players based on whether or not they threw a fastball (all types), cutter/slider, curveball, and change-up. Unusual pitch types (e.g. knuckleball, eephus) were excluded from this analysis. To determine a pitcher's arsenal, we consider all pitches thrown versus the Cubs and Pirates during the 2012–2014 seasons. Because there is some misclassification in pitch types, we consider a pitch type to be a part of a pitcher's arsenal if it was observed in at least 5% of pitches, where this level was chosen for convenience. For example, the most common arsenal was the complete arsenal (fastball, cutter/slider, curveball, and changeup) and was observed for 33.3% and 31.2% of the at-bats for McCutchen and Castro in the dataset. We denote the categorical variable indicating pitcher arsenal as R_i in the remainder of the manuscript to distinguish this specific variable from the potential vector of variables U_i . With these assumptions, then $P(D_{ij} | \bar{D}_{ij-1}, \bar{S}_{ij-1}, \bar{O}_{ij-1}, U_i, A_i) = P(D_{ij} | C_{ij}, R_i, A_i)$, where C_{ij} is the observed ball-strike count on the j th pitch of the i th plate appearance. Rather than using a parametric model for this (multivariate) distribution, we use the empirical distribution of pitch characteristics for each batter, conditional on pitch count and arsenal.

3.3 Model for swinging

We model the probability of swinging at a pitch given past history of the plate appearance using a logistic generalized additive model (GAM) with covariates for location, movement, speed, and ball/strike count (a factor variable for all 12 possible combinations of balls and strikes). Location refers to the two dimensional point (distance from the ground and distance from center of plate) where the pitch crosses home plate. Movement is the two dimensional distance from where the pitch actually crosses the plate to where the same pitch with no spin would have crossed the plate (difference in horizontal and vertical directions). Speed refers to the speed when the pitch is released. Because we do not expect the effect of speed, horizontal and vertical location, and horizontal and vertical movement to be linearly related to the log odds of swinging

at a pitch, we use low rank thin plate regression splines to flexibly model the effect of the continuous covariates (Wood, 2003). Note that we use a single two-dimensional smoother for the location (vertical and horizontal) and movement (vertical and horizontal direction). Thin plate regression splines, unlike other smoothers, are isotropic. That is, rotation of the covariate coordinate system does not change the result of smoothing which is a realistic model for spatial coordinates such as location and movement. Thin plate regression splines are in a defined sense the optimal smoother of any given basis dimension/rank. Additional details on thin plate regression splines can be found in Wood (2006). A tensor product was used to allow the location and movement to interact. The tuning parameter for the smooth penalty was chosen using the generalized cross validation criterion. These models, and the outcome models in Section 3.4, were fit using the `mgcv` package in R (Wood, 2004, 2015).

Given that location and movement are not linearly related to the log odds and flexible basis expansions are needed, it is impractical to fit a model with the entire pitch and outcome history in the model for the j th pitch. To account for pitch sequencing, we considered for inclusion in this model whether or not the previous pitch was the same type (e.g. slider, four-seam fastball, etc.), the distance between where the previous pitch and the current one crossed the plate, and the difference in pitch speed between the previous and current pitches after adjusting for pitch location, break, and ball-strike count. However, we found that none of these terms were significant for either batter and were not retained in the final model. Thus, the final fitted model assumes $P(S_{ij} | \bar{D}_{ij}, \bar{S}_{ij-1}, \bar{O}_{ij-1}, U_i, A_i) = P(S_{ij} | D_{ij}, C_{ij}, A_i)$. To fit these models, we can pool the data across all pitches of a plate appearance and fit a single model rather than separate models for the first pitch, second pitch, etc.

3.4 Model for intermediate outcomes

As we do not collect information about the location/trajectory of batted balls, the intermediate outcomes, O_{ij} , are discrete and the possible values are ball, called strike, whiff (swinging strike), foul, in-play out, and hit (where we could further subdivide hits into 1B, 2B, 3B, or HR). The probability of each outcome is modeled using a series of conditional logistic GAMs illustrated in Figure 3. M1 is described in Section 3.3. Conditional on the fact the batter does not swing, M2 models the probability that the pitch is called a strike. A simplified approach would deterministically compare the location of the pitch to the strike zone; however, we found slight differences between players

in which borderline pitches were actually called strikes and that using a deterministic strike zone results in poor model calibration. Given the player swings, M3 models the probability the player makes contact. Similarly, given the batter makes contact, M4 models the probability the ball is in play, and given a ball is in play, M5 fits the probability it is a hit. The probability of any particular intermediate outcome can be easily found by multiplying the appropriate conditional probabilities. For example, the probability of a foul on the j th pitch of plate appearance i is

$$\begin{aligned} P(O_{ij} = \text{foul} | D_{ij}, C_{ij}, A_i) &= P(O_{ij} = \text{Foul} | \text{contact}_{ij} \\ &= 1, S_{ij} = 1, D_{ij}, C_{ij}, A_i) \times P(\text{contact}_{ij} = 1 | \\ &S_{ij} = 1, D_{ij}, C_{ij}, A_i) P(S_{ij=1} = 1 | D_{ij}, C_{ij}, A_i), \end{aligned}$$

where the conditional probabilities on the right hand side are found from model M4, M3, and M1.

All models (M2–M5) included location, movement, and speed as covariates using thin plate splines as a basis expansion. Location and movement were allowed to interact using a tensor product spline. Similar to M1 (swing) model, M2, M3, and M5 include the count as a factor variable (note: Castro swung at few pitches on a 3-0 count and whiffed on all of them, so for Castro, 3-0 counts are grouped with 2-0 counts for M3 and M5). Lastly, M4 (foul) for both batters includes an indicator of a 0-2 count. Covariates were chosen so that the model-based estimates of batting performance for McCutchen and Castro were closest to the observed values in Table 1.

