Searching for the GOAT of tennis win prediction

Stephanie Ann Kovalchik*

Searching for the GOAT of tennis win prediction

DOI 10.1515/jqas-2015-0059

Abstract: Sports forecasting models – beyond their interest to bettors – are important resources for sports analysts and coaches. Like the best athletes, the best forecasting models should be rigorously tested and judged by how well their performance holds up against top competitors. Although a number of models have been proposed for predicting match outcomes in professional tennis, their comparative performance is largely unknown. The present paper tests the predictive performance of 11 published forecasting models for predicting the outcomes of 2395 singles matches during the 2014 season of the Association of Tennis Professionals Tour. The evaluated models fall into three categories: regression-based, point-based, and paired comparison models. Bookmaker predictions were used as a performance benchmark. Using only 1 year of prior performance data, regression models based on player ranking and an Elo approach developed by FiveThirtyEight were the most accurate approaches. The FiveThirtyEight model predictions had an accuracy of 75% for matches of the most highly-ranked players, which was competitive with the bookmakers. The inclusion of career-to-date improved the FiveThirtyEight model predictions for lower-ranked players (from 59% to 64%) but did not change the performance for higher-ranked players. All models were 10–20 percentage points less accurate at predicting match outcomes among lower-ranked players than matches with the top players in the sport. The gap in performance according to player ranking and the simplicity of the information used in Elo ratings highlight directions for further model development that could improve the practical utility and generalizability of forecasting in tennis.

Keywords: betting; probit models; sports forecasting; validation.

1 Introduction

Predicting wins is a preoccupation of every sport. Tennis is no exception. For over a century, individuals have been fascinated and perplexed by the problem of forecasting match outcomes of the most popular of racquet sports, as is evident in this commentary from the 1898 issue of the *Lawn Tennis Guide* concerning the likely outcome of that year's draw for the Canadian International Tournament,

> 'It would have puzzled the canniest card in the business all the same to pick a winner out of these gentleman, any one of which was liable to develop a streak and win out.'

Real progress in tackling the "puzzle" of tennis forecasting began only decades ago with the publication of the first mathematical models to predict match outcomes. Since that time, many more models have been developed and interest in statistical approaches to forecasting wins in tennis has continued to grow.

Although multiple statistical models now exist for predicting tennis wins, few have been rigorously tested or compared against alternative approaches. Just as in sport itself, a high standard of performance should be the ultimate goal of a prediction model. However, there is currently little known about the validity and comparative utility of existing forecasting models.

The purpose of the present paper was to study the performance of published models that predict singles match outcomes in professional tennis. The accuracy and discriminatory power of 11 different forecasting approaches were evaluated in a large dataset of singles match outcomes for the 2014 season of the Association of Tennis Professionals (ATP) Tour. Performance differences by type of surface and tournament level were also investigated. Through the comparative validation of existing prediction models, this study aims to identify the major determinants of win ability in professional tennis and highlight ways to improve the performance of existing models.

2 Review of tennis prediction models

2.1 Notation

The symbol π_{ij} will be used to denote the probability that the *i*th player wins a tennis match when facing the *j*th opponent. The player with the higher ranking, according

***Corresponding author: Stephanie Ann Kovalchik,** Tennis Australia, Melbourne Park, Olympic Blvd, Melbourne, VIC 3000, Victoria, Australia, Tel.: +61 4 5050 9098, e-mail: s.a.kovalchik@gmail.com; and Institute of Sport Exercise and Active Living, Victoria University, Melbourne, Victoria, Australia

to ATP Emirates World Rankings at the time of the match, will be designated as the *i*th player.

2.2 Models

A literature review was conducted to identify published models for forecasting wins in tennis. The first stage of the search strategy queried Google Scholar for articles containing "tennis" and at least one word containing "predict" or "forecast." Hits were reviewed for relevance. Articles were considered relevant if they mentioned a strategy for match prediction in the abstract. In the second stage of the review process, the citations of the relevant articles were also reviewed.

From this search strategy, 17 articles underwent a detailed reading to determine their eligibility. Articles were excluded that did not present enough information to predict match wins (2 articles), did not present a forecasting model (2 articles), presented a previously published model (2 articles), or required within-match updating (1 article). Ten articles remained after applying these exclusion criteria. In addition to the models identified in the literature, there was one additional prediction method developed by analysts at FiveThirtyEight (www.fivethirtyeight.com) that, owing to its popularity and reproducibility, was also included.

The models included in the validation study are described in Table 1. The majority of approaches fall into three categories of models: regression, point-based, and paired comparison models. Regression models directly model the winner of the match, and most proposed approaches within this class have been based on the probit family and included player rankings as a predictor. Point-based approaches model the probability of winning on serve, then derive a prediction for the match outcome from an algebraic formula under the assumption that points are independent and identically distributed (IID). The point-based models extend the earliest approach to implementing the IID model, where only a player's average ability on serve was utilized (Newton and Keller 2005), using more sophisticated models for the probabilities of winning a point on serve and return. Two models – an application of the Bradley-Terry model (1952) proposed by McHale and Morton (2011) and the FiveThirtyEight model – are examples of paired comparison approaches. Finally, as a standard of reference for predictive performance, the study includes predictions from the bookmaker consensus model of Leitner, Zeileis, and Hornik (2009). The following sections provide a technical description of each forecasting approach.

2.2.1 Regression-based

Probit models. Three of the evaluated models predict match outcomes in professional tennis using a probit regression model. The basic form of the probit model is, $\pi_{ij} = \Phi(x_{ij}'\beta)$, where Φ denotes the cumulative density function of a standard normal variable and x_{ij} represents a vector of predictors that could include player, opponent, or match characteristics. The models differ in the set of predictors they consider. The predictors for each model are listed in Table 2 and show that the only common predictor across all of the models was player rank (or seeding).

