SEONZ
SPORT & EXERCISE SCIENCE
NEW ZEALAND

The Journal of Sport and Exercise Science, Vol. 8, Issue 1, 43-64 (2024)

www.jses.net

JSES
ISSN: 2703-240X

An expected wins approach using Fisher's Exact Test to identify the bogey effect in sports: An application to tennis

Rory P. Bunker[1*], Calvin Yeung[1], Keisuke Fujii[1,2,3]

[1]*Graduate School of Informatics, Nagoya University, Japan*
[2]*RIKEN Advanced Intelligence Project, Japan*
[3]*PRESTO Japan Science and Technology Agency, Japan*

ARTICLE INFO

Received: 31.10.2023
Accepted: 05.06.2024
Online: 22.11.2024

Keywords:
Tennis
Bogey
Betting odds
Fisher's Exact Test
Elo ratings

ABSTRACT

In sports, so-called "bogey" players or teams tend to beat a particular opposition with regularity despite being of similar ability. Although the existence of bogey players is widely discussed and debated among sports fans and the media, methods that could be used to identify the bogey effect have received little attention in the literature. This study proposes a statistical procedure to identify bogey players using a publicly available men's Association of Tennis Professionals (ATP) and Women's Tennis Association (WTA) dataset, which is also split into Grand Slam and non-Grand Slam matches. The proposed method iterates over all unique player pairs and applies Fisher's Exact Test to a contingency table containing expected wins and actual win distributions for historical matches between the players in the player pair. To compute expected wins, betting odds- or Elo ratings-implied probabilities for each player are aggregated over all matches between the player pair for each player. If the Fisher's Exact Test result is statistically significant (that is, actual wins and expected wins do not follow the same distribution), and the expected wins and actual win counts are contradictory, we suggest that the bogey effect exists between the two players. The obtained results suggest that the bogey player effect exists in professional tennis but is rare (and even rarer in Grand Slams), and expected wins obtained using betting odds-implied win probabilities more closely matched actual wins than Elo ratings, which resulted in fewer bogey player pairs being identified with betting odds. The number of bogey player pairs identified is intuitively found to be inversely related to the predictability of matches.

1. Introduction

The existence of the bogey effect, in which players (or teams in the case of team sports) tend to beat a particular opposition with regularity despite being of similar ability, is widely discussed among sports fans and the media. For instance, France at one point was considered the bogey team of New Zealand in Rugby World Cup tournaments (Bruce, 2014), while in soccer, Italy was considered the bogey team of Germany (Wilms, 2013). While tennis, which is considered in the current study, appears to have had less attention in the media than soccer, there has been some discussion of potential bogey players in online forums and in articles (Niall, 2013; Wood, 2017). Loosely speaking, a bogey player/team, or "Angstgegener" (translated as "feared opponent") as it is known in the German language, tends to habitually beat another specified player/team despite appearing to be of equal, or even lesser, strength on paper. Although the bogey phenomenon has been mentioned in a small number of academic studies in, for example, education (Bruce, 2014) and sociology (Chiweshe, 2021; Poulton, 2004), and in doctoral theses (Awerbuch, 2009; Wilms, 2014), it has been largely unexplored in the sports science and sports statistics disciplines.

*Corresponding Author: Rory P. Bunker, Graduate School of Informatics, Nagoya University, Japan, rorybunker@gmail.com

Related to the bogey effect are the concepts of streaks, form, hot hand, stability (non-stationarity), autocorrelation, and "hot and cold nights". Streaks can be considered both in terms of individual player actions (e.g., home runs in baseball, three-pointers in basketball) or match-winning streaks. Form, also known as the "hot hand" phenomenon in basketball, assumes that future outcomes can be determined, at least partly, based on the most recent outcomes and that players or teams having successful streaks impact their future successes (Ayton & Fischer, 2004; Bar-Eli, Avugos, & Raab, 2006). Carlson and Shu (2007) found that across five diverse studies, including one related to shooting in basketball, the third repeated event within a sequence is critical to the subjective belief that a streak is happening. The most common techniques used in such analyses have been Wald-Wolfowitz Runs Tests and autocorrelation tests (Carlson & Shu, 2007; Peel & Clauset, 2015; Raab, 2012; Stone, 2012). Hales (1999) argued that autocorrelation, the correlation between the outcomes of consecutive events, and non-stationarity, the probability of success fluctuating over time, should be considered separately because the two concepts represent different underlying mechanisms of the hot hand effect. Specifically, (positive) autocorrelation, which is often measured using the correlation coefficient or the runs test, suggests that success in one event increases the likelihood of success in a subsequent event, thus indicating a hot hand effect. If autocorrelation is present, it may indicate that performance is influenced by recent success/failure, and a measure of positive association between shot outcomes may be appropriate. Non-stationarity may be caused by several factors (e.g., form, fatigue, or external factors such as opponent performance) and the chi-square test can be used to detect changes in the probability of success over time. Non-stationarity indicates the existence of fluctuations in player performance, and a time-varying ability parameter (e.g., time-varying Bradley-Terry parameters as per Cattelan, Varin, & Firth, 2013) may be appropriate to model the data. Steeger, Dulin, and Gonzalez (2021) distinguished between streaks and momentum; the former referring to observed sequences of events each of which may or may not have dependence between them, while momentum suggests that a dependence exists between events that are similar. In their seminal paper, Gilovich, Vallone, and Tversky (1985) refer to basketball shooters having "hot" and "cold" nights (i.e., strong and poor performance, respectively), and analysed whether stability exists across matches in terms of shooters having more hot or cold nights than would be expected by chance; that is, how the variability in match shooting percentages that is observed compares with the expected variability according to a player's record overall. The "gambler's fallacy", also known more generally as negative recency, is the belief that in sequences comprised of binary random events, runs of a specific outcome will be balanced by corrective action; that is, a tendency for the other outcome (Estes, 1964 and Ayton & Fisher, 2004, as cited in Steeger, Dulin, & Gonzalez, 2021). Positive recency, also known as the hot-hand fallacy, is the inclination to predict future outcomes the same as recent outcomes (Ayton & Fischer, 2004, as cited in Steeger, Dulin, & Gonzalez, 2021). Baboota and Kaur (2019) engineered streak and time-weighted streak features for soccer match result prediction that consider the results of a single, and a form feature that considers the results between specific pairs of teams. At first glance, it can be tempting to attribute a long streak of wins to a particular team to the bogey effect. Tottenham, for example, won only one game out of 37 against Chelsea between 1990 and 2006, and Watford did not beat Manchester City in the 30-year period from 1989 to 2019. However, it may have been the case that these results may all have been expected; thus, the question then becomes how expectedness can be accounted for. The media tend to use the "bogey" term relatively loosely, without providing additional contextual information that would help determine whether such results are actually unexpected. This study attempts to clarify this by introducing a bogey player identification method that uses Fisher's Exact Test to determine whether the match results between a particular pair of tennis players deviate from what would be expected, given the betting odds or Elo ratings of the two players in each player pair.

Prior studies that have focused on bogey effect identification specifically have attempted to use statistical techniques including the Wald-Wolfowitz runs test (Bunker, 2022) and the Aylmer test (Hankin & Bunker, 2016), using data from Tennis and Rugby League. Using runs tests, however, also tended to identify result sequences evenly split between unexpected wins and unexpected losses, which is not indicative of the bogey effect, and the Aylmer test, since it was applied to the match results of all pairs simultaneously, was subject to the multiple comparisons issue. In the current study, betting odds, which were used to identify unexpected results in Bunker (2022), as well as Elo ratings are used to compute expected wins for each player pair. In particular, (implied) win probabilities for both betting odds and Elo ratings are aggregated to calculate expected wins for each player in the player pair. Leitner, Zeileis, and Hornik (2009) proposed a bookmaker consensus model that aggregates bookmaker expectations into a prediction for tennis match results. In the proposed method, the odds that we use in the dataset already represent the average across multiple bookmakers, and we aggregate the odds-implied win probabilities of each player in the player pair. Fisher's Exact Test (FET) is then applied to a contingency table consisting of the computed expected wins distribution and the actual win distribution for the player pair. The bogey effect is considered to exist between two players if FET yields a statistically significant result, and the expected wins and actual wins contradict. The method proposed circumvents the multiple comparisons problem since it is applied iteratively to each unique player pair, and match results and betting odds are independent since odds are determined by bookmakers prior to matches and thus represent their assessment of the probabilities of the match outcomes, and bets placed on certain outcomes do not affect match results, while match result is determined purely by the teams' or players' in-match performance. While odds can provide insights into bookmaker expectations, in the absence of match-fixing, they do not have any effect on match results. As well as proposing a novel method for bogey effect identification, the method is demonstrated on publicly available datasets consisting of 33,976 men's Association of Tennis Professionals (ATP) matches from 2005 to 2020 and 27,094 Women's Tennis Association (WTA) matches from 2007 to 2020, as well as subsets of the original datasets comprising only Grand Slam and non-Grand Slam matches. The analysis presented in this paper is also relevant in the context of sports psychology in that some players (teams) struggling more against a particular opposition player (team) may be a psychological phenomenon.

We hypothesise that the betting odds are more accurate for Grand Slam matches since, first, bookmakers would carry out more research and perform more modelling before setting the

odds for the match outcome, and second, the betting volume is likely to be higher on Grand Slam matches compared to non-Grand Slam matches, and this additional information from betting volume increases the accuracy of the odds. Since betting odds accuracy is inversely related to the number of bogey players identified by odds in our proposed method, both of these factors would contribute to fewer bogey players being identified.

On the other hand, Elo ratings are affected only by match results and are updated dynamically over time based on these match outcomes. Form and strength would also be accounted for by bookmakers, and they may even use Elo ratings or other ratings in their models to set odds, however, this is only a subset of the information they would use to set betting odds.

Through a regression-based analysis of 51,881 tennis matches, Barrutiabengoa, Corredor, and Muga (2022) found that, even after controlling for surprise factor uncertainty and the amount of media attention, the prices that bookmakers quote are higher for women's matches than men's, which suggests that the betting volume (and, therefore, betting odds accuracy) on women's matches could be lower compared to men's. Vaughan Williams, Liu, Dixon, and Gerrard (2021) compared the performance of betting odds, rankings, standard and surface-specific Elo ratings, and weighted rating composites, including and excluding the betting odds, in predicting men's and women's professional tennis matches and found that betting odds performed well in general, and standard Elo ratings performed well for women's tennis. The authors found that Elo and betting odds performed better than rankings, which supports the use of these two variables in our proposed method.

Kovalchik (2016) found that the accuracy of predictive models when predicting match results is markedly different for lower-ranked and top-ranked players, finding that match outcome prediction models are 10% to 20% less accurate for matches among lower-ranked players than matches among top-ranked players. Yue, Chou, Hsieh, and Hsiao (2022) showed that win probability with respect to ranking difference fluctuates to a greater degree (i.e., predictability decreases) when the ranking difference increases (Figure 2, Yue et al., 2022) due to smaller number of samples at larger ranking differences (Figure 3, Yue et al., 2022). The findings of Yue et al. (2022) imply then that matches between top-ranked players have a small ranking difference (Elo rating difference) and are thus easier to predict, while matches between top-ranked and lower-ranked players, which have a large ranking (Elo rating) difference, are more difficult to predict. In this study, we partition the original dataset into Grand Slam and non-Grand Slam matches. Grand Slams generally consist of matches among top-ranked players, thus, based on the findings of Kovalchik (2016) and Yue et al. (2022), we would expect our proposed method to identify fewer bogey player pairs for the Grand Slam match dataset compared to the non-Grand Slam match dataset.

Elo ratings, which are updated over time based on historical match results, do not incorporate factors such as the court surface and what hand the players play with, while betting odds do incorporate such factors through the odds set by bookmakers supplemented by fan knowledge reflected in betting volume, which bookmakers use to tweak their initial odds. We would expect that using Elo ratings in the proposed method will generally identify more bogey players than betting odds.

2. Methods

2.1. Data

2.1.1. Datasets

Publicly available data from professional ATP (men's) and WTA (women's) tennis was sourced for this study. In particular, we use the same datasets used by Angelini, Candila, and De Angelis (2022), which were originally sourced from the website tennis-data.co.uk. Among many other variables, the dataset contains match data from ATP and WTA tournaments and Grand Slams, as well as bookmaker odds, player rankings, and ATP/WTA player points. The datasets contain 33,976 men's ATP matches from 2005 to 2020 and 27,094 women's WTA matches from 2007 to 2018. The ATP and WTA datasets were downloaded as RData files from "Appendix C. Supplementary materials" in the paper by Angelini, Candila, and De Angelis (2022). An R script was then created to clean the data using the `clean()` function in the welo R package (Candila, 2023) and to convert and export the cleaned data to a CSV file. The datasets were passed through the welo R package's `clean()` function, which reduced the number of matches in the final dataset. This data-cleaning function reduced the original number of matches from 38,868 to 33,976 for the ATP dataset and from 30,706 to 27,094 for the WTA dataset.[1]

For further analysis, the ATP and WTA datasets consisting of all matches were each divided into two additional datasets consisting of Grand Slam matches and non-Grand Slam matches (Figure 1), and the proposed method will also be applied to these four additional datasets.