To estimate the counterfactual slugging percentage, we need to estimate the expected number of total bases given the j th pitch resulted in a ball in-play and $\bar{D}_{ij}, \bar{S}_{ij}, \bar{O}_{ij-1}$. To do so, we use a generalized additive log-linear regression model for the expected number of bases which included the same factors for location, movement, and speed as the other models. Additionally, an indicator for 3-0 count was included in the model for McCutchen.

As with the model for the probability of swinging, we evaluated whether or not pitch sequencing was a significant predictor in each of these models by investigating the effect of whether or not the previous pitch was the same type, the distance between where the previous pitch and the current one crossed the plate, and the difference in pitch speed between the previous and current pitches. None of these terms were significant predictors for either batter for any of the models and, therefore, were not retained. Overall, these models assume that $P [O_{ij} | \bar{D}_{ij}, \bar{S}_{ij}, \bar{O}_{ij-1}, U_i, A_i] = P [O_{ij} | D_{ij}, S_{ij}, C_{ij}, A_i]$.

To visualize the effect these assumptions have on the causal pathway, we present a DAG in Figure 4.

Figure 3: Set of generalized additive logistic regression models (M1–M5) used to model outcomes O_{ij} and swinging S_{ij} .

Figure 4: Directed acyclic graph (DAG) denoting causal pathways for batter A with pitch characteristics D_j and swing choice S_j (the “treatments”) and intermediate outcomes O_j (“time-varying confounders”) of pitch j with count C_j facing a pitcher with arsenal R . We assume that, conditional on the count C_j , treatments and outcomes are independent of the full history $(\bar{D}_{j-1}, \bar{S}_{j-1}, \bar{O}_{j-1})$. Only pitch characteristics D_j depend upon the arsenal R .

3.5 Estimating counterfactual outcomes

As noted in Section 2.5, estimating the counterfactual probability of any outcome of a plate appearance (e.g. $P[O_i^{(a_D, a_S)} = \text{hit}]$) requires integrating over the distribution of intermediate outcomes as in Equation 1. In many applications of the G-computation algorithm, this would require evaluating an integral of very large dimension and usually necessitates some approximation (e.g. Monte Carlo integration). The assumption that the outcome of the j th pitch depends on the past pitches only through the count simplifies the dimension of the integral considerably. For example, we show in the Supplementary Material that Equation 1 simplifies to

$$\begin{aligned}
 & P \left[O_i^{(a_D, a_S)} = \text{hit} \mid A_i = a_H \right] \\
 &= \sum_j P \left[O_{ij}^{(a_D, a_S)} = \text{hit} \mid A_i = a_H \right] \\
 &= \sum_j \int \int \int \int P[R_i = r \mid A_i = a] \\
 &\quad \times P \left[C_{ij}^{(a_D, a_S)} = c_j \mid R_i = r, A_i = a \right] \\
 &\quad \times P_D^{(g_j)} \left[D_{ij} = d_j \mid C_{ij}^{(a_D, a_S)} = c_j, R_i = r, A_i = a_H \right] \\
 &\quad \times P_S^{(g_j)} \left[S_{ij} = s_j \mid D_{ij} = d_j, C_{ij}^{(a_D, a_S)} = c_j, A_i = a_H \right] \\
 &\quad \times P \left[O_{ij}^{(d_j, s_j)} = \text{hit} \mid C_{ij}^{(a_D, a_S)} = c_j, A_i = a_H \right] \\
 &\quad dr \, dc_j \, dd_j \, ds_j.
 \end{aligned} \tag{8}$$

The probability distribution for the counterfactual count of the j th pitch, $P[C_{ij}^{(a_D, a_S)} | R_i = r, A_i = a_H]$, can be defined recursively based on the probability of the count of the previous pitch and the probability of the outcome of the previous pitch. For example, $P[C_{i1}^{(a_D, a_S)} = 0-0 | R_i = r, A_i = a_H] = 1$ and then the probability of having count b -s on pitch j is given by:

$$\begin{aligned} P[C_{ij}^{(a_D, a_S)} = b-s | R_i = r, A_i = a_H] \\ = P[O_{ij-1}^{(a_D, a_S)} = \text{ball} | C_{ij-1}^{(a_D, a_S)} = (b-1)-s, R_i = r, \\ A_i = a_H] P[C_{ij-1}^{(a_D, a_S)} = (b-1)-s | R_i = r, A_i = a_H] \\ + \left\{ P[O_{ij-1}^{(a_D, a_S)} \in \{\text{whiff, called strike, foul}\} | C_{ij-1}^{(a_D, a_S)} \right. \\ = b-(s-1), R_i = r, A_i = a_H] \\ \times P[C_{ij-1}^{(a_D, a_S)} = b-(s-1) | R_i = r, A_i = a_H] \left. \right\} \\ + I[s = 2] P[O_{ij-1}^{(a_D, a_S)} = \text{foul} | C_{ij-1}^{(a_D, a_S)} = b-s, R_i = r, \\ A_i = a_H] P[C_{ij-1}^{(a_D, a_S)} = b-s | R_i = r, A_i = a_H]. \end{aligned} \quad (9)$$