Table 1: Summary of published models for forecasting outcomes in tennis.

Shorthand	Model description (source)
Regression-based	
Logistic	Logistic with player ranking (Klaassen and Magnus 2003)
Basic probit	Probit model with player seeding (Boulier and Stekler 1999)
Probit plus	Probit model with player ranking and demographics (Del Corral and Prieto-Rodriguez 2010)
Prize probit	Probit model with ranking, prize earnings, and demographics (Gilsdorf and Sukhatme 2008)
Point-based	
Basic IID	Match win probability based on IID points (Newton and Keller 2005)
Opponent-adjusted IID	IID with opponent-adjusted point probabilities (Barnett and Clarke 2005)
Low-level IID	IID with low-level point probabilities (Spanias and Knottenbelt 2012)
Common opponent IID	IID with common opponent point probabilities (Knottenbelt, Spanias, and Madurska 2012)
Paired comparison	
Bradley-Terry	Bradley-Terry model of player abilities (McHale and Morton 2011)
FiveThirtyEight	Elo rating predictions
BCM	Bookmaker consensus model (BCM) (Leitner, Zeileis, and Hornik 2009)

IID, Independent identically distributed.

Table 2: Predictors for regression models of tennis match wins.

Variable	Logistic	Basic probit	Probit plus	Prize probit
Difference in seeds	+			
Difference in ranks		+	+	
Previous tournament result			+	
Former top 10 player			+	
Difference in age			+	+
Difference in height			+	
Handedness			+	
Potential prize earnings				+
Head-to-head wins				+
Head-to-head losses				+
Difference in rank points				+
Difference in career wins				+
Rounds remaining			+	+
Grand Slam indicator			+	+
Masters 1000 indicator				+

Boulier and Stekler (1999) proposed the simplest probit model (Basic Probit). It has the difference in seedings and an indicator for an unseeded opponent as the only predictors. An implicit assumption of this model is that an unseeded player has a 50–50 chance of winning a match when facing another unseeded player. The authors found a difference in the model coefficients for men's singles play at Wimbledon, so separate parameters for the seeding variables were obtained for Wimbledon.

Gilsdorf and Sukhatme (2008) were interested in the explanatory role of monetary incentives for win probabilities. In their model (Prize Probit), the main predictor was the gap between the top prize money for the tournament and the earnings for a loss at the current stage of the match. Additional predictors were player and opponent age, head-to-head results, remaining rounds, difference in ATP points, difference in career match wins for matches played on the same surface, and an indicator for a match at the Masters level or above. This model had a predictive accuracy of 63% when applied to ATP match outcomes for the 2000–2001 season. Models that included interactions between prize money and other predictors had a minimal improvement on performance.

Del Corral and Prieto-Rodriguez (2010) compared three models with different sets of predictors. The largest model included the difference in player ranks, an indicator for career rank within the top ten, player physical characteristics (height, age, and handedness), and tournament factors (round, tournament level). A second model excluded the rank predictors, and a third model excluded player physical characteristics. In terms of a Brier score, the predictive performance of the full model was superior when applied to within- and out-of-sample datasets for men's and women's match outcomes.

Logistic model. Klaassen and Magnus (2003) considered a logit model to predict match outcomes at the beginning of a match. Letting R_i be the rank of the ith player, who is favored to win the match, and R_j the rank of the jth opponent, the logistic model they propose is

$$log(\pi_{ij}/(1-\pi_{ij})) = \theta log(R_i/R_j). \quad (1)$$

The log ratio of player rankings was the only predictor included in the model.

2.2.2 Point-based

Point-based models refer to a general class of tennis forecasting models that begin by specifying probabilities of winning a point on serve and return. Under an IID assumption for point outcomes, the probability of winning a match is a function of the probabilities of winning a point on serve and return and can be written down in closed form. Newton and Keller (2005) were two of the earliest authors to give a full treatment of the point-based model along with formulae for predicting a match win given a player's probabilities of winning a point on serve and return, with the possibility of a tiebreaker to determine set wins. The authors did not propose a model for the serve and return probabilities but only suggested considering some adjustment for opponent ability. In this paper, a player's average frequency of point wins on serve and return in the prior 12 months are the inputs to the Newton and Keller formulae for match wins.

Subsequent to Newton and Keller's seminal paper on the IID model for tennis win probabilities, several authors have considered approaches for estimating the probabilities of winning a point on serve and return. Barnett and Clarke (2005) proposed a tournament-specific average adjusted for player advantage on serve and opponent advantage on return. Player advantage was estimated by how much better (or worse) a player's serve and return ability was compared to the average tour player. The authors illustrate the use of the model for one singles match but did not evaluate its performance more generally.

Spanias and Knottenbelt (2012) use a state model of the possible events leading to a win on serve (e.g. ace, point win on first serve, etc.) and combine the probabilities of these states to get an overall estimate of the probability of winning a point on serve. In this "low-level" point model, estimates for the probabilities of each state are averaged over some period of recent play for each competitor with adjustment for opponent strength in the same fashion

as the method proposed by Barnett and Clarke (2005). A similar approach can be used for estimating return win probabilities. The authors examined the prediction error of their approach for predicting matches of the 2011 ATP season and found that estimates based on 12 months of the most recent match outcomes outperformed estimates with 6 or 18 months of data.

Knottenbelt, Spanias, and Madurska (2012) also adjusted for the competitive level of a player's opponent in the estimated chance a player wins a point on serve or return, but the authors used a different angle than Barnett and colleagues. Specifically, these authors used statistical data from a subset of matches, within a predetermined period, which only included matches against opponents who have been faced by both players being modeled, thus a "common opponent model." The aim of this approach is to obtain the average serve and return performance of each player while eliminating the bias arising from the different mix of opponents each player has faced within the predetermined period. Knottenbelt et al. considered using the common-opponent average themselves or as inputs to the opponent-adjusted model of Barnett and colleagues. Both approaches were found to yield a positive return on investment for bets on the 2011 Grand Slams when estimates were stratified by surface.