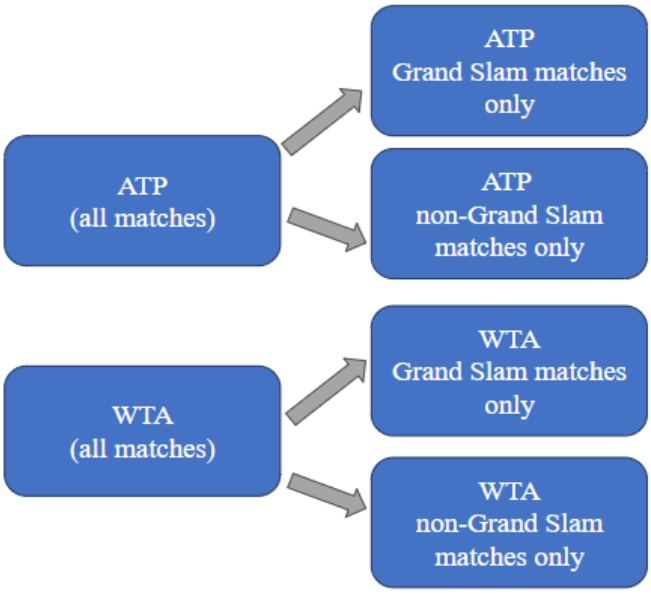


Figure 1: Datasets used upon which the proposed method is applied.

[1] The `clean()` function performs ten cleaning operations, which are described under the clean function details on page 4 in the welo package manual: https://cran.r-project.org/package=welo. The default options of the function were used.

2.1.2. Descriptive statistics

To compare the temporal variability of betting odds (winner, loser, and combined) and Elo ratings, we plot the coefficient of variation (CV), which is calculated by dividing the standard deviation by the mean and is indicative of the level of dispersion around the mean and accounts for the different scale of variance/standard deviation of odds and Elo ratings, of each for both the ATP and WTA datasets (Figure 2A and Figure 3A). As can be seen, the variability of Elo ratings is much lower overall compared to betting odds. It is notable from Figure 3 that the variability of betting odds for women's tennis declined over the time period 2005 and 2020. The variability of Elo ratings declined for both ATP and WTA in the latter period, however, WTA Elo rating variability declined from 2013 onwards, whereas ATP Elo rating variability declined later—from 2016 onwards. The mean Elo rating generally increased over the time period for both ATP and WTA. Another noticeable feature of Figures 2C and 3C is that the mean betting odds and betting odds CV for the WTA, especially for the loser odds, declined over the sample period. On the other hand, the mean betting odds and betting odds CV for the ATP had no discernible trend over the sample period. This perhaps indicates that bookmakers, on account of an evening in the level of competition in the WTA, as evidenced by the decline in Elo rating CV from 2013 onwards, began to lower the price on the likely losers in WTA matches. The remaining figures for the Grand Slam and non-Grand Slam datasets for the ATP and WTA are in the Supplemental material (Supplementary Figures 1 to 4).

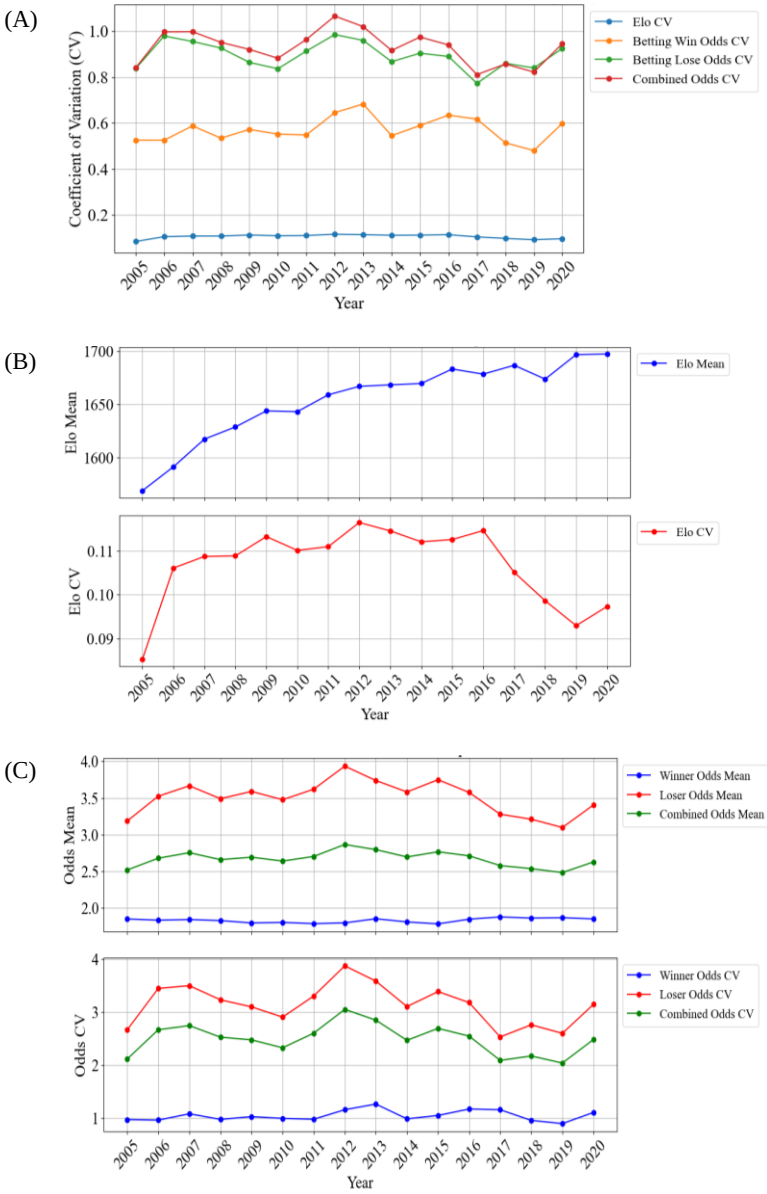


Figure 2: Coefficient of Variation (CV) of Betting Odds and Elo Ratings for each year (A), as well as the more granular view of the mean and CV for each year of Elo rating and Betting Odds (B and C, respectively) for the ATP dataset.

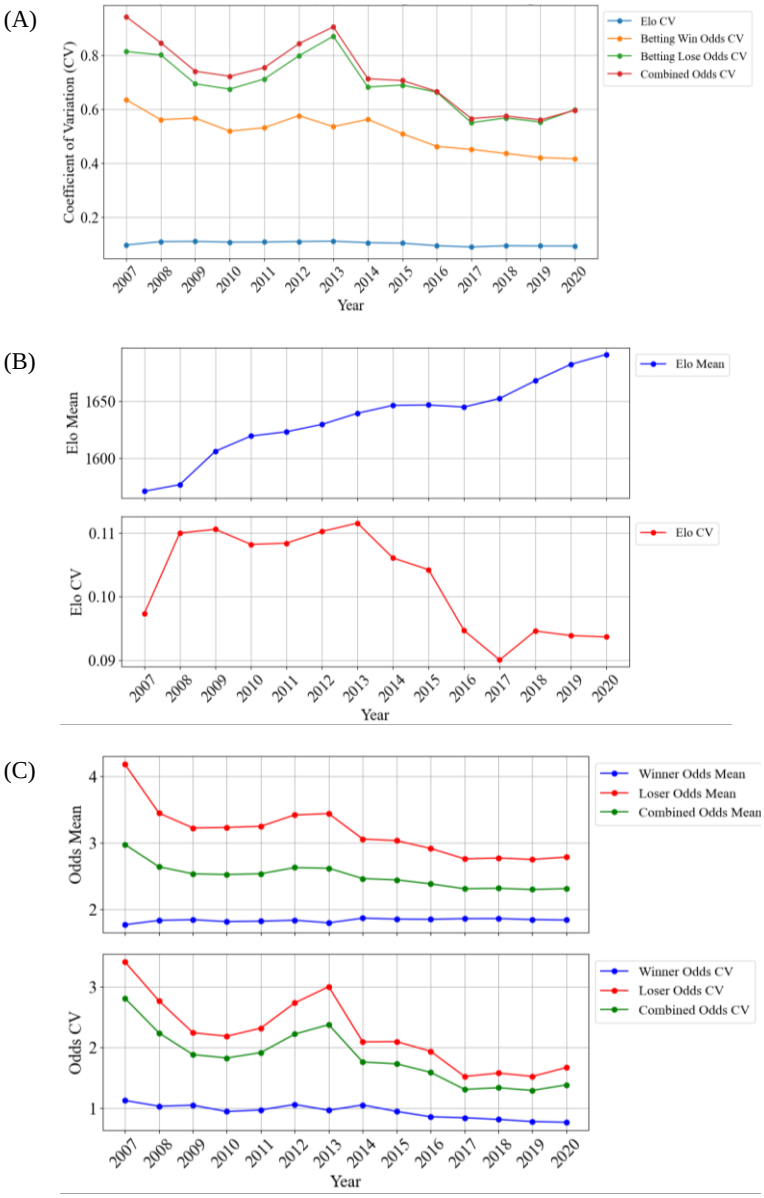


Figure 3. Coefficient of Variation (CV) of Betting Odds and Elo Ratings for each year (A), as well as the more granular view of the mean and CV for each year of Elo rating and Betting Odds (B and C, respectively) for the WTA dataset.

2.2. Proposed method

The proposed method consists of three steps, which are depicted in Figure 4 and are outlined in the following three subsections.

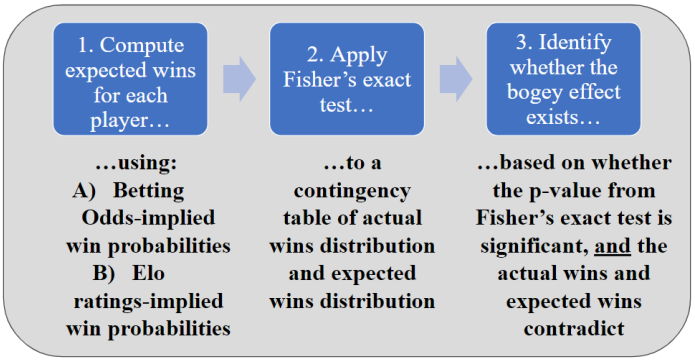


Figure 4: Three steps of the proposed method.

2.2.1. Computing expected wins

As mentioned, expected wins are determined using two approaches: using betting odds- and Elo ratings-implied probabilities. So that the win probabilities for each player in each match add to one, the betting odds-implied win probabilities are obtained by dividing the reciprocal of the decimal betting odds by a normalization factor (also known as the over-round), which is simply the sum of the two decimal betting odd reciprocals. Mathematically, the win probabilities for each player, i and j, in match t, can thus be derived from the decimal betting odds by:

$$p_t^i = \frac{\frac{1}{o_t^i}}{\frac{1}{o_t^i}+\frac{1}{o_t^j}} \quad (1)$$

and

$$p_t^j = \frac{\frac{1}{o_t^j}}{\frac{1}{o_t^i}+\frac{1}{o_t^j}} \quad (2)$$

where O_t^i and O_t^j are the decimal betting odds for a player i win and player j win, respectively, in match t. The denominator in each of these two expressions is the normalization factor (over-round).

The second way expected wins are determined is based on Elo ratings-implied win probabilities. It can be shown that the Bradley-Terry strength/ability parameter can be expressed as a function of the Elo rating (Coulom, 2007). In the Bradley-Terry model, the probability that i beats j is given by:

$$\frac{s_i}{s_i+s_j} \quad (3)$$

where the Elo rating of i is defined as

$$R_i = 400log_{10}(s_i) \quad (4)$$

or equivalently

$$s_i = 10^{\frac{R_i}{400}} \quad (5)$$

Therefore, the Elo-implied estimated win probability for player i over player j in match t is given by:

$$p_t^i = \frac{10^{\frac{R_t^i}{400}}}{10^{\frac{R_t^i}{400}}+10^{\frac{R_t^j}{400}}} \quad (6)$$

Similarly, the Elo-implied estimated win probability for player j over player i in match t is given by:

$$p_t^j = \frac{10^{\frac{R_t^j}{400}}}{10^{\frac{R_t^i}{400}}+10^{\frac{R_t^j}{400}}} \quad (7)$$

Note that, unlike betting odds, normalization is not required as it is already the case with Elo-rating estimated probabilities that $p_t^i + p_t^j = 1$. The before- and after-match Elo ratings were computed by passing the cleaned ATP and WTA datasets (see 2.1. Data) to the `welofit()` function in the welo package (the default options were used). For estimating win probabilities, we use the before-match Elo ratings. Since the betting odds and Elo ratings already account for historical information, it is reasonable to ignore the temporal order of the match result data (i.e., we do not need to consider the match results as a temporally ordered sequence).