That is, provided that we can estimate $P[O_{ij-1}^{(a_D, a_S)} = \text{ball} | C_{ij-1}^{(a_D, a_S)} = (b-1)-s, R_i = r, A_i = a_H]$, $P[O_{ij-1}^{(a_D, a_S)} = \text{whiff, called strike, foul} | C_{ij-1}^{(a_D, a_S)} = b-(s-1), R_i = r, A_i = a_H]$, and $P[O_{ij-1}^{(a_D, a_S)} = \text{foul} | C_{ij-1}^{(a_D, a_S)} = b-s, R_i = r, A_i = a_H]$, then we can estimate $P[C_{ij}^{(a_D, a_S)} = b-s | R_i = r, A_i = a_H]$ recursively. We now discuss how each of those terms may be estimated. We note that if the assumptions in Section 2.6 are valid, the first term in the right hand side of Equation 9 is given by

$$\begin{aligned} P[O_{ij-1}^{(a_D, a_S)} = \text{ball} | C_{ij-1}^{(a_D, a_S)} = (b-1)-s, R_i = r, A_i = a_H] \\ = \int P_D^{(g_{j-1})} [D_{ij-1} = d_{j-1} | C_{ij-1}^{(a_D, a_S)} = (b-1)-s, \\ R_i = r, A_i = a_H] \\ \times P_S^{(g_{j-1})} [S_{ij-1} = 0 | D_{ij-1} = d_{j-1}, C_{ij-1}^{(a_D, a_S)} \\ = (b-1)-s, A_i = a_H] \\ \times P[O_{ij-1}^{(a_D, a_S)} = \text{ball} | D_{ij-1} = d_{j-1}, S_{ij-1} = 0, \\ C_{ij-1}^{(a_D, a_S)} = (b-1)-s, A_i = a_H] dd_{j-1} \\ = \int P[D_{ij-1} = d_{j-1} | C_{ij-1} = (b-1)-s, \\ R_i = r, A_i = a_D] \\ \times P[S_{ij-1} = 0 | D_{ij-1}, C_{ij-1} = (b-1)-s, A_i = a_S] \\ \times P[O_{ij-1} = \text{ball} | D_{ij-1} = d_{j-1}, S_{ij-1} = 0, \\ C_{ij-1} = (b-1)-s, A_i = a] dd_{j-1} \end{aligned} \quad (10)$$

demonstrating how the conditional distribution of potential outcomes (e.g. $O_{ij-1}^{(a_D, a_S)} = \text{ball}$) under regimes a_D and a_S with batter a_H hitting is identified from observed data. We discussed in Sections 3.2–3.4 how to estimate each of the conditional probabilities of the observed data. In particular, the estimated distributions are all discrete so the integrals simplify to summations. Therefore, we can estimate $P[O_{ij-1}^{(a_D, a_S)} = \text{ball} | C_{ij-1}^{(a_D, a_S)} = (b-1)-s, R_i = r, A_i = a_H]$ using

$$\begin{aligned} \hat{P}[O_{ij-1}^{(a_D, a_S)} = \text{ball} | C_{ij-1}^{(a_D, a_S)} = (b-1)-s, \\ R_i = r, A_i = a_H] \\ = \sum_{d_{j-1} \in \mathcal{D}_{(b-1)-s, r}} \hat{P}[D_{ij-1} = d_{j-1} | C_{ij-1} = (b-1)-s, \\ R_i = r, A_i = a_D] \\ \times \hat{P}[S_{ij-1} = 0 | D_{ij-1} = d_{j-1}, C_{ij-1} = (b-1)-s, \\ A_i = a_S] \\ \times \hat{P}[O_{ij-1} = \text{ball} | D_{ij-1} = d_{j-1}, S_{ij-1} = 0, \\ C_{ij-1} = (b-1)-s, A_i = a_H], \end{aligned} \quad (11)$$

where $\mathcal{D}_{c,r}$ indicates the support of pitch characteristics for ball-strike count c from pitchers with arsenal r , and $\hat{P}[\cdot]$ are the estimated probabilities from the models described in Section 3 to the data. A similar approach can be used to obtain estimates of $P[O_{ij-1}^{(a_D, a_S)} = \text{whiff, called strike, or foul} | C_{ij-1}^{(a_D, a_S)} = b-(s-1), R_i = r, A_i = a_H]$ and $P[O_{ij-1}^{(a_D, a_S)} = \text{foul} | C_{ij-1}^{(a_D, a_S)} = b-s, R_i = r, A_i = a_H]$ so that we can estimate $P[C_{ij}^{(a_D, a_S)} = b-s | R_i = r, A_i = a_H]$ by plugging in estimated quantities using the observed data for each term in the recursive formula given in Equation 9.

Returning to Equation 8, using the G-computation algorithm and substituting the estimated conditional probabilities of the observed data for the conditional distribution of the potential outcomes gives

$$\begin{aligned} \hat{P}[O_i^{(a_D, a_S)} = \text{hit} | A_i = a_H] \\ = \sum_j \sum_{r \in \mathcal{R}} \sum_{c_j \in \mathcal{C}} \sum_{d_j \in \mathcal{D}_{c_j, r}} \sum_{s_j \in \{0,1\}} \hat{P}[R_i = r | A_i = a_D] \\ \times \hat{P}[C_{ij}^{(a_D, a_S)} = c_j | R_i = r, A_i = a_H] \\ \times \hat{P}[D_{ij} = d_j | C_{ij} = c_j, R_i = r, A_i = a_D] \\ \times \hat{P}[S_{ij} = s_j | D_{ij} = d_j, C_{ij} = c_j, A_i = a_S] \\ \times \hat{P}[O_{ij} = \text{hit} | D_{ij} = d_j, S_{ij} = s_j, C_{ij} = c_j, A_i = a_H], \end{aligned} \quad (12)$$

where $\mathcal{C} = \{0-0, 1-0, 0-1, \dots, 3-2, \text{ plate appearance complete}\}$ is the set of all possible pitch counts and $\mathcal{A}$ is the set of all arsenals. A very similar approach can be used to estimate $P\left[O_i^{(ad, as)} = \text{out} \mid A_i = a_H\right]$ and $P\left[O_i^{(ad, as)} = \text{walk} \mid A_i = a_H\right]$ to estimate the counterfactual batting average, on-base percentage, and slugging percentage which we describe in detail in the Supplementary Material.