2.2.3 Paired comparison

One alternative class of approaches to the regression and point-based models is the paired comparison model. A popular type of paired comparison model is the Bradley-Terry model, which was applied to forecasting outcomes in tennis by McHale and Morton (2011). With this approach, each player has a latent match win ability, α_i. In the Bradley-Terry formulation the odds that the ith player beats the jth player is α_i/α_j, which corresponds to the following match win probability

$$\pi_{ij} = \alpha_i / (\alpha_i + \alpha_j), \tag{2}$$

the ratio of a player's ability to the sum of the abilities in a given match. McHale and Morton estimate the abilities of professional tennis players from a likelihood of games won and lost between player and opponent, with an exponential decay function to weigh more recent matches more heavily. The ability parameters are obtained by maximizing the following log-likelihood

$$l(\boldsymbol{\alpha}) = \sum_i \sum_{j \in M_i} \exp(\varepsilon(t - t_j))[g_{ij} log(\alpha_i) + (\bar{g}_j - g_{ij}) \log(\alpha_{k(i)}) - \bar{g}_j log(\alpha_i + \alpha_{k(i)})]. \tag{3}$$

In Equation (3), g_{ij} is the total games won by the ith player in the jth match ($j \in M_i$), $\bar{g}_j$ are the total games played in the jth match, $k(i)$ is an index for the opponents, and ε is the half-life of an exponential decay function based on the time of the current match, t, and the previous match outcomes, t_j. The likelihood can be stratified by surface to obtain surface-specific abilities or surface effects can be incorporated through weighting.

Of all the methods considered in this paper, the approach that has been the most used in the media is the prediction method developed by the data journalists at FiveThirtyEight. Their approach is a variant of the popular Elo rating system (1978) that was first developed to rate player strength in chess and has since been adopted as a dynamic measure of strength in multiple sports (Stefani 2011). The FiveThirtyEight version (2015) uses the following recursive formula for updating the ith player's rating,

$$E_i(t+1) = E_i(t) + K * (\hat{\pi}_{ij}(t) - W_i(t)). \tag{4}$$

Here, t refers to the tth match, $W_i(t)$ is an indicator of whether the player won, $E_i(t)$ is the player's Elo rating at the start of the match, and $\hat{\pi}_{ij}(t)$ is the Elo-based prediction for a match win against the jth opponent. The parameter K is a function of the player's career matches played at time t. If we denote these matches as $m(t)$, then $K = 250/(m(t)+5)^{0.4}$. For Grand Slam tournaments, K is multiplied by 1.1 to give outcomes at the majors 10% greater weight than all other tournaments.

The Elo prediction that the ith player wins the tth match, $\hat{\pi}_{ij}(t)$, is equal to

$$\hat{\pi}_{ij}(t) = \left(1 + 10^{\frac{(E_j(t) - E_i(t))}{400}}\right)^{-1}, \tag{5}$$

which approaches 1 as the rating difference, $E_j(t) - E_i(t)$, becomes more negative. For their first match, players begin with a rating of $E(1) = 1500$.

2.2.4 Bookmakers consensus model

For any given professional tennis match there will be multiple bookmakers who publish winning odds. These odds can be regarded as an expert opinion about a player's win probability against a specific opponent. Leitner, Zeileis, and Hornik (2009) proposed aggregating these expectations into an overall prediction for match outcomes. Given K bookmakers for a given match, the estimated probability that the favored player will win is

$$\text{logit}\,(\hat{\pi}_{ij}) = K^{-1} \sum_{k=1}^{K} \text{logit}\,(\hat{\pi}_{ijk}). \tag{6}$$

The authors refer to this model as the bookmaker consensus model (BCM). The derivation of the individual $\hat{\pi}_{ijk}$ will depend on how odds are reported by the bookmaker and how one corrects for the bookmaker's profit margin (or overround). The present paper uses the correction method proposed by Shin (1993), which has been shown to outperform basic normalization and regression-based corrections (Štrumbelj 2014).

3 Methods

3.1 Parameter estimation

To allow for a fair comparison among approaches, all methods were estimated with data from a 52 week period. The period of inclusion was defined on a rolling basis with respect to the week of the current match so that the recency of the match outcomes used was consistent across the forecasts for the validation matches. The only exception to the 52 week rolling input data was for model predictors that specifically called for a longer look-back period (e.g. career wins). Player ability estimates for the Bradley-Terry model were based on game wins and were stratified by surface. Elo ratings were computed for match results of ATP players beginning at the start of the 2013 season (1 year before the earliest matches in the validation data described in the next section) up to the date of the validation match being predicted. All Grand Slam matches and ATP Tour matches above the Challenger level were included. The BCM included the pre-match odds of 7 bookmakers reported on www.tennis-data.co.uk.

Owing to the stratification by surface of the Bradley-Terry models, there was concern that estimates based on a single season might not be sufficiently stable. Thus, an additional version of the Bradley-Terry predictions were evaluated that were based on a 2-year rolling window (Bradley-Terry 2) rather than the 1-year rolling window (Bradley-Terry 1) described above. When 2 years of input data was used, a decay function was introduced that weighted outcomes that were more than 1 year from the date of the validation match with a weight that was one-half the weight of outcomes within 1 year of the validation match.

Elo ratings are traditionally based on career wins and losses. While using only 1 year of prior performance data would make the Elo predictions more directly comparable to the other approaches, it would not replicate the implementation that has been used in practice. For this reason, the study also includes Elo predictions based on career-to-date outcomes (FiveThirtyEight 2) in addition to those based on a single year of prior performance data (FiveThirtyEight 1).