2.2.2. Fisher's Exact Test

A Fisher's Exact Test is applied to compare each pair of players' expected wins distribution with their actual match result distribution. The bogey effect is assumed to represent a violation of expectation. The obtained expected wins distribution for each player in each player pair comprises one part of the contingency table to which the Fisher's Exact Test, which has the advantage of being able to be used for small sample sizes, is applied. The expected wins for each player in a given player pair are obtained by simply summing the estimated win probabilities across all matches they have played against that specific player, which represents the expected number of matches each player should win given the betting odds/Elo ratings distribution. The other part of the contingency table is the actual win distribution between the two players. The null hypothesis, H0, and alternative hypothesis, H1, are described as follows:

H_0: The expected and actual win distributions are the same, or there is no significant difference between them.
H_1: The expected and actual win distributions are not the same, or there is a significant difference between them.

The contingency table for the Fisher exact test for one pair of players, player i and player j, is shown in Table 1. In Table 1, N represents the total number of historical matches played between player i and player j. Thus, $\sum_{t=1}^{N} p_t^i$ represents the aggregation of the betting odds- or Elo-implied win probabilities for player i over all historical matches against player j. Similarly, $\sum_{t=1}^{N} p_t^j$ represents the aggregation of the betting odds (or Elo) implied win probabilities for player j over all historical matches against player i. The cell a_t^i in Table 1 is a binary variable that takes the value of 1 if player i won against player j in match t and 0 otherwise, and analogously, a_t^j is a binary variable that takes the value of 1 if player j beat player i in match t and 0 otherwise. Therefore, $\sum_{t=1}^{N} a_t^i$ represents the total number of actual wins player i has had over player j in their N past matches and $\sum_{t=1}^{N} a_t^j$ the total number of actual wins player j has had over player i in their N past matches.

Table 1: Contingency table for Fisher's Exact Test for a given player pair over their *N* historical matches.

	Betting Odds- or Elo ratings-implied win probabilities	Actual match result
Player i wins	$\sum_{t=1}^{N} p_t^i$	$\sum_{t=1}^{N} a_t^i$
Player j wins	$\sum_{t=1}^{N} p_t^j$	$\sum_{t=1}^{N} a_t^j$

2.3. Bogey effect identification

To identify the bogey effect between a pair of players, the method iteratively computes the Fisher's Exact Test *p*-values for each player pair. Having a $p \leq \alpha$, where α is a specific significance level (in this study we consider $\alpha = 0.05$ and $\alpha = 0.1$, which means that we have 95% or 90% confidence that a bogey player pair exists), is a necessary but not sufficient condition for the bogey effect to exist. The *p*-value is calculated from the Fisher's Exact Test that is applied to the contingency table structure for the *N* historical matches between a particular player pair as per Table 1. If Fisher's Exact Test yields a statistically significant result *and* the expected wins and actual wins contradict—that is, player A was expected to win more matches than player B but actually player A won fewer, or player B was expected to win more matches than player A but actually player B won fewer—we suggest that the bogey effect exists between the two players. Using the notation in Table 1, we suggest that the bogey effect exists between a pair of players player i and player j, based on their N past matches, if the following holds true:

$$\begin{aligned} &\text{IF} \quad p \leq \alpha \\ &\text{AND} \quad [(\textstyle\sum_{t=1}^{N} p_t^i > \sum_{t=1}^{N} p_t^j \text{ AND } \sum_{t=1}^{N} a_t^j > \sum_{t=1}^{N} a_t^i) \\ &\text{OR} \quad (\textstyle\sum_{t=1}^{N} p_t^j > \sum_{t=1}^{N} p_t^i \text{ AND } \sum_{t=1}^{N} a_t^i > \sum_{t=1}^{N} a_t^j)] \end{aligned}$$

3. Results

3.1. Number of bogey player pairs identified with the proposed method using Elo ratings and betting odds for each dataset

A summary of the results for all datasets, at the 90% and 95% level of confidence, is shown in Table 2. An initial observation from Table 2 is that, regardless of the level of significance used and whether betting odds or Elo ratings are used, the number of bogey player pairs identified is very small relative to the number of player pairs in the datasets. This suggests that the bogey player effect is very rare in professional tennis.

For the whole ATP dataset, of the 18,241 distinct ATP player pairs, 4 and 18 significant bogey pairs were identified at the 90% significance level using betting odds and Elo ratings, respectively. For the whole WTA dataset, of the 15,844 distinct WTA player pairs, 7 and 13 significant bogey pairs were identified using betting odds and Elo ratings, respectively, at the 90% level of significance.

Table 2: Summary of the number of significant player pairs with the bogey effect identified for each of the six datasets, using aggregated betting odds- and Elo ratings-implied probabilities for computing expected wins, at the 95% and 90% significance levels.

		95% level of significance		90% level of significance	
	Player pairs (*n*)	Betting Odds	Elo Ratings	Betting Odds	Elo Ratings
All					
ATP	18,241	0	3	4	18
WTA	15,844	1	2	7	13
Grand Slam					
ATP	5,485	0	0	1	0
WTA	5,075	0	0	0	0
Non-Grand Slam					
ATP	15,735	0	3	6	12
WTA	13,372	1	2	5	8

Notes: The player pairs that are significant at the 95% significance level are also significant at the 90% significance level. For example, for ATP data with Elo ratings, the 18 significant bogey player pairs at the 90% level of significance also include the 3 player pairs that are significant at the 95% level.

Since the total number of player pairs differs across datasets (see the third column in Table 2), in order to make a like-for-like comparison between datasets, the number of bogey player pairs identified is scaled by the total number of player pairs in the dataset in Figure 5. In particular, Figure 5 shows the number of bogey effect player pairs identified for each dataset using betting odds- and Elo ratings-implied probabilities, as a percentage of the total number of player pairs in each dataset, at the 95% (Figure 5A) and 90% (Figure 5B) significance levels.

Some observations can be made from Figure 5 and Table 2. The proposed method with betting odds-implied probabilities obtained fewer bogey player pairs (as a percentage of total player pairs) for all datasets apart from the ATP Grand Slam match dataset. Across the ATP Grand Slam and WTA Grand Slam datasets with both Elo ratings and betting odds-implied probabilities used to compute expected wins, and the ATP Grand Slams dataset with betting odds-implied probabilities used, only one bogey player pair was identified. These results lend support to what was expected: that Grand Slams are closely followed by bookmakers and thus odds are set accurately, and the small differences in Elo ratings between matches among generally top-ranked players at Grand Slam tournaments result in fewer bogey player pairs being identified compared to in the non-Grand Slam match dataset. While the proposed method using betting odds identified relatively fewer bogey player pairs in men's than women's tennis in the all-match datasets, this wasn't the case with the Grand Slam or non-Grand Slam subsets. Somewhat consistent with Vaughan Williams et al. (2021), Elo ratings appear to perform well in the WTA since there were relatively fewer bogey player pairs identified in the WTA using Elo ratings compared to the ATP.

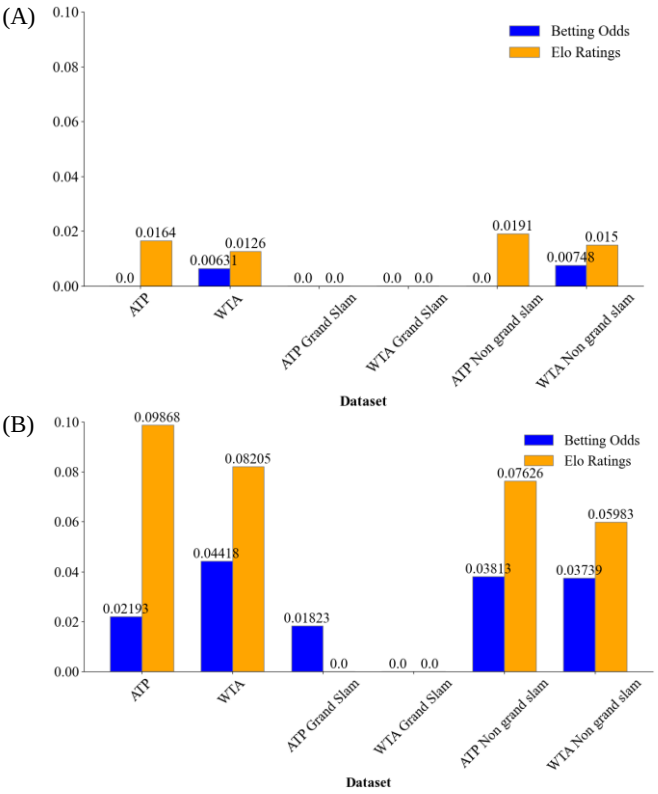


Figure 5: The number of bogey effect player pairs identified for each dataset using betting odds- and Elo ratings-implied probabilities, as a percentage of the total number of player pairs in each dataset, at the 95% (A) and 90% (B) significance levels (data from Table 2).

3.2. Visualising obtained bogey player pairs for a particular dataset

Since the order of the players in each player pair is not material, we visualise bogey player pairs on the absolute difference in actual wins against the absolute difference in expected wins. The size of these points is scaled based on the FET *p*-value, and significant player pairs based on the FET that are not bogey player pairs can be distinguished by shape. Figures 6 to Figure 9 depict the significant bogey (and significant non-bogey) player pairs in this manner for the all-match ATP and WTA datasets, obtained with Elo ratings- and betting odds-implied probabilities (Supplementary Figures 5 to 8; Supplementary Tables 6 to 9 show the same plots and corresponding data for the Grand Slam and non-Grand Slam datasets for the ATP and WTA). These figures correspond to the data in Supplementary Tables 1, 2, 3, and 4 in the Supplemental material.

In general, Figures 6 to Figure 9 exhibit a cluster of points towards the bottom-left hand corner with an absolute actual win difference between the players in the bogey player pair of around two and an absolute expected wins difference of around three. There is often an additional cluster of points with roughly the same absolute expected wins difference but higher absolute actual wins difference values. In all figures there was also an outlier player pair with higher absolute actual and expected wins differences, however, these were generally significant but not identified as a bogey player pair since their actual and expected wins did not contradict.

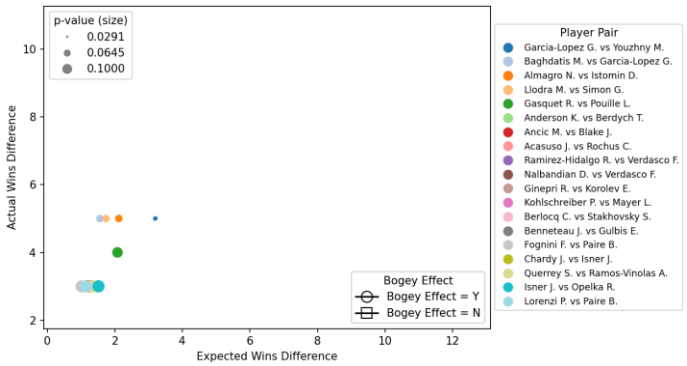


Figure 6: Visualizing, on axes representing the absolute difference in actual wins and expected wins for the player pair, the statistically significant (with at least 90% confidence) and bogey player pairs from the ATP tennis dataset, with aggregated Elo ratings-implied probabilities used to compute expected wins. The data used to create this plot is presented in Supplementary Table 1 in the Supplemental material.

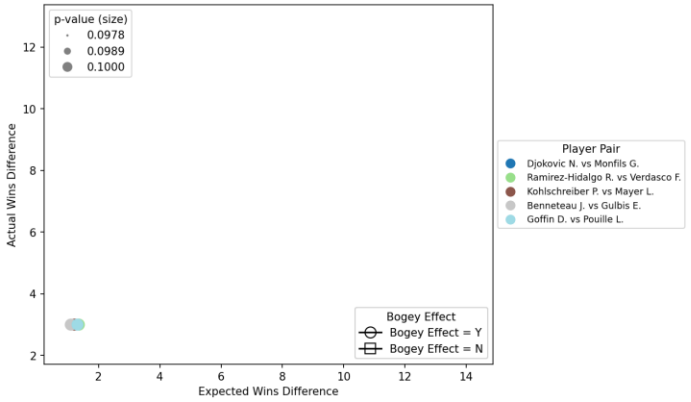


Figure 7: Visualizing, on axes representing the absolute difference in actual wins and expected wins for the player pair, statistically significant (with at least 90% confidence) and bogey player pairs from the ATP tennis dataset, with aggregated betting odds-implied probabilities used to compute expected wins. The data used to create this plot is presented in Supplementary Table 2 in the Supplemental material.