3.6 Estimating uncertainty

Standard errors for the estimated counterfactual batting average, on-base percentage, and slugging percentage were estimated using the nonparametric bootstrap (1,000 bootstrap resamples). Smoothing parameters were held fixed across bootstrap resample datasets.

4 Results

The observed batting average, on-base percentage, and slugging percentage for Castro and McCutchen during the 2012–2014 seasons against right-handed pitching among the plate appearances included in our analysis are given in Table 1. As noted in the introduction, Andrew McCutchen produced uniformly better offensive statistics than Starlin Castro throughout the 2012–2014 seasons. We first highlight differences between the two players in terms of their propensity to swing at pitches and the outcomes given that they swing based on the models described in Sections 3.3 and 3.4 before describing the anticipated outcomes under different counterfactual scenarios. Because the effect of location and movement is modeled using a thin plate regression spline basis expansion the regression coefficients are not interpretable, so we present the estimated models graphically.

4.1 Plate discipline

Figures 5 and 6 give the model-based probability of swinging at pitches at different locations with the typical break and speed of a four-seam fastball and slider, respectively, on 0-0 and 3-2 counts. Early in the count Castro is much less aggressive than McCutchen with four-seam fastballs and, to a lesser extent, sliders in the middle of the plate (Figures 5A and 6A). The results emphasize that the differences in plate discipline between the two players are not only that Castro swings at more pitches outside the

Table 1: Observed batting statistics facing right-handed pitchers for Starlin Castro and Andrew McCutchen in 2012–2014 MLB seasons among the plate appearances with complete PITCHf/x data and excluding plate appearances resulting in a sacrifice bunt, sacrifice fly, or batter hit by the pitch.

	Castro	McCutchen
Total at-bats	1401	1384
Total at-bats + BB	1463	1560
Batting average	0.273	0.310
On-base percentage	0.303	0.388
Slugging	0.407	0.512

Batting average, on-base percentage, and slugging percentage calculated using the three-season totals.

zone, but is less aggressive early in the count on presumably desirable pitches. This difference has important implications for batting performance which we formalize in Section 4.3. First, not surprisingly, pitches in the middle of the plate are most likely to result in hits if swung at. Second, pitches in the middle of the plate which are not swung at are, of course, very likely to be called strikes. Not swinging at these pitches in the middle of the plate puts the batter behind in the count which allows pitchers to pitch outside the zone, an area where Castro is likely to swing but has little success.

The difference between players' choice to swing grows more pronounced deeper in the count, as seen in Figures 5B and 6B. McCutchen remains quite disciplined even on full counts although expands the region where he is likely to swing slightly beyond the strike zone to protect against taking a called third strike. Castro by contrast is not very selective, swinging even at pitches far outside the zone. He is particularly susceptible to swinging at sliders low and away with greater than 80% chance of swinging at sliders just 6 inches off the ground. Although the comparison between the players is less dramatic for four-seam fastballs, Castro is still much more likely to swing at pitches both inside and up in the strike zone.

4.2 Batting ability

Examining the model-based differences between McCutchen and Castro helps explain some of the counterintuitive intermediate results. For example, despite McCutchen's superior batting performance (Table 1), he whiffs on more pitches than Castro (10.4% versus 9.3%). However, after adjusting for pitch location, movement speed, and count, the models indicate that McCutchen tends to whiff more frequently on breaking pitches that are down, away, and outside the strike zone and fastballs

Figure 5: Modeled probability of swinging at four-seam fastballs for Castro and McCutchen for 0-0 count (A) or for 3-2 count (B). The perspective here is from the batter/catcher and the green box indicates the strike zone.

which are up in the strike zone than Castro (see Figure 7). This approach is actually preferable early in the count (before the batter has two strikes) as the percentage of time those pitches end up as (extra-base) hits if put in play is much lower than if the pitch were in other parts of the strike zone (Figure 8). That is, it may be preferable to whiff on pitches which are unlikely to result in hits with the hope that more hittable pitches would happen later in the count.

Figure 8 also demonstrates that even after adjusting for the movement, speed, and count of the pitch, McCutchen has much greater power over a wider range of the strike zone than Castro. This is particularly true for fastballs. In other words, even after adjusting for the characteristics of the pitches the batter swings at, there remain significant differences in the ability of the two batters.

4.3 Counterfactual results

Observing these differences between players qualitatively, we compare batting average, on-base percentage, and slugging percentage for these players with (possibly) unobserved combinations of McCutchen and Castro's plate discipline and the distribution of pitches thrown to each player. In Tables 2, 3, and 4 we display the results for 6 possible counterfactual combinations alongside the model-based predicted outcomes for the observed combinations (i.e. all Castro and all McCutchen). The model-based predictions of batting average, on-base percentage, and slugging percentage are very close to the observed values (see Table 1) suggesting that the modeling assumptions we made were reasonable.