A small percentage of matches for some of the methods did not have adequate data to estimate the method's parameters. In these cases, the missing data was imputed to the mean among players in the same rank group as the player with the missing value. For this imputation, players were grouped into two rank categories: those with an ATP rank below 100 and those with a rank of 100 or greater.

Data on model predictors and match outcomes were gathered from publicly available websites using the author's R package deuce (www.github.com/skoval/deuce). All analyses were conducted in the R language (R Core Team 2015). Estimates for the regression models were obtained from the modeling function glm and estimates for the Bradley-Terry model were fit with functions from the package BradleyTerry2 (Turner and Firth 2012). The predictions for each method and the programs used for their evaluation are provided as supplementary material.

3.2 Validation data

The performance of the models was tested against outcomes for 2395 ATP singles matches played during the 2014 season, excluding 105 matches that were walkovers or ended with a retirement (Table 3). Among the matches in the independent validation dataset, 20% occurred at a Grand Slam, 55% were played on hard court surfaces, and slightly more than half included a top 30 player. Higher ranked players won 68% of matches.

Table 3: Validation dataset of 2014 ATP singles matches.

Characteristic	Count	Percentage
Total matches	2395	100
Highest-ranked wins	1631	68.1
Series		
Grand Slams	482	20.1
Masters	549	22.9
Other	1364	57.0
Surface		
Clay	790	33.0
Grass	287	12.0
Hard	1318	55.0
Highest-ranked player		
Top 30	1235	51.6
Lower-ranked	1160	48.4

3.3 Performance

Four properties of model performance were evaluated: prediction accuracy, calibration, log-loss, and discrimination. Prediction accuracy is the percentage of correct predictions when wins are assigned to the player in a matchup with the higher probability. Calibration is also a measure of the model accuracy but focuses on the expected wins across a number of matches. A probabilistic forecast is well calibrated if, when considering all matches for which the predicted probability is π, the observed proportion of wins is close to π. To determine if this condition is met, the calibration ratio is used. This ratio is the sum of the win probabilities of the higher-ranked player across all matches divided by the number of matches won by the higher-ranked player. When a model is well-calibrated, it will have a calibration ratio close to 1. A model with a calibration ratio greater than 1 tends to overestimate the wins of the highest-ranked player; a model with a ratio less than 1 tends to underestimate the wins of the highest-ranked player.

A performance measure that has a direct connection with betting is the *log-loss function*. For N matches, the function is equal to

$$\text{Log-loss} = -N^{-1}\sum_i [y_i log(\hat{\pi}_i) + (1-y_i) log(1-\hat{\pi}_i)] \quad (7)$$

where y_i is an indicator of whether the higher-ranked player won the ith match and $\hat{\pi}_i$ is the corresponding prediction that the higher-ranked player wins. It can be shown that the log-loss function converges to the Kelly criterion when a bettor places π_i percent of their wealth on the ith match and $\hat{\pi}_i \to \pi_i$. A characteristic of the log-loss function is that there is a high penalty for incorrect predictions made with high confidence. Thus, it is ideally suited for a betting context where one wants to minimize overconfidence on incorrect bets (Yuan et al. 2015).

The final performance measure considered is model discrimination. In a diagnostic setting, discrimination is a measure of a test's ability to give positive results for true cases and negative results for non-cases. For the present paper, discrimination was measured as the mean prediction for matches higher-ranked players won minus the mean prediction for matches they lost (i.e. upsets), which is equal to the integrated discrimination improvement (IDI) measure used in diagnostic settings (Pencina, D'Agostino, and Vasan 2008). Higher values of the IDI indicate greater discriminatory power, as this would indicate that model predictions tend to be more certain for wins than for upsets.

4 Results

Across methods, prediction accuracy ranged from 59% to 72% (Table 4). Models within the regression-based and point-based classes performed similarly, predicting with 64–65% accuracy on average. Within the class of regression-based models, models with player ranking were the most accurate (Prize Probit, Probit Plus, Logistic) and there was little evidence that additional predictors beyond ranking improved accuracy. Within the class of point-based models, the Opponent Adjusted approach, with 67% accuracy, had the superior performance. The FiveThirtyEight model based on 1 year of performance data was more accurate than the Bradley-Terry model using the same input data (67% vs. 62%) and had accuracy that was comparable to the best-performing regression and point-based models.

Both of the paired comparison models improved accuracy when the amount of input data was increased. Across all approaches, only two models had prediction accuracies of 70% or greater: the BCM with 72% accuracy (the highest of all models) and the FiveThirtyEight model with 70% accuracy.

The calibration findings highlighted evidence of bias for several models. These included one regression model (Basic Probit), all point-based models save for the Opponent Adjusted model, and the Bradley-Terry model (Table 4). In each case, the bias showed that the models tended to underestimate the higher-ranked player's probability of winning, predicting more upsets than were actually observed in the validation data. The bias was most severe for the Bradley-Terry model, where the calibration ratio was 0.80 whether using 1 or 2 years of data.

The BCM had the lowest log-loss of all models, indicating that it was the least vulnerable to overconfident predictions than any other modeling approach. The FiveThirtyEight model (both with 1 year or career-to-date data) and the regression models that included player ranking were the next best-performing in terms of log-loss, all having log-losses of approximately 0.60. Relative to these methods, the point-based models and Bradley-Terry models gave more overconfident predictions.

There were only three models that had more than a 10% separation in the mean probabilities of expected wins and upsets (discrimination). These were the BCM, with a discrimination power of 14%; the Low Level model, with 12%; and the FiveThirtyEight model using career-to-date data with 11% (Table 4). As a class, the discriminatory ability of point-based models was approximately twice that of the regression-based models.

Table 4: Summary of prediction performance in 2014 ATP validation data by method type.