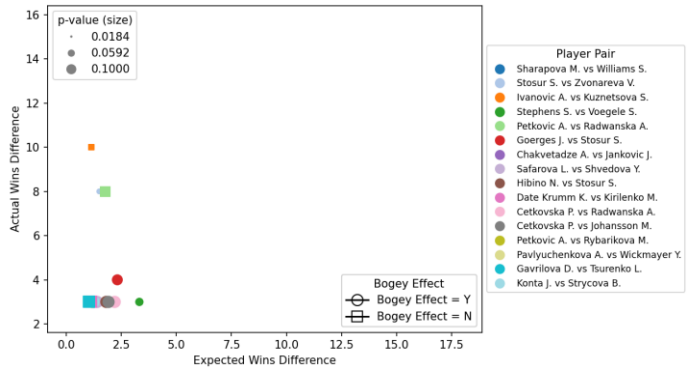


Figure 8: Visualizing, on axes representing the absolute difference in actual wins and expected wins for the player pair, statistically significant (with at least 90% confidence) bogey player pairs from the WTA tennis dataset, with aggregated Elo ratings-implied probabilities used to compute expected wins. The data used to create this plot is presented in Supplementary Table 3 in the Supplemental material.

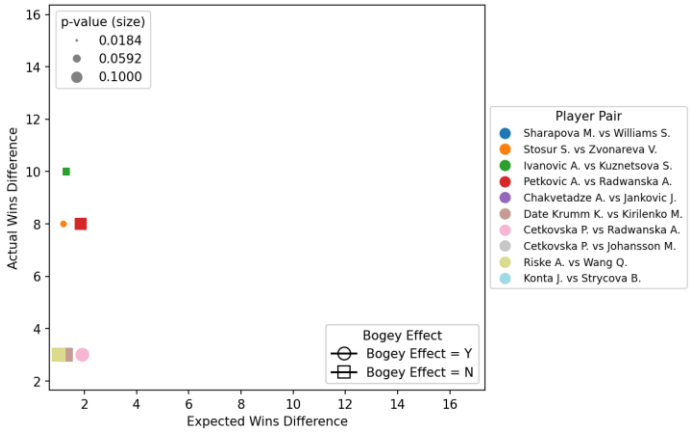


Figure 9: Visualizing, on axes representing the absolute difference in actual wins and expected wins for the player pair, statistically significant (with at least 90% confidence) bogey player pairs from the WTA tennis dataset, with aggregated Betting odds-implied probabilities used to compute expected wins. The data used to create this plot is presented in Supplementary Table 4 in the Supplemental material.

3.3. Analysing the overlap in the bogey player pairs identified by Elo ratings and betting odds

Figure 10A shows the number of bogey player pairs identified by betting odds based on whether they were also identified by Elo ratings (or only betting odds). Figure 10B shows the number of bogey player pairs identified by Elo ratings based on whether the player pairs were also identified by betting odds (or only by Elo ratings).

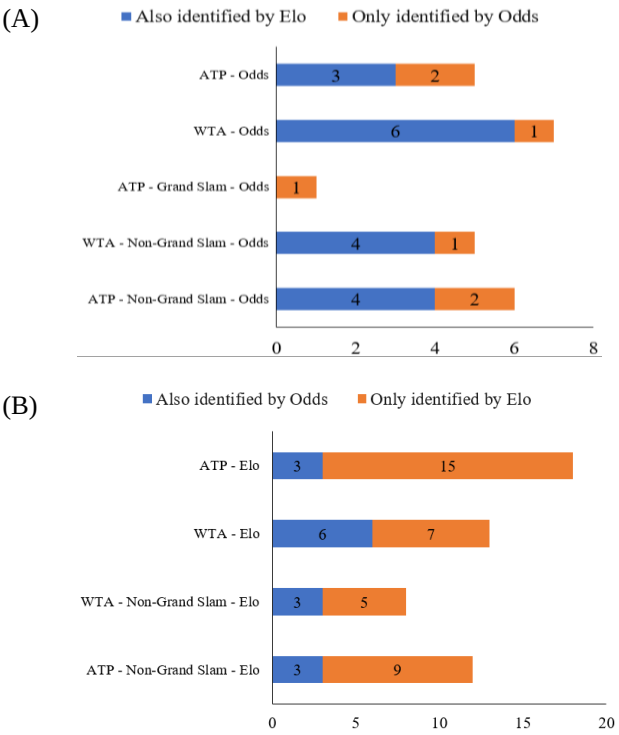


Figure 10: The number of bogey player pairs identified by (A) betting odds, split into whether they were also identified by Elo ratings or only by betting odds, and (B) Elo ratings, split into whether they were also identified by betting odds or only Elo ratings.

For example, for the whole ATP dataset, of the significant bogey player pairs identified by Elo ratings (Supplementary Table 1), three of these (Ramirez-Hidalgo R. vs Verdasco F.; Kohlschreiber P. vs Mayer L.; and Benneteau J. vs Gulbis E.) were also identified by betting odds (Supplementary Table 2). Fifteen bogey player pairs that were identified by Elo ratings were only identified by Elo but were not identified by betting odds.

Figure 10A suggests that the bogey player pairs identified by the proposed method with betting odds were largely also identified with Elo ratings. However, Figure 10B suggests that the bogey player pairs identified by the proposed method with Elo ratings were largely not also identified with betting odds (the WTA all-match dataset is one notable exception).

3.4. Visualising the expected win distribution violation quantification for each dataset

Figure 11 shows average expected and actual win probability, as well as the differences, for player pairs containing and not containing bogey players. The underlying data for this plot is shown in Supplementary Table 12 in the Supplemental material. For each player pair type, whether the player pair is a bogey player pair or not, the average expected wins and average actual wins were calculated by scaling the actual and expected wins based on the total number of matches between the players in the player pair. Taking the difference between these two values provides a means of quantifying the degree to which the expected win distribution is violated. The red values in Figure 11 denote the average difference in expected and actual win probabilities for player pairs without a bogey player, while the blue values denote the average difference in expected and actual win probabilities for player pairs that don't involve a bogey player. Three of the datasets/methods on the far-right have no bogey player pairs (see Table 2) and therefore only two points show for these. As we would expect based on our proposed method's design, for player pairs that do contain a bogey player, there is a large violation of the expected wins distribution in terms of the average difference in expected and actual wins (the blue values), and are many times larger than the violation of the expected distribution for player pairs not containing a bogey player (red values). It is also notable that while the values representing the violation of the expected wins distribution for bogey player pairs were all relatively similar, ranging from 0.665 to 0.708, whereas the values representing the violation of the expected wins distribution values for bogey player pairs—while smaller in magnitude—ranged from 0.0389 to -0.0382.

4. Discussion

This study proposed a bogey player identification method that involves computing an expected wins distribution using the summed implied win probabilities based on Elo ratings and betting odds, constructing a contingency table containing actual and expected wins between each player in a player pair, and applying Fisher's Exact Test to this contingency table. If a significant result from Fisher's Exact Test was obtained, and the actual wins and expected wins contradict, the bogey effect was deemed to exist between the two players in the player pair.

The obtained results suggest that although the bogey player effect exists in professional tennis, it is very rare. The proposed

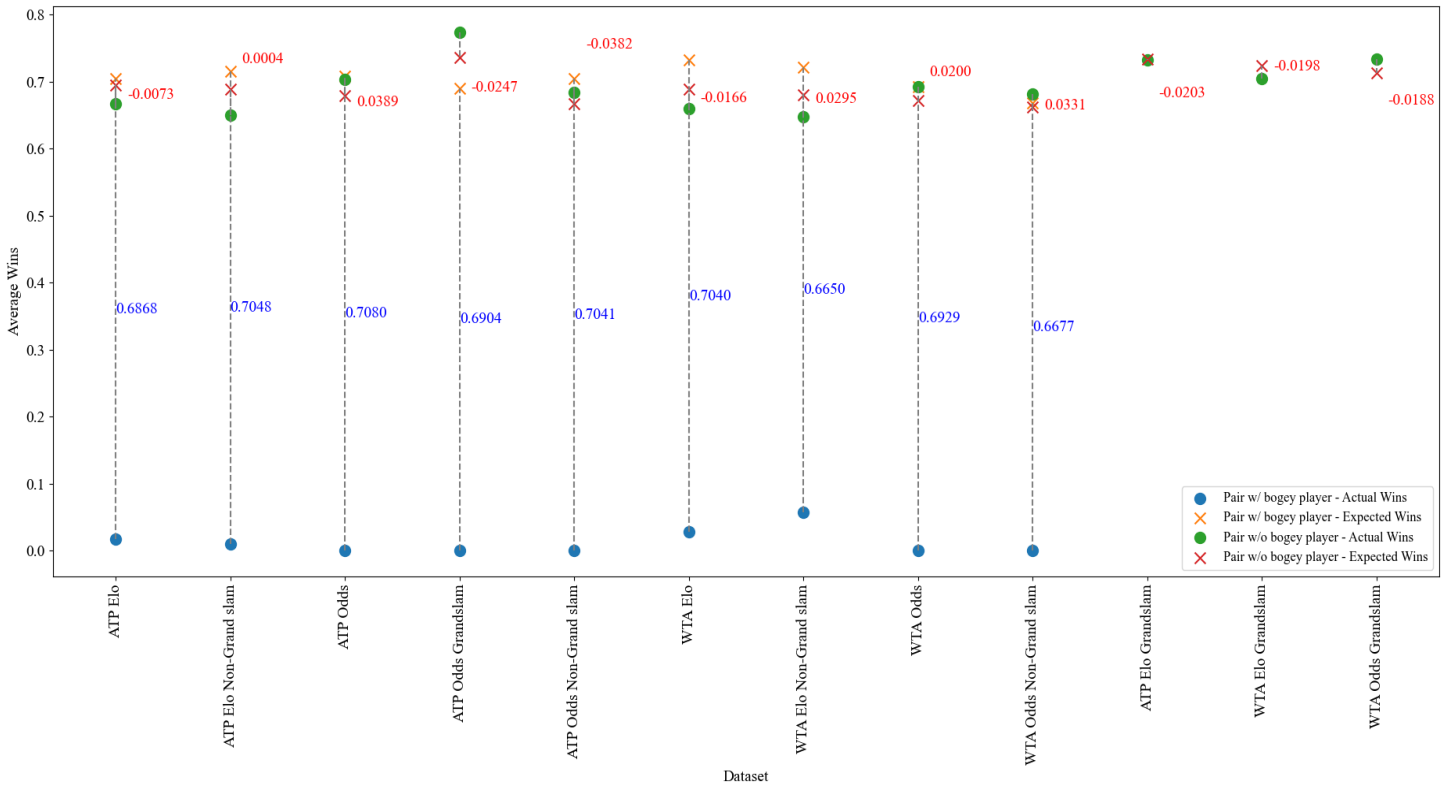


Figure 11: For each dataset and method (Elo/odds), this plot shows the average expected wins and average actual wins scaled based on the number of matches between players in a player pair, as well as their differences for player pairs that contain and do not contain bogey players (at the 90% level of confidence) (see Supplementary Table 12 in the Supplemental material for the underlying data in this plot).

method with betting odds used to compute expected wins obtained fewer bogey player pairs, as a percentage of total player pairs in the dataset, for all datasets except for the ATP Grand Slam dataset. Only one significant (at the 90% level of confidence) bogey player pair was identified across the four Grand Slam datasets: ATP and WTA Grand Slam datasets using Elo ratings and betting odds. The results conformed with our prior expectation that Grand Slams are closely analysed by bookmakers and so odds are set accurately, which results in fewer bogey player pairs in general. Furthermore, when the proposed method is used with Elo ratings to compute expected wins, the small differences in Elo ratings in matches among generally top-ranked players at Grand Slams mean that fewer bogey player pairs are identified compared to non-Grand Slams. Although betting odds identified relatively fewer bogey player pairs in ATP than WTA tennis in the ATP and WTA datasets as a whole, this did not hold for the Grand Slam or non-Grand Slam datasets. Since there were relatively fewer bogey player pairs identified in the WTA using Elo ratings compared to the ATP, Elo ratings could be said to be of better predictive value for the WTA than the ATP, a result that is consistent with Vaughan Williams et al. (2021). Surprisingly, when visualising obtained FET-significant player pairs and bogey player pairs for the various datasets, certain patterns emerged in terms of the clusters of bogey player pairs' absolute differences in actual and expected wins. However, these patterns/clusters did not appear useful for identifying which of the pairs are actually bogey player pairs. When analysing the overlap in the bogey player pairs that were identified by Elo ratings and betting odds, while betting odds generally identified fewer bogey player pairs than Elo ratings, the majority of the bogey player pairs identified by betting odds were also identified by Elo ratings. On the other hand, the majority of bogey player pairs identified by Elo ratings were only identified by Elo ratings but not by betting odds. This suggests perhaps that betting odds may be generally a more reliable means of computing expected wins and thus identifying bogey player pairs. When visualising the expected win distribution violation quantification for the various datasets by considering the average difference in expected and actual wins for player pairs containing and not containing bogey players, we validated that, for bogey player pairs, there was a large violation of the expected wins distribution in terms of the average difference in expected and actual wins, and these differences – which were relatively similar across the different datasets/methods (Elo and odds) – were many times larger than the violation of the expected distribution for non-bogey player pairs.