Figure 6: Modeled probability of swinging at sliders for McCutchen and Castro for 0-0 count (A) or for 3-2 count (B). The perspective here is from the batter/catcher and the green box indicates the strike zone.

For batting average and slugging percentage, the distribution of pitch characteristics does not seem to greatly impact outcomes. However, OBP tends to be slightly higher with McCutchen's pitch selection than Castro's. This is perhaps counter-intuitive. One might expect that a player (e.g. Castro) who swings at many breaking pitches outside the strike zone would be more likely to face those types of pitches. Another batter (e.g. McCutchen) who infrequently swings at those pitches may walk more often if he faced more of those pitches. However, the percentage of pitches on each count that each batter sees outside the strike zone is fairly consistent among the two batters except for on 3-0 and 3-1 counts when McCutchen sees more pitches outside the zone. That is, many pitchers choose to "pitch around" or intentionally walk McCutchen. For

a given plate discipline and hitting ability combination, seeing McCutchen's pitch selection yields a 0.005 to 0.013 increase in OBP over Castro's pitch selection (Table 3).

As expected, plate discipline has a large effect on a player's OBP, the one metric that we consider that accounts for walks. We estimate that Castro's OBP would increase by 0.040 (se = 0.006) if he were to adopt the plate discipline of McCutchen (Table 3). Such an increase accounts for 50% of the total difference of 0.080 in the model-based estimates of Castro and McCutchen's OBP.

Perhaps more surprising was the substantial effect of plate discipline on batting average and slugging percentage, two metrics which do not account for walks. About half of the 0.037 difference between Castro and McCutchen's batting average could be accounted for by

Figure 7: Modeled probability of a whiff given the batter chooses to swing for McCutchen and Castro for 1-1 count for four-seam fastballs (A) or sliders (B). The perspective here is from the batter/catcher and the green box indicates the strike zone. Yellow dots indicate the observed whiffs for each pitch type across all counts. The probability estimates are more precise in regions with more observed whiffs.

Castro adopting McCutchen's plate discipline which is estimated to lead to a 0.017 increase ($se = 0.004$) in batting average (Table 2). Adopting only McCutchen's plate discipline is expected to increase Castro's slugging percentage by 0.028 ($se = 0.008$), or 27% of the difference of the model-based estimate of the difference in slugging between Castro and McCutchen (0.102, Table 4). Previous work by Baumer (2008) has shown algebraically that an increased walk rate would be associated with a decreased batting average for the vast majority of players. The fact that the anticipated batting average would increase if Castro would adopt the plate discipline of McCutchen suggests that plate discipline also impacts other factors of the plate appearance including strike out, home run, and

batted-ball in play rate. Two factors likely contribute to the effect of plate discipline on these metrics. First, there is general acknowledgment that swinging (and missing) on pitches outside the strike zone early in the count (see Figure 6A) can lead to "pitcher-friendly" counts, which portend worse outcomes for all batters. Second, being less aggressive on "ideal" pitches in the middle of the strike zone early in the count (Figure 5A), likely contributes to worse outcomes for Castro.

Additionally, the approach we present here allows one to compare the differences in these hitting metrics holding the effects of pitch selection and plate discipline constant, which may be considered a more pure comparison of hitting ability. For example, if we assume that both

Figure 8: Expected total bases (i.e. slugging percentage) for McCutchen and Castro for 1-1 count for four-seam fastballs (A) or sliders (B). The perspective here is from the batter/catcher and the green box indicates the strike zone.

hitters adopted McCutchen's plate discipline and received pitches like McCutchen, the slash line for Castro was estimated to be 0.288/0.355/0.437 (Tables 2–4, row 4) compared to 0.308/0.382/0.510 (Tables 2–4, row 8) with none of those metrics significantly different at the 0.05 significance level.

The effect of plate discipline, pitch selection, and hitting ability appears to be more than additive for each of the metrics that we considered. For example, we expect that if Castro's plate discipline would improve to that of McCutchen, assuming he continued to see the same pitch distribution, his batting average would improve by 0.017. Keeping the plate discipline and pitch selection constant but changing the batter to McCutchen would improve batting average by 0.018. Changing both plate discipline and

hitting ability would be expected to improve batting average by 0.046, more than the sum of these two effects. Similar patterns are observed for on-base and slugging percentage.

5 Discussion

We have demonstrated the utility of a causal outcome framework in a baseball setting to understand how different factors impact batting performance. This allows for a more direct comparison of players who face different types of pitches and separates the impact of pitch selection, plate discipline, and hitting ability on a player's offensive

Table 2: Estimated batting average for counterfactual outcomes for specified batter's pitch selection, plate discipline, and hitting ability.

Pitch selection	Plate discipline	Hitting ability	Estimated batting average	Difference from all castro	Difference from all mcutchen
Castro	Castro	Castro	0.270 (0.011)	0.000 (0.000)	—0.037 (0.017)
McCutchen	Castro	Castro	0.271 (0.012)	0.000 (0.005)	—0.037 (0.017)
Castro	McCutchen	Castro	0.288 (0.012)	0.017 (0.004)	—0.020 (0.018)
McCutchen	McCutchen	Castro	0.288 (0.013)	0.017 (0.006)	—0.020 (0.018)
Castro	Castro	McCutchen	0.289 (0.012)	0.018 (0.016)	—0.019 (0.006)
McCutchen	Castro	McCutchen	0.284 (0.012)	0.013 (0.016)	—0.024 (0.004)
Castro	McCutchen	McCutchen	0.316 (0.013)	0.046 (0.017)	0.008 (0.005)
McCutchen	McCutchen	McCutchen	0.308 (0.012)	0.037 (0.017)	0.000 (0.000)

Bootstrap standard errors presented in parentheses. Differences of more than 2 standard errors are bolded.