Method	Prediction accuracy	Calibration (95% CI)	Log-loss	Discrimination
Regression-based				
Mean	65	0.96 (0.94, 0.99)	0.61	5
Basic probit	59	0.92 (0.88, 0.97)	0.63	3
Prize probit	68	0.98 (0.93, 1.03)	0.61	4
Probit plus	67	0.98 (0.93, 1.03)	0.60	7
Logistic	67	0.97 (0.92, 1.02)	0.60	6
Point-based				
Mean	64	0.91 (0.88, 0.93)	0.66	9
Basic IID	63	0.89 (0.85, 0.93)	0.67	7
Opponent adjusted	67	0.98 (0.93, 1.03)	0.63	9
Low level	64	0.87 (0.83, 0.91)	0.68	12
Common opponent	63	0.89 (0.84, 0.93)	0.66	7
Paired comparison				
Bradley-Terry 1	62	0.80 (0.76, 0.83)	0.67	3
Bradley-Terry 2	65	0.80 (0.76, 0.84)	0.65	3
FiveThirtyEight 1	67	0.98 (0.93, 1.03)	0.60	9
FiveThirtyEight 2	70	1.03 (0.98, 1.08)	0.59	11
BCM	72	0.98 (0.93, 1.03)	0.55	14

Figure 1 summarizes the patterns in the four measures of performance across methods using the BCM as the performance benchmark. The patterns in performance between the regression-based and point-based models were complex. While the regression-based models that include player ranking tended to outperform the point-based approaches on measures of accuracy (accuracy, calibration, and log-loss), point-based models had superior discriminatory ability. One point-based model, the Opponent Adjusted approach, was exceptional in having both excellent accuracy and good discriminatory ability.

The FiveThirtyEight model using 1 year of data had similar accuracy as the Opponent Adjusted approach but superior log-loss and discriminatory ability. The inclusion

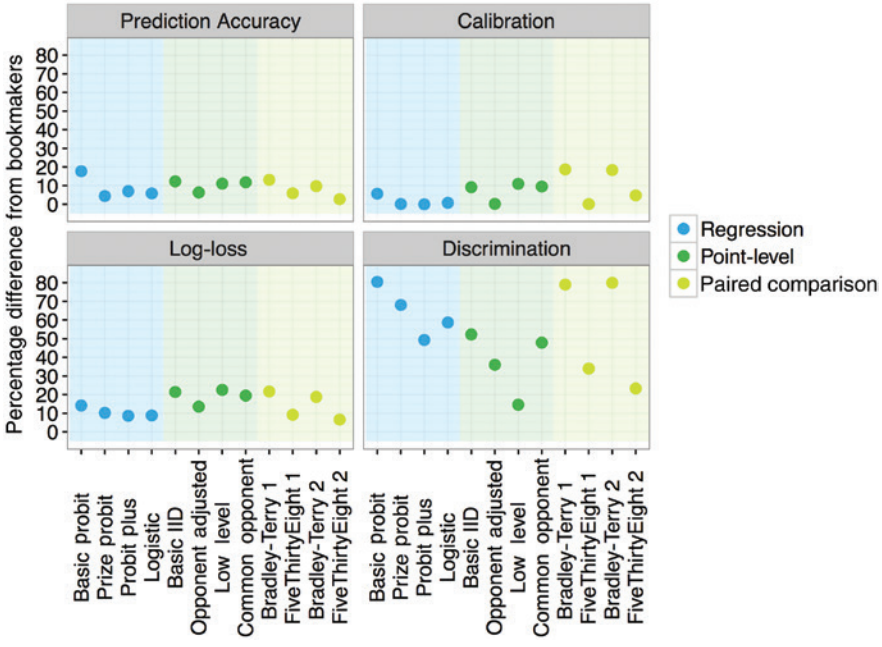

Figure 1: Performance summary – prediction accuracy, calibration, log-loss, and discrimination – for 2014 ATP validation data, as the percentage difference in performance from the bookmakers consensus model.

Table 5: Method differences in prediction accuracy in 2014 ATP validation data by player ranking, tournament level, and surface.

Prediction accuracy	Regression-based		Point-based		Bradley-Terry		FiveThirtyEight		BCM
	Mean	Range	Mean	Range	1	2	1	2	
Player ranking									
Top 30	74	2	70	6	67	71	75	75	76
Lower-ranked	56	18	58	5	57	58	59	64	67
Tournament level									
Grand Slams	73	7	70	6	65	65	75	74	78
Masters 1000	66	11	65	5	65	66	70	72	72
250 or 500	62	10	62	4	60	64	64	67	70
Surface									
Clay	63	8	61	4	60	63	66	67	70
Grass	67	11	65	5	60	63	67	72	76
Hard	67	10	66	5	64	66	69	71	72

of career win-loss improved the prediction accuracy and discrimination of the FiveThirtyEight model predictions, resulting in performance that was the closest to the bookmaker predictions on all dimensions of performance. The performance of the FiveThirtyEight model with career-to-date data differed from the BCM by 3–7% on measures of accuracy and by 20% on discriminatory ability.

Differences in the accuracy of predictive methods were found according to player rank, tournament level, and surface characteristics (Table 5). The largest differences were found between matches in which the highest-ranked player was ranked 30 or higher versus matches of lower-ranked opponents. All models performed significantly better at predicting the outcomes of matches of higher-ranked players by 10–20 percentage points. Though smaller in magnitude, it was also found that predictions for Grand Slam matches were more accurate than predictions for lower-tiered tournaments; and predictions for grass and hard court tournaments tended to be more accurate than for clay court tournaments.