Analysing a particular player's performance, whether they are prone to being a bogey player or being the bogey player of another, can be useful for player-level performance analysis and match preparation. For instance, Stosur is a WTA player who appeared both as a bogey player and as a non-bogey player in bogey player pairs, which is interesting from a practical performance analysis perspective, for example, for her opponents and coaching staff to analyse further.

The proposed method is flexible in that it can be applied not only to tennis but to other sports, directly to sports with two outcomes and with some modifications to sports with more than

two outcomes. In future work, the method could therefore be applied to other sports with two outcomes (e.g., basketball), and it could be extended to sports with three outcomes, for example, soccer, by using extensions of Fisher's Exact Test that can handle contingency tables with more than two columns/rows, for example, the Freeman-Halton extension of Fisher's Exact Test (Freeman & Halton, 1951). For instance, to include a draw outcome, in addition to summing the win probability for each opposition as in the current study, the probability of a draw multiplied by 0.5 for both players would be summed, and 0.5 could be used to represent an actual draw result. Rating systems other than Elo ratings could also be trialled. Finally, other subsets of the original dataset other than Grand Slam/non-Grand Slam, for example, based on court surface, time period, player hand, and player rank group could be considered. For example, based on Figure 3 above, splitting the WTA dataset into 2005 to 2012 and 2013 to 2020 would be an obvious partition of the original dataset. Also, subsets of the original dataset based on Elo rating differences would also be interesting to investigate. For instance, since bookmakers have only a limited (or no match) history to go on in the case of matches among two low-ranked players, odds may be more difficult to set, and therefore their value for prediction may be lower and thus more bogey player pairs would be identified. Matches among low-ranked players would, however, have a small ranking difference in terms of Elo ratings and, therefore, based on the findings of Yue et al. (2022) would be more predictable and we could hypothesise that this would result in fewer bogey player pairs compared to when odds are used.

Conflict of Interest

The authors declare no conflict of interests.

Acknowledgment

This work was supported by JSPS under Grant [number 20H04075] and JST Presto under Grant [number JPMJPR20CA].

Data availability

The dataset that supports the findings of this study was obtained, as described in the "Data" subsection of "Materials and Methods", from the openly available data in the Appendix of the online version of the paper by Angelini, Candila, and De Angelis (2022), at doi:10.1016/j.ejor.2021.04.011 under "Supplementary Data S1". This is open data under the CC BY license http://creativecommons.org/licenses/by/4.0/

Code

The code is available at the following GitHub repository: https://github.com/rorybunker/bogey-phenomenon-sport/

References

Angelini, G., Candila, V., & De Angelis, L. (2022). Weighted Elo rating for tennis match predictions. *European Journal of Operational Research, 297*(1), 120–132.

Ayton, P., & Fischer, I. (2004). The hot hand fallacy and the gambler's fallacy: Two faces of subjective randomness? *Memory & Cognition, 32*(6), 1369–1378.

Awerbuch, W. (2009). *Anwendung von Data Mining zu statistischen Auswertungen und Vorhersagen im Sport* [Master's thesis, TU Darmstadt]. TU Darmstadt Knowledge Engineering website. https://ke-tud.github.io/lehre/arbeiten/diplom/2009/Awerbuch_Wladimir.pdf

Baboota, R., & Kaur, H. (2019). Predictive analysis and modelling football results using machine learning approach for English Premier League. *International Journal of Forecasting, 35*(2), 741–755.

Barrutiabengoa, J. M., Corredor, P., & Muga, L. (2022). Does the betting industry price gender? Evidence from professional tennis. *Journal of Sports Economics, 23*(7), 881–906.

Bar-Eli, M., Avugos, S., & Raab, M. (2006). Twenty years of "hot hand" research: Review and critique. *Psychology of Sport and Exercise, 7*(6), 525–553.

Bunker, R. (2022). The Bogey Phenomenon in Sport. In J. J. Reade (Eds.), *IX Mathsport International 2022 Proceedings* (pp.15–21). https://www.mathsportinternational.com/MathSport2022Proceedings.pdf

Candila, V. (2023). welo: An R package for weighted and standard Elo rates. *Statistica Applicata-Italian Journal of Applied Statistics*, (1).

Carlson, K. A., & Shu, S. B. (2007). The rule of three: How the third event signals the emergence of a streak. *Organizational Behavior and Human Decision Processes, 104*(1), 113–121.

Chiweshe, M. K. (2021). Frenemies: Understanding the interconnectedness of rival fan identities in Harare, Zimbabwe. In K. Bandyopadhyay (Eds.), *Face to Face: Enduring Rivalries in World Soccer* (pp. 191–203). Routledge.

Cattelan, M., Varin, C., & Firth, D. (2013). Dynamic Bradley–Terry modelling of sports tournaments. *Journal of the Royal Statistical Society Series C: Applied Statistics, 62*(1), 135–150.

Coulom, R. (2007). Computing "elo ratings" of move patterns in the game of go. *ICGA Journal, 30*(4), 198–208.

Estes, W. K. (1964). Probability learning. In A. Melton (Ed.), *Categories of human learning* (pp. 89–128). Academic Press/Elsevier.

Freeman, G. H., & Halton, J. H. (1951). Note on an exact treatment of contingency, goodness of fit and other problems of significance. *Biometrika, 38*(1/2), 141–149.

Gilovich, T., Vallone, R., & Tversky, A. (1985). The hot hand in basketball: On the misperception of random sequences. *Cognitive Psychology, 17*(3), 295–314.

Hales, S. D. (1999). An epistemologist looks at the hot hand in sports. *Journal of the Philosophy of Sport, 26*(1), 79–87.

Hankin, R. K. S., & Bunker, R. (2016). Bogey teams in sport. Technical report, Auckland University of Technology.

Kovalchik, S. A. (2016). Searching for the GOAT of tennis win prediction. *Journal of Quantitative Analysis in Sports, 12*(3), 127–138.

Leitner, C., Zeileis, A., & Hornik, K. (2009). Is Federer stronger in a tournament without Nadal? An evaluation of odds and seedings for Wimbledon 2009. *Austrian Journal of Statistics, 38*(4), 277–286.

Niall, J. (2013). It's Murray v Djokovic. The Sydney Morning *Herald.* https://www.smh.com.au/sport/tennis/its-murray-v-djokovic-20130125-2dcuv.html

Peel, L., & Clauset, A. (2015). Predicting sports scoring dynamics with restoration and anti-persistence. In C. Aggarwal, Z. Zhou, A. Tuzhilin, H. Xiong, & X. Wu (Eds.), *2015 IEEE International Conference on Data Mining* (pp. 339–348). IEEE. https://doi.org/10.1109/ICDM36327.2015

Poulton, E. (2004). Mediated patriot games: The construction and representation of national identities in the British television production of Euro'96. *International Review for the Sociology of Sport*, *39*(4), 437–455.

Raab, M. (2012). Simple heuristics in sports. *International Review of Sport and Exercise Psychology, 5*(2), 104–120.

Steeger, G. M., Dulin, J. L., & Gonzalez, G. O. (2021). Winning and losing streaks in the National Hockey League: Are teams experiencing momentum or are games a sequence of random events? *Journal of Quantitative Analysis in Sports, 17*(3), 155–170.

Stone, D. F. (2012). Measurement error and the hot hand. *The American Statistician*, *66(*1), 61–66.

Vaughan Williams, L., Liu, C., Dixon, L., & Gerrard, H. (2021). How well do Elo-based ratings predict professional tennis matches? *Journal of Quantitative Analysis in Sports*, *17*(2), 91–105.

Wilms, S. (2014). Mental Training im Sport: Möglichkeiten der Anwendung in der Sozialen Arbeit [Doctoral dissertation, Hochschule für angewandte Wissenschaften Hamburg]. HAW Hamburg Repository. https://reposit.haw-hamburg.de/handle/20.500.12738/6563

Wood, L. (2017). Jo-Wilfried Tsonga hopes to avoid Australian Open bogey Kei Nishikori. *Herald Sun.* https://www.heraldsun.com.au/sport/tennis/jowilfried-tsonga-hopes-to-avoid-australian-open-bogey-kei-nishikori/news-story/2dfaab3f641494447405ae65fc7e7592

Yue, J. C., Chou, E. P., Hsieh, M. H., & Hsiao, L. C. (2022). A study of forecasting tennis matches via the Glicko model. *PLoS ONE*, *17*(4), 1–12. https://doi.org/10.1371/journal.pone.0266838

Supplementary materials

Supplementary Table 1: Statistically significant (with at least 90% confidence) and bogey player pairs from the ATP tennis dataset, with aggregated Elo ratings-implied probabilities used to compute expected wins.

Player1 (p1) vs player2 (p2)	*p*-value FET	Expected wins (p1/p2)	Actual wins (p1/p2)	Bogey effect (Y/N)
Garcia-Lopez G. vs Youzhny M.	0.0291	1.9/5.1	6/1	Y
Baghdatis M. vs Garcia-Lopez G.	0.0476	3.28/1.72	0/5	Y
Almagro N. vs Istomin D.	0.0476	3.56/1.44	0/5	Y
Llodra M. vs Simon G.	0.0476	1.63/3.37	5/0	Y
Gasquet R. vs Pouille L.	0.0801	4.04/1.96	1/5	Y
Anderson K. vs Berdych T.	0.0932	3.07/8.93	0/12	N
Ancic M. vs Blake J.	0.1	0.91/2.09	3/0	Y
Acasuso J. vs Rochus C.	0.1	2.25/0.75	0/3	Y
Ramirez-Hidalgo R. vs Verdasco F.	0.1	0.85/2.15	3/0	Y
Nalbandian D. vs Verdasco F.	0.1	2.1/0.9	0/3	Y
Ginepri R. vs Korolev E.	0.1	2.16/0.84	0/3	Y
Kohlschreiber P. vs Mayer L.	0.1	2.15/0.85	0/3	Y
Berlocq C. vs Stakhovsky S.	0.1	2.01/0.99	0/3	Y
Benneteau J. vs Gulbis E.	0.1	0.85/2.15	3/0	Y
Fognini F. vs Paire B.	0.1	2.01/0.99	0/3	Y
Chardy J. vs Isner J.	0.1	0.87/2.13	3/0	Y
Querrey S. vs Ramos-Vinolas A.	0.1	2.17/0.83	0/3	Y
Isner J. vs Opelka R.	0.1	2.26/0.74	0/3	Y
Lorenzi P. vs Paire B.	0.1	0.93/2.07	3/0	Y

Notes: As mentioned in the manuscript when describing the proposed method, the FET needs to have a statistically significant *p*-value but also the expected wins and actual wins need to contradict to be suggestive of the bogey effect between a particular player pair. In Supplementary Table 1 above, there is one case where the expected wins and actual wins did not contradict. In particular, the FET *p*-value for Anderson vs Berdych was significant with 1- $\alpha = 1 - 0.1 = 90\%$ confidence (p-value $= 0.0932 \leq \alpha$), however, the expected wins and actual wins do not contradict. Out of the $N = 12$ historical matches between Anderson and Berdych in the dataset, Anderson was expected, based on the sums of the respective players' Elo ratings-implied win probabilities, to win 3.07 of the matches, while Berdych was expected to win 8.93 of the matches. Berdych did better than expected, achieving 12 wins while Anderson achieved zero. Thus, although this was a statistically significant result, since the actual wins and expected wins did not contradict, this is not a case where the bogey effect was deemed to exist.

Supplementary Table 2: Statistically significant (with at least 90% confidence) and bogey player pairs from the ATP tennis dataset, with aggregated betting odds-implied probabilities used to compute expected wins.

Player1 (p1) vs player2 (p2)	*p*-value FET	Expected wins (p1/p2)	Actual wins (p1/p2)	Bogey effect (Y/N)
Djokovic N. vs Monfils G.	0.0978	10.97/3.03	14/0	N
Ramirez-Hidalgo R. vs Verdasco F.	0.1	0.82/2.18	3/0	Y
Kohlschreiber P. vs Mayer L.	0.1	2.11/0.89	0/3	Y
Benneteau J. vs Gulbis E.	0.1	0.95/2.05	3/0	Y
Goffin D. vs Pouille L.	0.1	2.16/0.84	0/3	Y

Note: All player pairs in Table 4 are deemed bogey player pairs apart from Djokovic N. vs Monfils G. since the expected and actual wins do not contradict.