Table 3: Estimated on-base percentage for counterfactual outcomes for specified batter's pitch selection, plate discipline, and hitting ability.

Pitch selection	Plate discipline	Hitting ability	Estimated on-base percentage	Difference from all Castro	Difference from all McCutchen
Castro	Castro	Castro	0.302 (0.012)	0.000 (0.000)	—0.080 (0.017)
McCutchen	Castro	Castro	0.310 (0.012)	0.008 (0.007)	—0.072 (0.017)
Castro	McCutchen	Castro	0.342 (0.013)	0.040 (0.006)	—0.040 (0.018)
McCutchen	McCutchen	Castro	0.355 (0.013)	0.053 (0.009)	—0.027 (0.017)
Castro	Castro	McCutchen	0.324 (0.012)	0.022 (0.016)	—0.058 (0.009)
McCutchen	Castro	McCutchen	0.329 (0.012)	0.027 (0.017)	—0.054 (0.006)
Castro	McCutchen	McCutchen	0.376 (0.013)	0.074 (0.017)	—0.006 (0.009)
McCutchen	McCutchen	McCutchen	0.382 (0.012)	0.080 (0.017)	0.000 (0.000)

Bootstrap standard errors presented in parentheses. Differences of more than 2 standard errors are bolded.

Table 4: Estimated slugging for counterfactual outcomes for specified batter's pitch selection, plate discipline, and hitting ability.

Pitch selection	Plate discipline	Hitting ability	Estimated slugging percentage	Difference from all Castro	Difference from all McCutchen
Castro	Castro	Castro	0.408 (0.023)	0.000 (0.000)	—0.102 (0.035)
McCutchen	Castro	Castro	0.407 (0.025)	—0.001 (0.014)	—0.103 (0.036)
Castro	McCutchen	Castro	0.436 (0.025)	0.028 (0.008)	—0.074 (0.035)
McCutchen	McCutchen	Castro	0.437 (0.028)	0.030 (0.017)	—0.072 (0.038)
Castro	Castro	McCutchen	0.462 (0.025)	0.054 (0.034)	—0.048 (0.014)
McCutchen	Castro	McCutchen	0.463 (0.024)	0.055 (0.033)	—0.047 (0.008)
Castro	McCutchen	McCutchen	0.514 (0.027)	0.106 (0.035)	0.004 (0.011)
McCutchen	McCutchen	McCutchen	0.510 (0.026)	0.102 (0.035)	0.000 (0.000)

Bootstrap standard errors presented in parentheses. Differences of more than 2 standard errors are bolded.

performance. Here, we are able to determine the impact of improvement in one or a combination of these aspects on the performance of Starlin Castro as compared to Andrew McCutchen with a particular focus on the effects of plate discipline. In particular, we found that the difference in plate discipline between these two players accounts for 50% and 46% of the difference between their batting average and on-base percentage, respectively, but only 27% of the difference in the slugging percentage.

The approach that we develop allows an analyst to have a clear understanding of the effect of changing the plate discipline of a batter on measures of batting performance rather than discuss the effect of a batter's swing choice in generalities. Importantly, our approach allows the effect of swinging at pitches outside the zone to vary between batters. That is, the effect of poor plate discipline may be substantially different between players as batters have differing abilities to hit pitches outside the

zone. Although we focused on common measures of batting performance including batting average, on-base percentage, and slugging percentage, the method that we have proposed could easily be extended to other measures of batting performance.

The approach that we take compares the batting outcomes under the assumption that a player adopts another's plate discipline throughout the entire plate appearance. However, we could examine the effect of altering a player's plate discipline in specific scenarios. For example, we may be interested in estimating the effect if Castro were able to adopt the plate discipline of McCutchen on two-strike counts only. Alternatively, we could examine the effect if Castro were to reduce the probability of swinging at pitches low and outside the strike zone (without necessarily adopting another player's pitch discipline). Each of these counterfactual scenarios can easily be estimated within the proposed framework.

In this paper, we estimate the effect of a pitch selection intervention under a specific intervention of the pitch distribution. In the mediation literature, this quantity is known as the controlled direct effect. Estimating this quantity may underestimate the total effect of plate discipline as the batter-pitcher relationship is inherently adversarial. As scouting reports are updated, pitchers would respond to changes in a batter's plate discipline. Batters would similarly make changes in response to facing different types of pitches, and so on. Our goal is to estimate the immediate expected outcomes if a player were to adopt a new plate discipline or pitchers were to change their strategy for facing that batter. Although we do not model the long-term adversarial relationship, our results should not be construed as to imply that a plate discipline or pitch selection strategy would remain fixed over time. Indeed, estimating how pitchers respond to changes in a players plate discipline (and vice versa) is an important area of future research. Once one has estimated the anticipated pitch distribution a batter would face, this could easily be incorporated into the G-computation algorithm described in this manuscript. For example, we might postulate that if Castro were to adopt McCutchen's plate discipline the proportion of pitches that he saw outside the strike zone would be reduced. We could consider a situation in which the percentage of pitches Castro would see outside the strike zone would be reduced by 20%. In that scenario, Castro's estimated slash line was 0.292/0.318/0.445. We note that such an adjustment is likely extreme as the percentage of pitches McCutchen sees outside the strike zone on each count is similar Castro. Nonetheless, several of these manual adjustments to Castro's current pitch selection strategy to reflect the adjustments pitchers could

make to Castro's new plate discipline could be performed as a sensitivity analysis.