Predictions for the highest-ranked players based on the FiveThirtyEight model had good accuracy, even with 1 year of data (Table 5). In fact, the additional career data did not increase the accuracy for top 30 players, which was 75% with 1 year or career-to-date win-loss data, but did improve the accuracy of predictions for lower-ranked

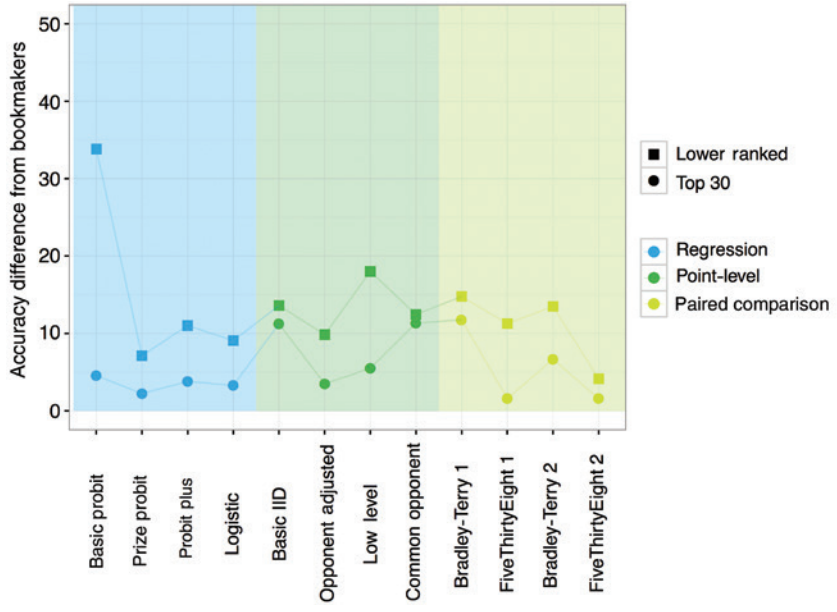

Figure 2: Percentage difference in prediction accuracy from the bookmakers consensus model according to the ranking of the highest-ranked player in the match.

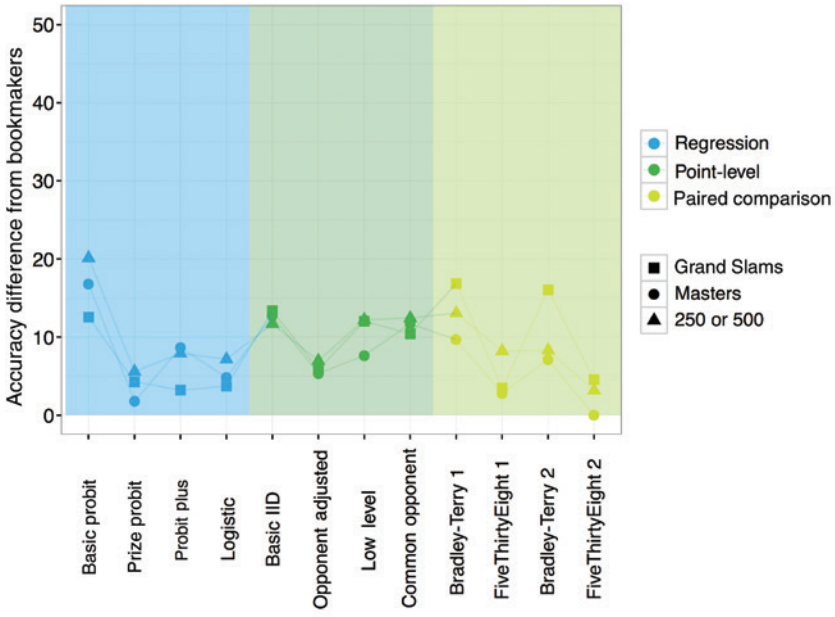

Figure 3: Percentage difference in prediction accuracy from the bookmakers consensus model according to the tournament level, from the highest tier (Grand Slams) to the lowest tier (250 and 500 Series).

players (increased from 59% to 64%). Among methods using 1 year of performance data, FiveThirtyEight model predictions had accuracy that was most consistent with the BCM predictions, and the agreement in accuracy was even greater with career-to-date data (FiveThirtyEight 2).

Figure 2 shows the prediction accuracy of each method by player ranking, using the level of accuracy of the BCM as the benchmark for comparison. The FiveThirtyEight model had nearly identical performance to the BCM for top 30 players the most similar performance to the BCM for both ranking groups, but was only 88% as accurate for lower-ranked players with 1 year of performance data and 96% as accurate with career-to-date data. While the comparative accuracy of the regression models with player ranking as a predictor was very similar, there were notable divergences among the point-based models. The Opponent Adjusted model had performance that was most comparable to the regression models. By contrast, the Low Level model was one of the most sensitive to the ranking level of the competitors, differing from the accuracy of the BCM by only 5% for matches with a top 30 player but differing by 18% for all other matches.

The comparative accuracy of the models generally paralleled the BCM across tournament levels (Figure 3). There were two exceptions to this overall trend. The Probit Plus model was the only regression-based model that had excellent accuracy for Grand Slam matches but approximately 10% worse accuracy for lower-tiered tournaments. The Low Level model was unusual among the point-based models in that it differed least from the BCM accuracy for Masters level tournaments. The FiveThirtyEight predictions using 1 year of data were the closest in accuracy to the BCM for Grand Slams and Masters tournaments but were less accurate for lower-tiered tournaments than the Prize Probit model. When career-to-date performance data was used, the accuracy of the FiveThirtyEight predictions for the lower-tiered tournaments improved but the accuracy for the higher-tiered tournaments did not notably change.

5 Discussion

This is the first study to compare the performance of proposed models for forecasting tennis match wins. Using a large set of recent professional tennis matches to test the predictive performance of published approaches, it was found that a predictive method based on Elo ratings was the closest competitor to bookmaker predictions, correctly predicting 70% of match outcomes and outperforming published alternatives that included regression-based models, point-based models, and a Bradley-Terry model. Among the regression models, the inclusion of player ranking as a predictor resulted in the best performance and additional predictors considered by previous authors did not improve performance. Among the point-based models, the approach that adjusts the serve win probability for player and opponent strength, according to how their average performance differs from the average of

the field, had the best predictive performance. While the regression-based and point-based models had comparable accuracy, point-based models generally had better discriminatory ability, putting them on par with the Elo and bookmaker predictions on this dimension of performance. All models were less accurate at predicting the outcomes of matches of lower-ranked players compared to matches of the top players in the sport.