Supplementary Table 3: Statistically significant (with at least 90% confidence) bogey player pairs from the WTA tennis dataset, with aggregated Elo ratings-implied probabilities used to compute expected wins.

Player1 (p1) vs player2 (p2)	*p*-value FET	Expected wins (p1/p2)	Actual wins (p1/p2)	Bogey effect (Y/N)
Sharapova M. vs Williams S.	0.0184	5.56/11.44	0/17	N
Stosur S. vs Zvonareva V.	0.0256	3.25/4.75	8/0	Y
Ivanovic A. vs Kuznetsova S.	0.0325	5.57/4.43	10/0	N
Stephens S. vs Voegele S.	0.0476	4.16/0.84	1/4	Y
Petkovic A. vs Radwanska A.	0.0769	3.12/4.88	0/8	N
Goerges J. vs Stosur S.	0.0801	1.84/4.16	5/1	Y
Chakvetadze A. vs Jankovic J.	0.1	0.87/2.13	3/0	Y
Safarova L. vs Shvedova Y.	0.1	2.2/0.8	0/3	Y
Hibino N. vs Stosur S.	0.1	0.59/2.41	3/0	Y
Date Krumm K. vs Kirilenko M.	0.1	0.93/2.07	3/0	N
Cetkovska P. vs Radwanska A.	0.1	0.4/2.6	3/0	Y
Cetkovska P. vs Johansson M.	0.1	2.46/0.54	0/3	Y
Petkovic A. vs Rybarikova M.	0.1	2.09/0.91	0/3	Y
Pavlyuchenkova A. vs Wickmayer Y.	0.1	2.09/0.91	0/3	Y
Gavrilova D. vs Tsurenko L.	0.1	2.02/0.98	0/3	N
Konta J. vs Strycova B.	0.1	2.12/0.88	0/3	Y

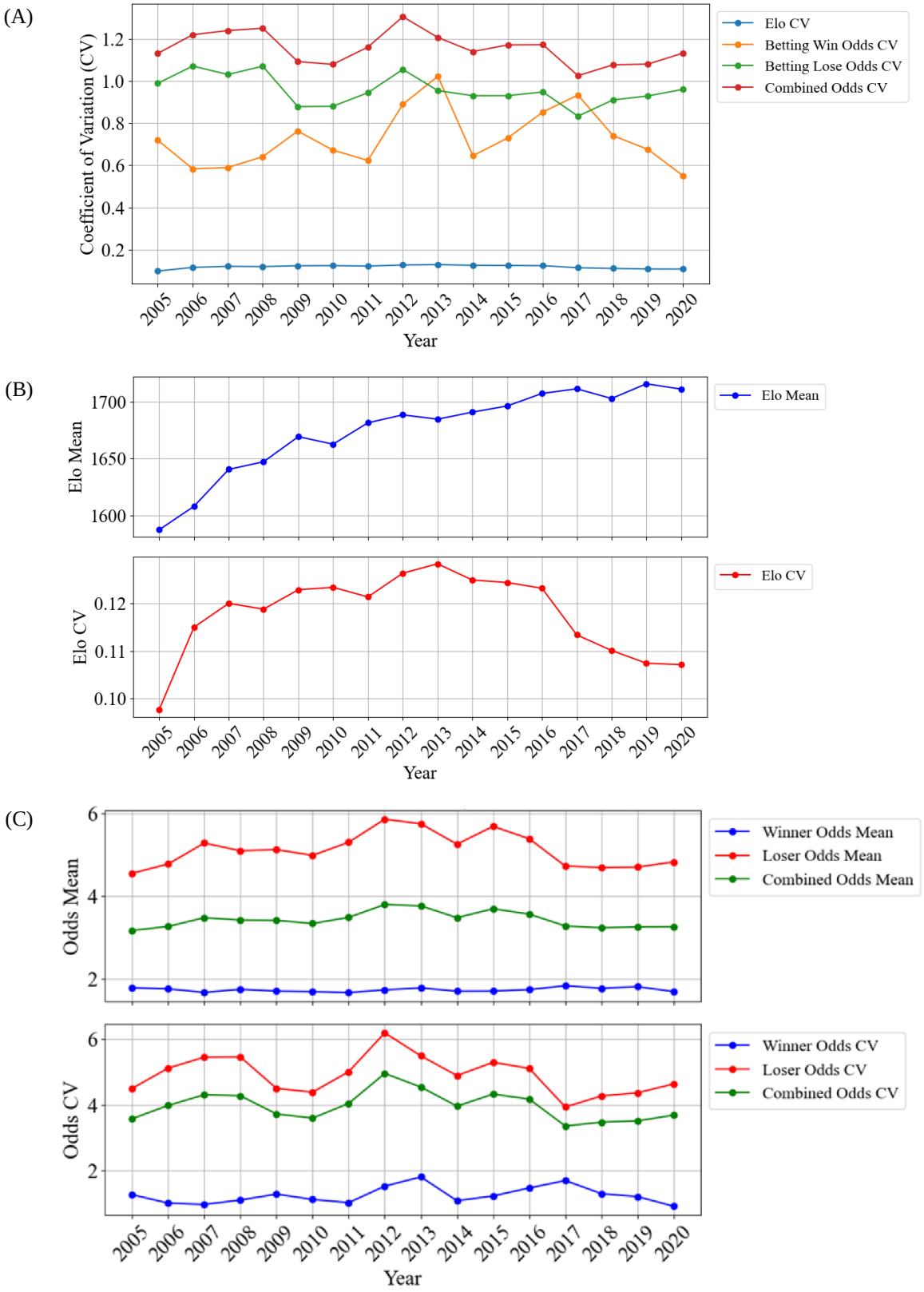
Supplementary Table 4: Statistically significant (with at least 90% confidence) bogey player pairs from the WTA tennis dataset, with aggregated Betting odds-implied probabilities used to compute expected wins.

Player1 (p1) vs player2 (p2)	*p*-value FET	Expected wins (p1/p2)	Actual wins (p1/p2)	Bogey effect (Y/N)
Sharapova M. vs Williams S.	0.0184	5.4/11.6	0/17	N
Stosur S. vs Zvonareva V.	0.0256	3.4/4.6	8/0	Y
Ivanovic A. vs Kuznetsova S.	0.0325	5.65/4.35	10/0	N
Petkovic A. vs Radwanska A.	0.0769	3.07/4.93	0/8	N
Chakvetadze A. vs Jankovic J.	0.1	0.92/2.08	3/0	Y
Date Krumm K. vs Kirilenko M.	0.1	0.85/2.15	3/0	N
Cetkovska P. vs Radwanska A.	0.1	0.54/2.46	3/0	Y
Cetkovska P. vs Johansson M.	0.1	2.12/0.88	0/3	Y
Riske A. vs Wang Q.	0.1	0.99/2.01	3/0	N
Konta J. vs Strycova B.	0.1	2.02/0.98	0/3	Y

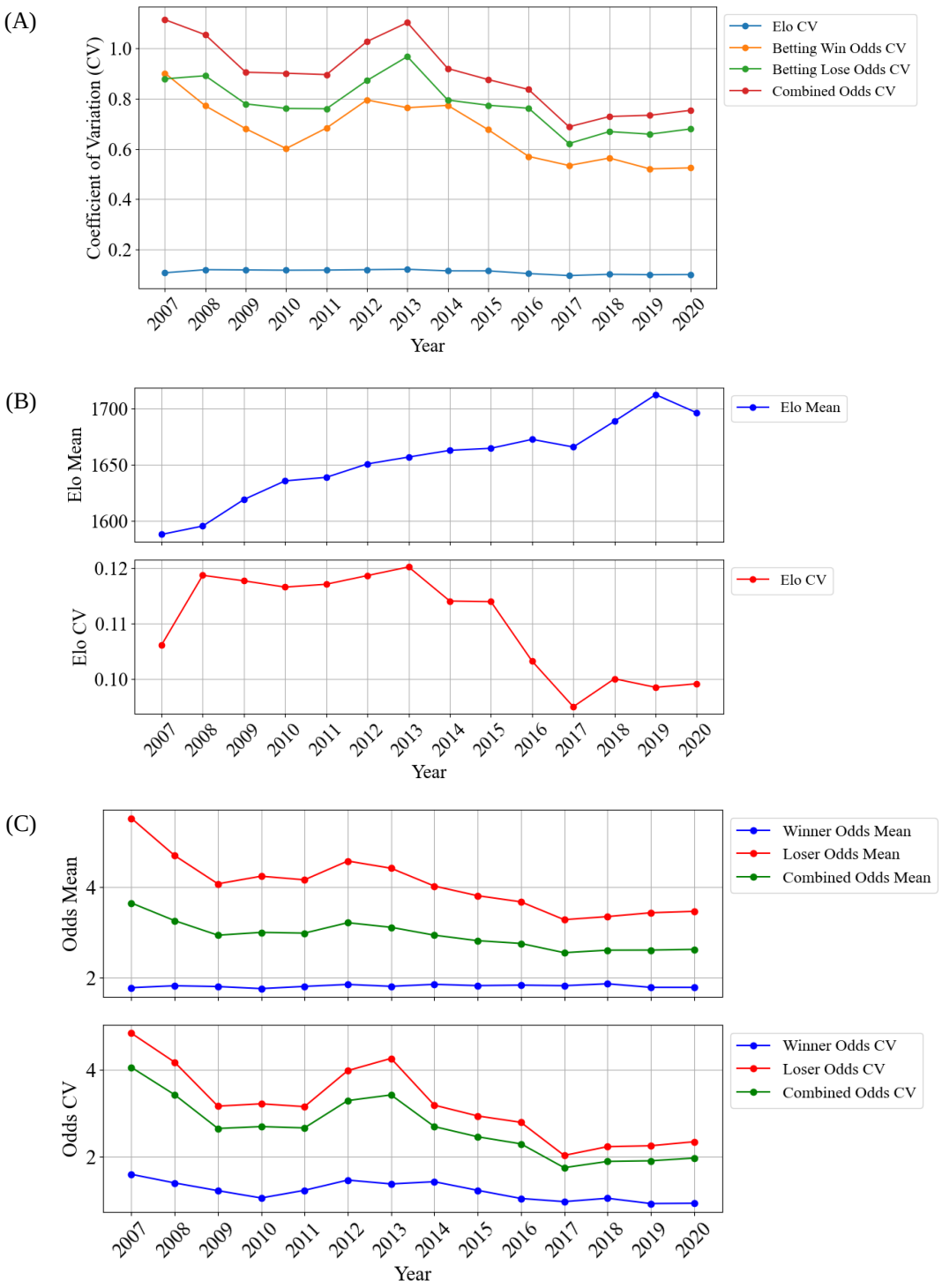
Supplementary Table 5: Statistically significant (with at least 90% confidence) bogey player pairs from the ATP tennis Grand Slams dataset, with aggregated Betting odds-implied probabilities used to compute expected wins.