As is the case in modeling any decision process, our models may not capture all the subtleties of pitching or hitting. For example, our approach treated "swinging" as a binary variable but how a batter swings may be different in different contexts (e.g. fighting off a pitch, swinging for power, or swinging for contact) which we did not consider in this manuscript. Additionally, the speed and break of pitches within an at-bat tend to be more similar than across different at-bats with different pitchers; our models for pitch selection did not account for this. We also did not explicitly account for the effects of pitch-sequencing in our models. Efforts to quantify the effect of pitch sequencing are relatively nascent, and we did not find strong evidence of pitch sequencing effects in our data. For example, there was little evidence that, within pitchers with the same arsenal, the location of the previous pitch was associated with location of subsequent pitch after adjusting for count. Furthermore, attempts to include information on the prior pitch (e.g. whether or not the pitch type was the same as the previous pitch, the distance between successive pitches, and the difference in speed between successive pitches) did not significantly improve the models for batter outcomes and, therefore, were not included in the final models. It is possible that with more data we could detect more subtle effects, but we feel confident that the effect of pitch sequencing on batter outcomes is relatively minimal at least for the two batters considered in this analysis. Finally, our modeling approach did not account for "game situation" variables including inning, current score, number of runners on-base, defensive alignment, etc. Preliminary analyses have shown that these factors are generally insignificant in the models that we fit. Furthermore, we found the model-based estimates of batting average, on-base percentage, and slugging percentage for the two players considered in this paper to be well-calibrated with their observed values suggesting that our assumptions were reasonable. However, there may be some residual confounding from failing to adjust for pitch sequencing and game situation characteristics.

The G-computation algorithm used here is also considerably more complicated than in most previous applications. Our analysis of multiple offense-related factors – pitch distribution and plate discipline – is analogous to estimating the effect of two separate treatments for the management of a disease. Most clinical applications using the G-computation algorithm only consider the effect of one treatment over time. However, management of chronic conditions, particularly those such as diabetes which often result in additional co-morbidities, will require

understanding the benefit of multiple treatments simultaneously. The framework presented here could serve as a model for such applications.

Though we focus here on just two players, this empirical comparison illustrates the great potential of the causal inference framework within a baseball setting. For example, we could use this approach to create standardized performance statistics for players across a team or league using a common pitch selection distribution to allow for more direct comparison. We could further envision using analysis to identify how improving different aspects of a player's batting would yield greater improvement in outcomes, which could then be used to determine a coaching strategy. Estimating the generalized additive models used in the G-computation formula takes approximately 240 seconds for each player with 3 years of plate appearance data using a single core on a machine with 32 GB of RAM and an Intel Xeon processor. Computational times could be improved by fitting these models in parallel if one wanted to fit models for multiple players in the league. We hope the framework developed here allows for a more informative understanding of the effect of various components of batting, including plate discipline, on measures of batter performance. By exemplifying the use of a causal inference framework for baseball, we also hope to expand the audience for this framework beyond biostatistics.