Although no model was able to beat the bookmakers, the Elo model developed by FiveThirtyEight was a close competitor and performed better than all other approaches in terms of accuracy and discrimination. The standard implementation of Elo is based on career-to-date wins and losses, while regression- and point-based models have typically been done with 1 or 2 years of prior performance data. When only 1 year of prior performance data was used, the FiveThirtyEight model performance was comparable to regression models with player ranking but remained the best-performing approach for predicting outcomes of matches of the highest-ranked players. Thus, the career-to-date information contributed to its edge for a subset of matches but not all.

These findings provide further rationale for the wide use of Elo-based predictions in the media and adds tennis to a growing list of sports for which Elo ratings have proven useful (Stefani 2011; Lasek, Szlávik, and Bhulai 2013). Still, the performance of the FiveThirtyEight might seem surprising given that the information it uses is fairly basic. The Elo ratings in this model only consider a player's past wins and losses. However, the dynamic nature of these ratings go further than the other approaches considered in this paper in how they reassess player ability and adjust for opponent strength at the time of the match. Unlike player rankings, that are updated weekly and rely on an arbitrary point system, Elo ratings are updated at the end of each match using a probability-based formula that weighs more recent performance more heavily and credits players for wins against more difficult opponents. This suggests that accounting for recency of play and the quality of opponents are critical elements in predicting the outcome of matches at the elite level with greater accuracy.

A key finding of the validation study was the lack of performance improvement for the most predictor-rich regression models. Indeed, a logistic model with the player and opponent ranking differential as its only predictor performed as well overall as the Prize Probit and Probit Plus models that incorporated additional tournament and player demographic variables. In analyses not shown, the predictor used in the logistic model was fit with a probit form, which did not change the prediction performance, showing that the logistic performance was due to the predictive strength of differential ranking and not the choice of distributional form.

There are several reasons why the difference in player-opponent ranking drives the performance of the regression-based models. Rankings represent a rolling weighted sum of a player's win-loss record in the previous 12 months, where the weights are a scaled point system derived by the tour that attempts to reflect the prestige of a match (Irons, Buckley, and Paulden 2014). Although non-probabilistic in nature, the ranking points are intended to have a high correlation with a player's recent ability on the biggest stages of the calendar. There is also a potential "rich get richer effect" with rankings due to tournament seeding, as tournaments are designed to help the highest ranked players advance, strengthening the correlation between player differences in rank and match wins.

Despite the strength of player ranking in published regression models, the superior performance of the FiveThirtyEight model suggests that alternative measures of player strength might improve the performance of regression methods. In particular, model-based measures of player ability that account for career matches and adjust for opponent difficulty could be promising alternatives to official rankings.

As a class, point-based models generally underestimated the win probability of the higher-ranked player in a match, the one exception being the Opponent Adjusted model. There are two main modeling decisions that determine the performance of the point-based model. First, the choice of model for winning a point on serve; second, the choice of approach for predicting a match win from point wins. All of the point-based methods rely on IID assumptions for predicting match wins. The IID model, which assumes a constant probability of winning on serve during a match, is known to be incorrect but has been argued to be a good approximation (Klaassen and Magnus 2001). However, the IID has not been comprehensively tested on more recent professional matches, which raises the possibility that relaxing the IID model assumptions might reduce the bias found for point-based models.

While the point-based models had, as a group, more evidence of bias, they also exhibited greater discriminatory ability. This suggests the conclusion of a trade-off in bias and discrimination with point-level information. However, the observation that the Opponent Adjusted model had good calibration and discriminatory ability shows that this is not an inherent trade-off of point-based methods, and it should be possible to improve the calibration of this class of methods without sacrificing their discriminatory strengths.

While the Bradley-Terry model had prediction accuracy that was similar to the point-based models, it showed the largest bias and the poorest discriminatory ability of all methods. Further, this bias was not remedied with the inclusion of an additional year of performance data. The Bradley-Terry model is the only approach considered here that utilizes game wins as the primary driver of player ability. Owing to the hierarchical nature of tennis, it is not necessary to win every point, game, or set to win a match. The superior performance of the Elo-based model over the Bradley-Terry model suggests that the focus on game-level measures of performance in place of overall match performance is a less reliable measure of player ability.

Strengths of the present study include the use of an independent validation dataset with many matches on all surfaces and stages of the ATP tour above the Challenger level. However, by focusing on only 1 year of data it is unclear whether these findings can be generalized to past or future generations of players. Another limitation in the generalizability of the findings is that the paper did not consider performance for the Women's Tennis Association. Further, the present paper focused on pre-match predictive performance and did not investigate the advantages of within-match updating, which is a potentially unique strength of point-based models.

This work highlights a number of directions for further research. Proposed improvements of the Elo rating system have been developed but have yet to be applied in tennis (Glickman 1999; Herbrich, Minka, and Graepel 2006). All of the evaluated prediction methods were less accurate at predicting match outcomes for lower-ranked players compared to matches of the best players in the sport. This property could hinder the practical utility of current prediction methods and the extension of these methods to the junior game. Further work is needed to identify stronger predictors of the performance of lower-ranked players. In this regard, it was notable that no published prediction method included information about player mental skills or shot-level characteristics, though these are both thought to be important determinants of match outcomes (Féry and Crognier 2001; Jones 2002). It is an open question whether either of these areas of performance could improve the predictive performance or generalizability of current methods.