Player1 (p1) vs player2 (p2)	*p*-value FET	Expected wins (p1/p2)	Actual wins (p1/p2)	Bogey effect (Y/N)
Isner J. vs Kohlschreiber P.	0.1	2.07/0.93	0/3	Y

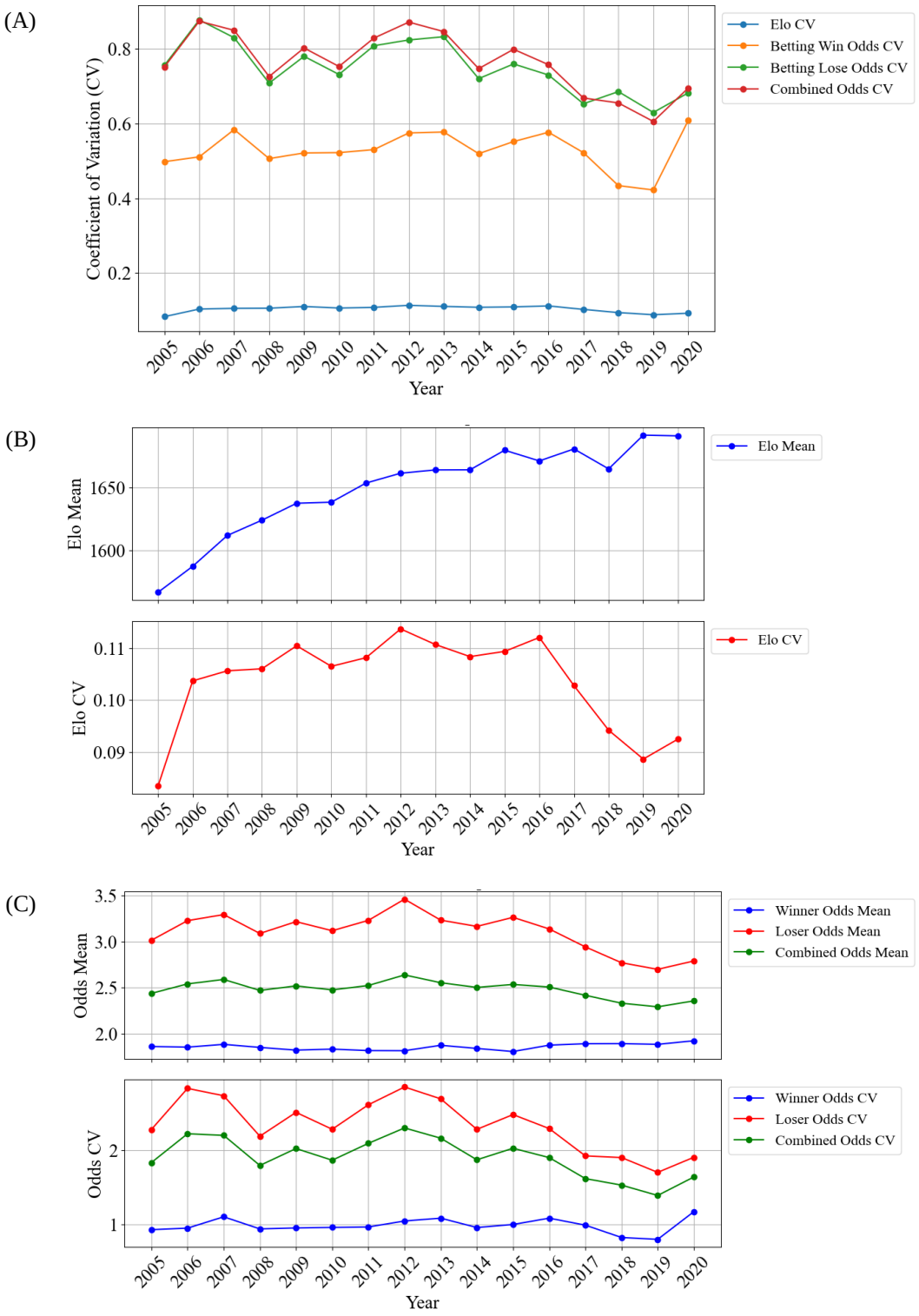
Notes: The ATP and WTA Grand Slams datasets with Elo ratings used to compute expected wins, and the WTA Grand Slams dataset with betting odds-implied probabilities used to compute expected wins, all yielded no bogey effect player pairs.



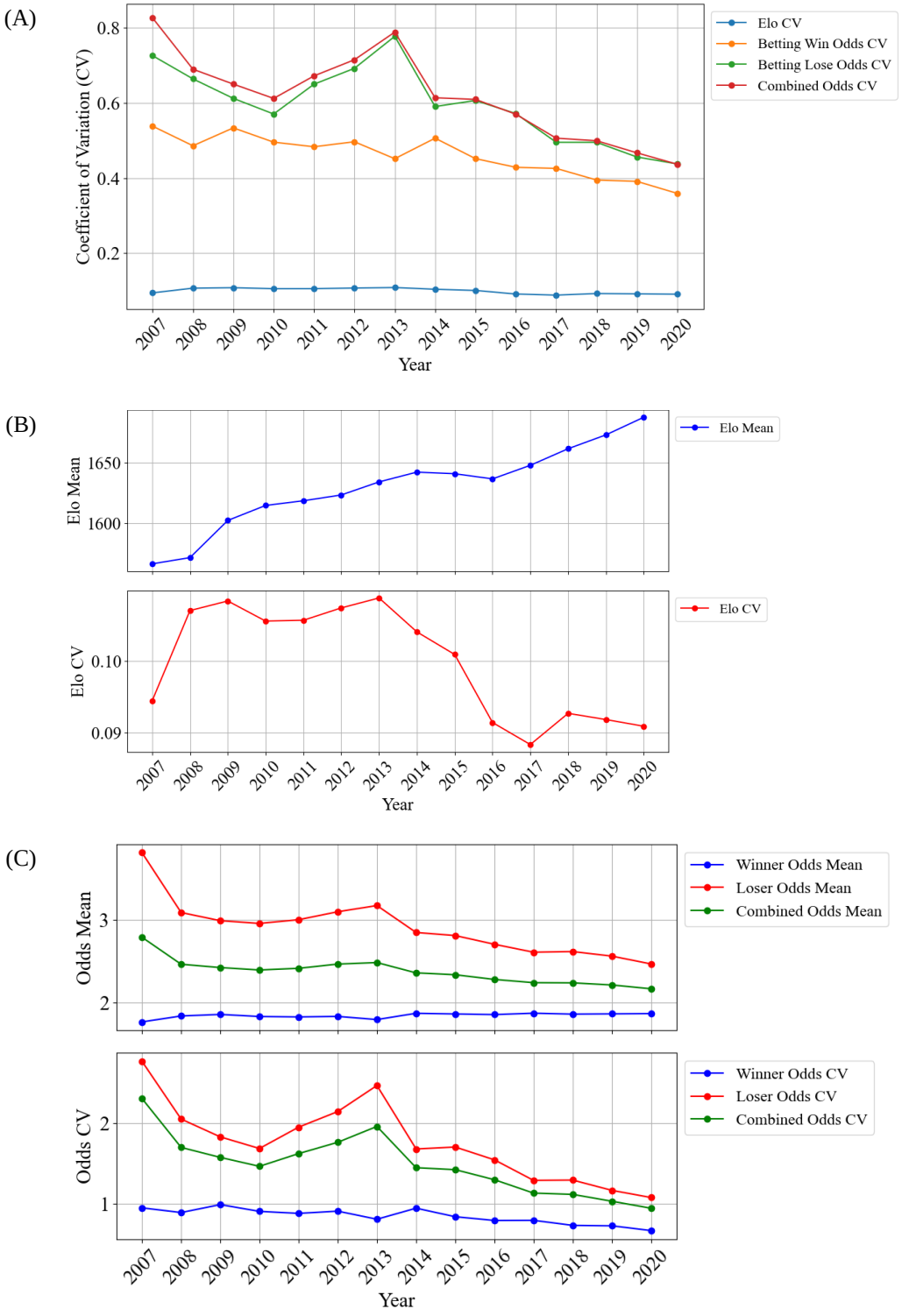
Supplementary Figure 1: Coefficient of Variation (CV) of Betting Odds and Elo Ratings for each year (A), as well as the more granular view of the mean and CV for each year of Elo rating and Betting Odds (B and C, respectively) for the ATP Grand Slam dataset.



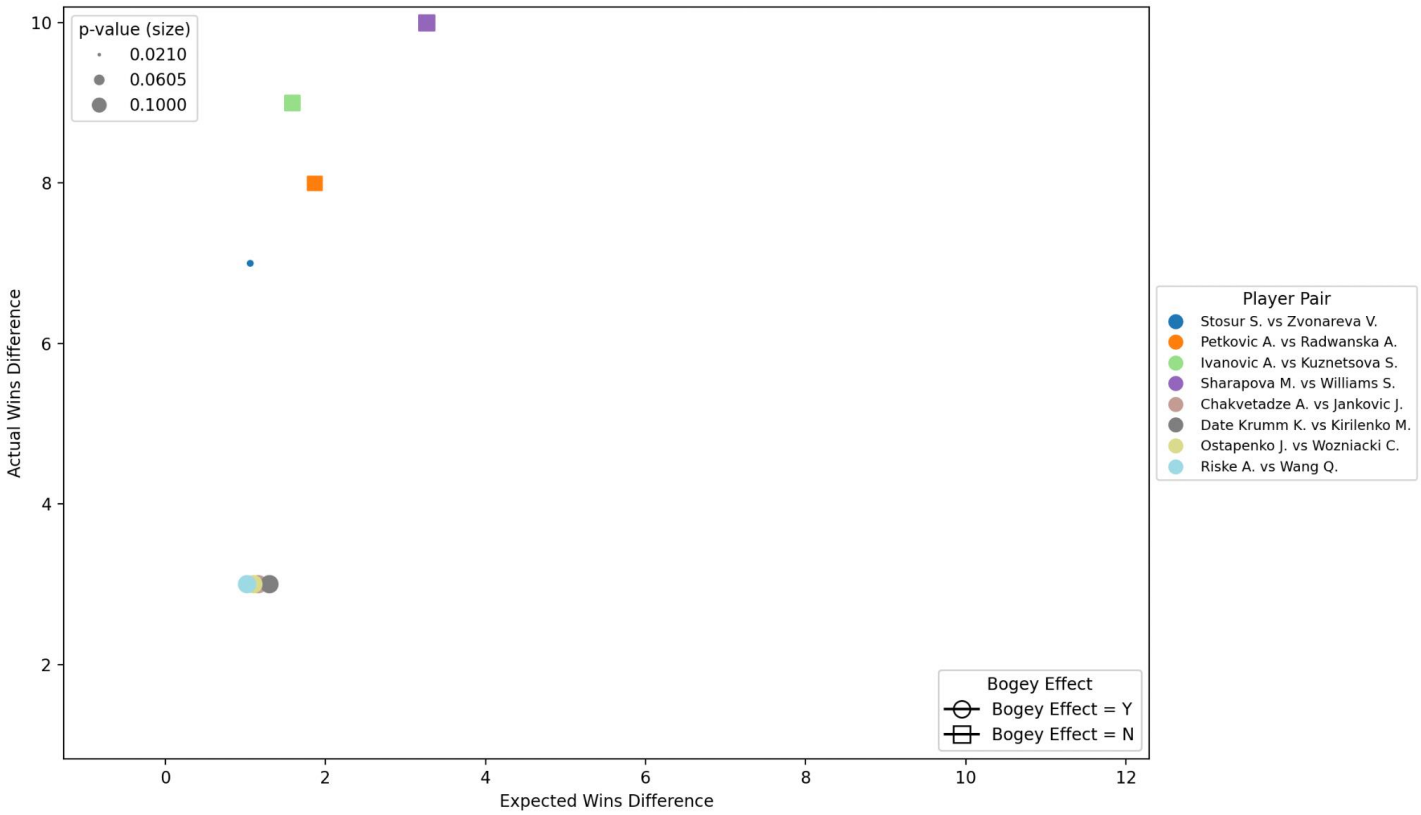
Supplementary Figure 2: Coefficient of Variation (CV) of Betting Odds and Elo Ratings for each year (A) as well as the more granular view of the mean and CV for each year of Elo rating and Betting Odds (B and C, respectively) for the WTA Grand Slam dataset.



Supplementary Figure 3: Coefficient of Variation (CV) of Betting Odds and Elo Ratings for each year (A), as well as the more granular view of the mean and CV for each year of Elo rating and Betting Odds (B and C, respectively) for the ATP non-Grand Slam dataset.