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. <https://doi.org/10.2202/1559-0410.1108>.
- Ahern, J., A. Hubbard, and S. Galea. 2009. "Estimating the Effects of Potential Public Health Interventions on Population Disease Burden: A Step-by-Step Illustration of Causal Inference Methods." *American Journal of Epidemiology* 169:1140–1147.
- Albert, J. 2006. "Pitching Statistics, Talent and Luck, and the Best Strikeout Seasons of All-time." *Journal of Quantitative Analysis in Sports* 2. <https://doi.org/10.2202/1559-0410.1014>.
- Almirall, D., T. Ten Have, and S. A. Murphy. 2010. "Structural Nested Mean Models for Assessing Time-Varying Effect Moderation." *Biometrics* 66: 131–139.
- Baumer, B. S. 2008. "Why On-Base Percentage is a Better Indicator of Future Performance than Batting Average: An Algebraic Proof." *Journal of Quantitative Analysis in Sports* 4. <https://doi.org/10.2202/1559-0410.1101>.
- Baumer, B. S., S. T. Jensen, and G. J. Matthews. 2015. "openWAR: An Open Source System for Evaluating Overall Player Performance in Major League Baseball." *Journal of Quantitative Analysis in Sports* 11:69–84.
- Berry, S. M., C. S. Reese, and P. D. Larkey. 1999. "Bridging Different Eras in Sports." *Journal of the American Statistical Association* 94:661–676.
- Cain, L. E., J. M. Robins, E. Lanoy, R. Logan, D. Costagliola, and M. A. Hernán. 2010. "When to Start Treatment? A Systematic Approach to the Comparison of Dynamic Regimes Using Observational Data." *The International Journal of Biostatistics* 6. <https://doi.org/10.2202/1557-4679.1212>.
- Carleton, R. 2007. "Is Walk the Opposite of Strikeout?" In *By the Numbers: The Newsletter of the SABR Statistical Analysis Committee* 17:3–7.
- Cole, S. R. and C. E. Frangakis. 2009. "The Consistency Statement in Causal Inference: A Definition or An Assumption?" *Epidemiology* 20: 3–5.
- Holland, P. W. 1986. "Statistics and Causal Inference." *Journal of the American Statistical Association* 81:945–960.
- James, B. 1986. *The Bill James Historical Baseball Abstract*. New York: Random House Inc.
- Kaplan, D. 2006. "A Variance Decomposition of Individual Offensive Baseball Performance." *Journal of Quantitative Analysis in Sports* 2. <https://doi.org/10.2202/1559-0410.1035>.
- Kvam, P. H. 2011. "Comparing Hall of Fame Baseball Players Using Most Valuable Player Ranks." *Journal of Quantitative Analysis in Sports* 7. <https://doi.org/10.2202/1559-0410.1337>.
- Lentzer, M. 2011. "A New Take on Plate Discipline – Redefining the Zone." URL <http://www.baseballprospectus.com/article.php?articleid=15216>.
- Miller, S. and J. Wojciechowski. 2015. *Baseball Prospectus 2015: The Essential Guide to the 2015 Season*. Nashville, TN: Wiley.
- Murphy, S. A., M. J. van der Laan, and J. M. Robins. 2001. "Marginal Mean Models for Dynamic Regimes." *Journal of the American Statistical Association* 96:1410–1423.
- Nieswiadomy, M. L., M. C. Strazicich, and S. Clayton. 2012. "Was There a Structural Break in Barry Bonds's Bat?" *Journal of Quantitative Analysis in Sports* 8. <https://doi.org/10.1515/1559-0410.1305>.
- Paine, N. 2014. "The Most Disciplined MLB Batters." URL <http://fivethirtyeight.com/datalab/the-most-disciplined-mlb-batters/>.
- Pearl, J. 2009. "Causal Inference In Statistics: An Overview." *Statistics Surveys* 3:96–146.
- Phillips, D. C. 2011. "You're Hurting My Game: Lineup Protection and Injuries in Major League Baseball." *Journal of Quantitative Analysis in Sports* 7. <https://doi.org/10.2202/1559-0410.1296>.
- Robins, J. 1986. "A New Approach to Causal Inference in Mortality Studies with a Sustained Exposure Period – Application to Control of the Healthy Worker Survivor Effect." *Mathematical Modelling* 7:1393–1512.
- Robins, J. M. 1997. "Causal Inference from Complex Longitudinal Data." Pp. 69–117 in *Latent Variable Modeling and Applications to Causality*. New York: Springer.
- Robins, J. M. and S. Greenland. 1994. "Adjusting for Differential Rates of Prophylaxis Therapy for PCP in High-Dose Versus Low-Dose AZT Treatment Arms in an AIDS Randomized Trial." *Journal of the American Statistical Association* 89:737–749.
- Robins, J. M., M. A. Hernan, and B. Brumback. 2000. "Marginal Structural Models and Causal Inference in Epidemiology." *Epidemiology* 11:550–560.

- Rubin, D. B. 1974. "Estimating Causal Effects of Treatments in Randomized and Nonrandomized Studies." *Journal of Educational Psychology* 66:688–701.
- Rubin, D. B. 1990. "Comment: Neyman (1923) and Causal Inference in Experiments and Observational Studies." *Statistical Science* 5:472–480.
- Schell, M. J. 2005. *Baseball's All-Time Best Sluggers: Adjusted Batting Performance from Strikeouts to Home Runs*. Princeton, NJ: Princeton University Press.
- Schmotzer, B. J., J. Switchenko, and P. D. Kilgo. 2008. "Did Steroid Use Enhance the Performance of the Mitchell Batters? The Effect of Alleged Performance Enhancing Drug Use on Offensive Performance from 1995 to 2007." *Journal of Quantitative Analysis in Sports* 4. <https://doi.org/10.2202/1559-0410.1130>.
- Sievert, C. 2014a. *pitchRx: Tools for Harnessing MLBAM Gameday Data and Visualizing PITCHf/x*. URL <http://cpsievert.github.com/pitchRx>, R package version 1.5.
- Sievert, C. 2014b. "Taming PITCHf/x Data with XML2R and pitchRx." *The R Journal* 6: 5–19.
- Tango, T. M., M. G. Lichtman, and A. E. Dolphin. 2007. *The Book: Playing the Percentages in Baseball*. Washington, DC: Potomoc Press.
- Taubman, S. L., J. M. Robins, M. A. Mittleman, and M. A. Hernán. 2009. "Intervening on Risk Factors for Coronary Heart Disease: An Application of the Parametric g-formula." *International Journal of Epidemiology* 38:1599–1611.
- Westreich, D., S. R. Cole, J. G. Young, F. Palella, P. C. Tien, L. Kingsley, S. J. Gange, and M. A. Hernán. 2012. "The Parametric g-formula to Estimate the Effect of Highly Active Antiretroviral Therapy on Incident AIDS or Death." *Statistics in Medicine* 31:2000–2009.
- Wood, S. N. 2006. *Generalized Additive Models: An Introduction with R*. Boca Raton, FL: Chapman and Hall/CRC.
- Wood, S. N. 2003. "Thin-Plate Regression Splines." *Journal of the Royal Statistical Society (B)* 65:95–114.
- Wood, S. N. 2004. "Stable and Efficient Multiple Smoothing Parameter Estimation for Generalized Additive Models." *Journal of the American Statistical Association* 99:673–686.
- Wood, S. N. 2015. *Mixed GAM Computation Vehicle with GCV/AIC/REML Smoothness Estimation*. R package version 1.8-7.
- Young, J. G., L. E. Cain, J. M. Robins, E. J. O'Reilly, and M. A. Hernán. 2011. "Comparative Effectiveness of Dynamic Treatment Regimes: An Application of the Parametric g-formula." *Statistics in Biosciences* 3:119–143.

Supplemental Material: The online version of this article offers supplementary material (<https://doi.org/10.1515/jqas-2016-0029>).