The recent media scandal on match-fixing in tennis calls attention to the reality that the performance of prediction methods in the sport is not simply an academic concern. For modelers to ensure that coaches and tennis officials are using the most appropriate available tools when evaluating tennis outcomes, rigorous validation should be a routine part of the development of tennis prediction methods. At present, it can be concluded that some published prediction models are more useful than others and all models have limited utility outside of the highest levels of the sport. The variation in model performance demonstrated in this study emphasizes the importance of comparative validation and the need for continued research to improve forecasting outcomes in tennis.

Acknowledgments: I am grateful to Jeff Sackmann of *Tennis Abstract* and the staff at the ATP who made this research possible by providing large amounts of tennis data to the public.

References

Barnett, T. and S. R. Clarke. 2005. "Combining Player Statistics to Predict Outcomes of Tennis Matches." *IMA Journal of Management Mathematics* 16(2):113–120.

Boulier, B. L. and H. O. Stekler. 1999. "Are Sports Seedings Good Predictors?: An Evaluation." *International Journal of Forecasting* 15(1):83–91.

Bradley, R. A. and M. E. Terry. 1952. "Rank Analysis of Incomplete Block Designs: I. The Method of Paired Comparisons." *Biometrika* 39(3):324–345.

Del Corral, J. and J. Prieto-Rodriguez. 2010. "Are Differences in Ranks Good Predictors for Grand Slam Tennis Matches?" *International Journal of Forecasting* 26(3):551–563.

Elo, A. E. 1978. *The Rating of Chessplayers, Past and Present*. New York: Arco.

Féry, Y.-A. and L. Crognier. 2001. "On the Tactical Significance of Game Situations in Anticipating Ball Trajectories in Tennis." *Research Quarterly for Exercise and Sport* 72(2):143–149.

FiveThirtyEight. 2015. "Serena Williams and the Difference between All-Time Great and Greatest of All Time." http://fivethirtyeight.com/features/serena-williams-and-the-difference-between-all-time-great-and-greatest-of-all-time/. Accessed on February 2, 2016.

Gilsdorf, K. F. and V. A. Sukhatme. 2008. "Testing Rosen's Sequential Elimination Tournament Model Incentives and Player Performance in Professional Tennis." *Journal of Sports Economics* 9(3):287–303.

Glickman, M. E. 1999. "Parameter Estimation in Large Dynamic Paired Comparison Experiments." *Applied Statistics* 48(3):377–394.

Herbrich, R., T. Minka, and T. Graepel. 2006. "Trueskill(tm): A Bayesian Skill Rating System." In *Advances in Neural Information Processing Systems*, pp. 569–576.

Irons, D. J., S. Buckley, and T. Paulden. 2014. "Developing an Improved Tennis Ranking System." *Journal of Quantitative Analysis in Sports* 10(2):109–118.

Jones, G. 2002. "What is this Thing called mental toughness? An investigation of Elite Sport Performers." *Journal of Applied Sport Psychology* 14(3):205–218.

Klaassen, F. J. and J. R. Magnus. 2001. "Are Points in Tennis Independent and Identically Distributed? Evidence from a

Dynamic Binary Panel Data Model." *Journal of the American Statistical Association* 96(454):500–509.

Klaassen, F. J. and J. R. Magnus. 2003. "Forecasting the Winner of a Tennis Match." *European Journal of Operational Research* 148(2):257–267.

Knottenbelt, W. J., D. Spanias, and A. M. Madurska. 2012. "A Common-Opponent Stochastic Model for Predicting the Outcome of Professional Tennis Matches." *Computers & Mathematics with Applications* 64(12):3820–3827.

Lasek, J., Z. Szlávik, and S. Bhulai. 2013. "The Predictive Power of Ranking Systems in Association Football." *International Journal of Applied Pattern Recognition* 1(1):27–46.

Leitner, C., A. Zeileis, and K. Hornik. 2009. "Is Federer Stronger in a Tournament without Nadal? An Evaluation of Odds and Seedings for Wimbledon 2009." *Research Report Series/Department of Statistics and Mathematics* 94.

McHale, I. and A. Morton. 2011. "A Bradley-Terry Type Model for Forecasting Tennis Match Results." *International Journal of Forecasting* 27(2):619–630.

Newton, P. K. and J. B. Keller. 2005. "Probability of Winning at Tennis I. Theory and Data." *Studies in Applied Mathematics* 114(3):241–269.

Pencina, M. J., R. B. D'Agostino, and R. S. Vasan. 2008. "Evaluating the Added Predictive Ability of a New Marker: From Area under the Roc Curve to Reclassification and Beyond." *Statistics in Medicine* 27(2):157–172.

R Core Team. 2015. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing.

Shin, H. S. 1993. "Measuring the Incidence of Insider Trading in a Market for State-Contingent Claims." *The Economic Journal* 103(420):1141–1153.

Spanias, D. and W. J. Knottenbelt. 2012. "Predicting the Outcomes of Tennis Matches using a Low-Level Point Model." *IMA Journal of Management Mathematics* 24(3):311–320.

Stefani, R. 2011. "The Methodology of Officially Recognized International Sports Rating Systems." *Journal of Quantitative Analysis in Sports* 7(4):10.

Štrumbelj, E. 2014. "On Determining Probability Forecasts from Betting Odds." *International Journal of Forecasting* 30(4):934–943.

Turner, H. and D. Firth. 2012. "Bradley-Terry Models in R: The BradleyTerry2 Package." *Journal of Statistical Software* 48(9):1–12.

Yuan, L.-H., A. Liu, A. Yeh, A. Kaufman, A. Reece, P. Bull, A. Franks, S. Wang, D. Illushin, and L. Bornn. 2015. "A Mixture-of-Modelers Approach to Forecasting NCAA Tournament Outcomes." *Journal of Quantitative Analysis in Sports* 11(1):13–27.

Supplemental Material: The online version of this article (DOI: 10.1515/jqas-2015-0059) offers supplementary material, available to authorized users.