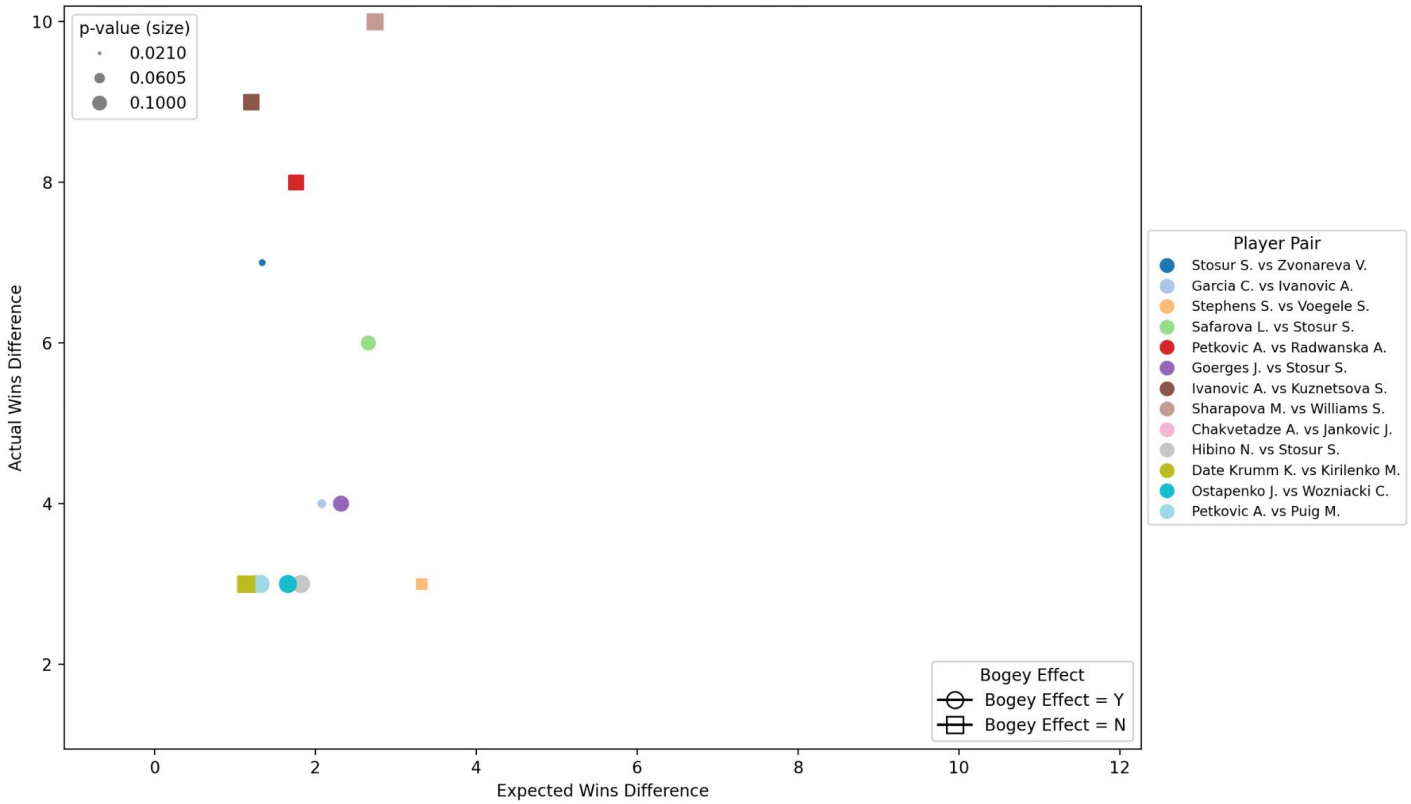
Supplementary Figure 4: Coefficient of Variation (CV) of Betting Odds and Elo Ratings for each year (A), as well as the more granular view of the mean and CV for each year of Elo rating and Betting Odds (B and C, respectively) for the WTA non-Grand Slam dataset.



Supplementary Figure 5: Visualizing, on axes representing the absolute difference in actual wins and expected wins for the player pair, statistically significant (with at least 90% confidence) bogey player pairs from the WTA tennis non-Grand Slams dataset, with aggregated betting odds-implied probabilities used to compute expected wins. The data used to generate this plot is presented below in Supplementary Table 6.

Supplementary Table 6: Statistically significant (with at least 90% confidence) bogey player pairs from the WTA tennis non-Grand Slams dataset, with aggregated betting odds-implied probabilities used to compute expected wins.

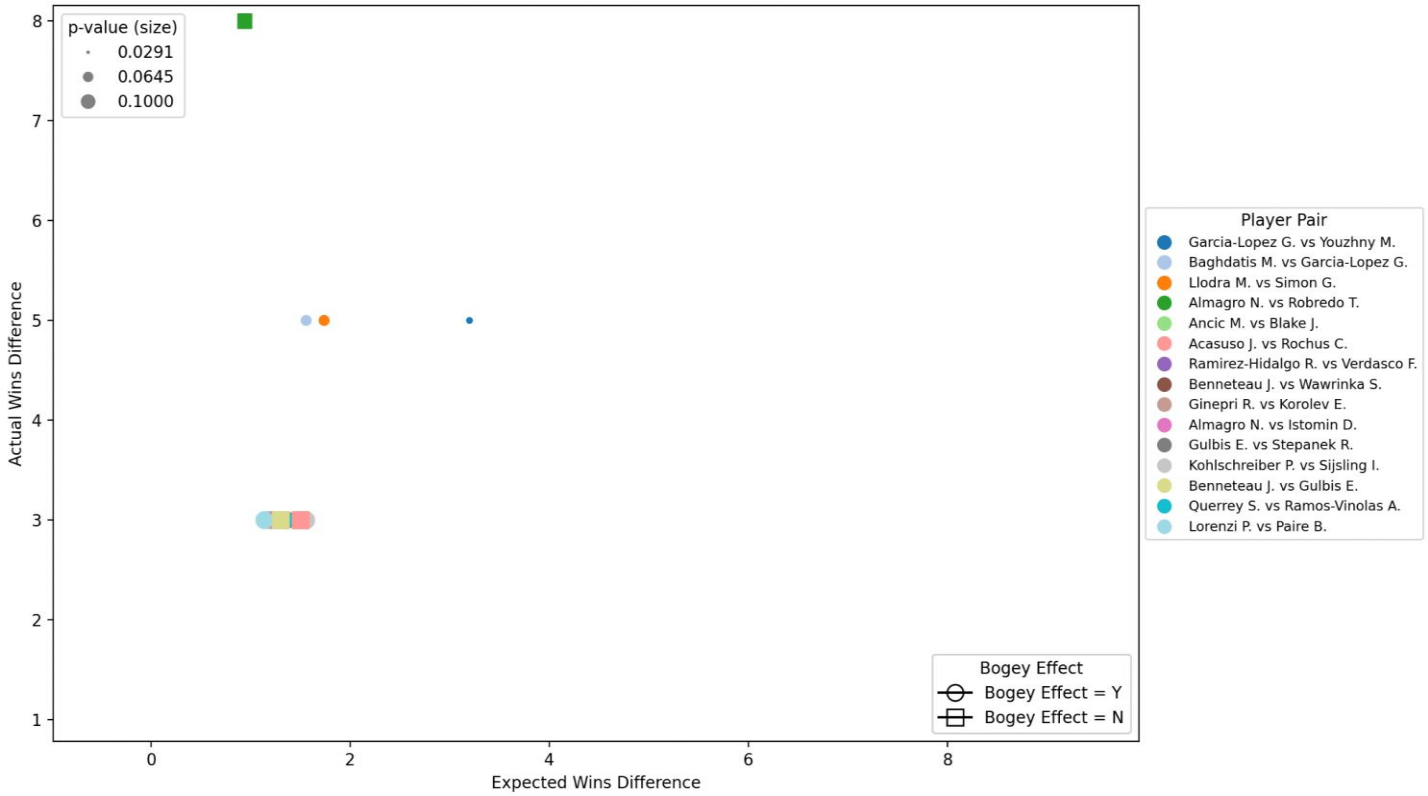
Player1 (p1) vs player2 (p2)	*p*-value FET	Expected wins (p1/p2)	Actual wins (p1/p2)	Bogey effect (Y/N)
Stosur S. vs Zvonareva V.	0.021	2.97/4.03	7/0	Y
Petkovic A. vs Radwanska A.	0.0769	3.07/4.93	0/8	
Ivanovic A. vs Kuznetsova S.	0.0824	5.29/3.71	9/0	
Sharapova M. vs Williams S.	0.0867	3.37/6.63	0/10	
Chakvetadze A. vs Jankovic J.	0.1	0.92/2.08	3/0	Y
Date Krumm K. vs Kirilenko M.	0.1	0.85/2.15	3/0	Y
Ostapenko J. vs Wozniacki C.	0.1	0.95/2.05	3/0	Y
Riske A. vs Wang Q.	0.1	0.99/2.01	3/0	Y



Supplementary Figure 6: Visualizing, on axes representing the absolute difference in actual wins and expected wins for the player pair, statistically significant (with at least 90% confidence) bogey player pairs from the WTA non-Grand Slams dataset, with aggregated Elo ratings-implied probabilities used to compute expected wins. The data used to generate this plot is presented below in Supplementary Table 7.

Supplementary Table 7: Statistically significant (with at least 90% confidence) bogey player pairs from the WTA non-Grand Slams dataset, with aggregated Elo ratings-implied probabilities used to compute expected wins.

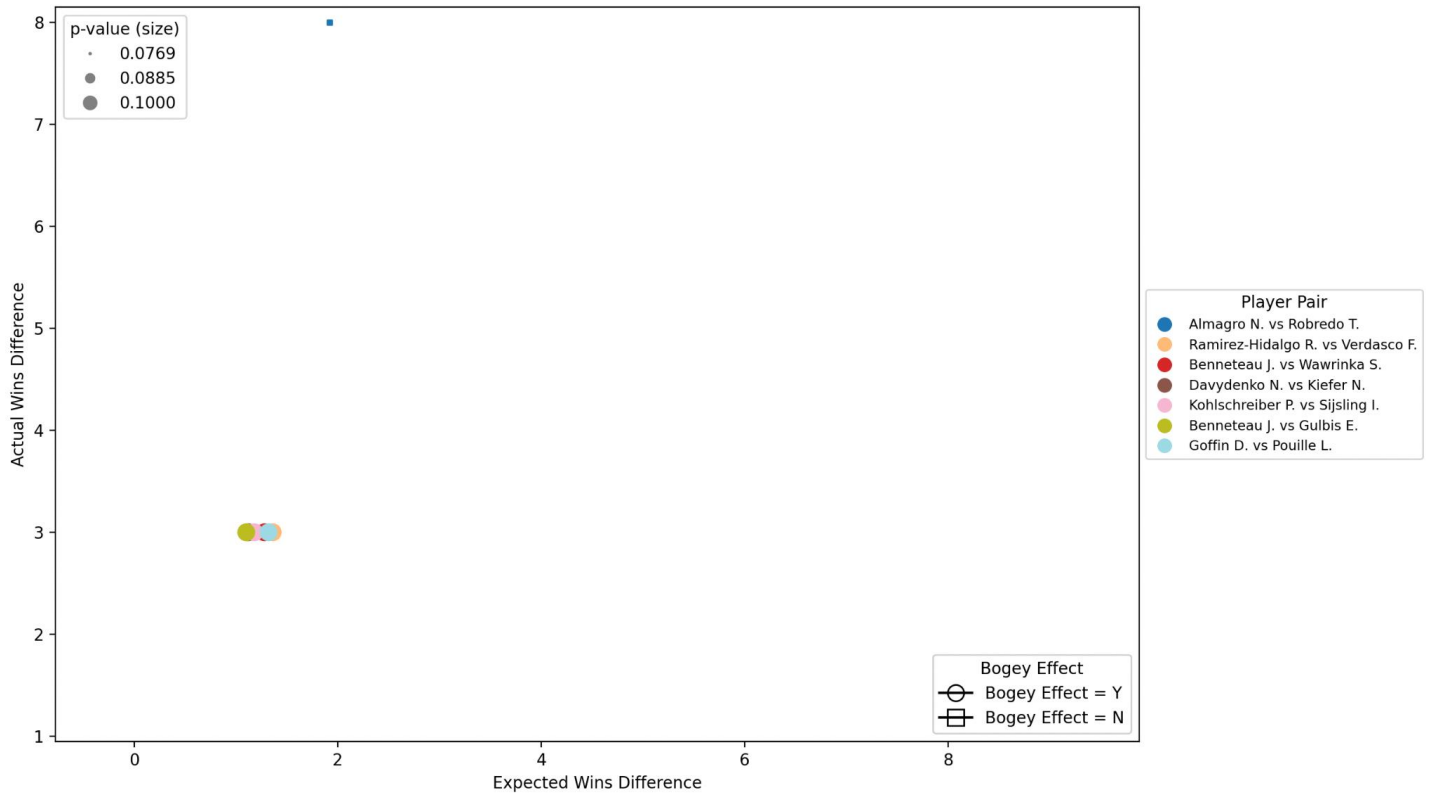
Player1 (p1) vs player2 (p2)	*p*-value FET	Expected wins (p1/p2)	Actual wins (p1/p2)	Bogey effect (Y/N)
Stosur S. vs Zvonareva V.	0.021	2.83/4.17	7/0	Y
Garcia C. vs Ivanovic A.	0.0286	0.96/3.04	4/0	Y
Stephens S. vs Voegele S.	0.0476	4.16/0.84	1/4	N
Safarova L. vs Stosur S.	0.0698	3.67/6.33	8/2	Y
Petkovic A. vs Radwanska A.	0.0769	3.12/4.88	0/8	N
Goerges J. vs Stosur S.	0.0801	1.84/4.16	5/1	Y
Ivanovic A. vs Kuznetsova S.	0.0824	5.1/3.9	9/0	N
Sharapova M. vs Williams S.	0.0867	3.63/6.37	0/10	N
Chakvetadze A. vs Jankovic J.	0.1	0.87/2.13	3/0	Y
Hibino N. vs Stosur S.	0.1	0.59/2.41	3/0	Y
Date Krumm K. vs Kirilenko M.	0.1	0.93/2.07	3/0	N
Ostapenko J. vs Wozniacki C.	0.1	0.67/2.33	3/0	Y
Petkovic A. vs Puig M.	0.1	2.16/0.84	0/3	Y



Supplementary Figure 7: Visualizing, on axes representing the absolute difference in actual wins and expected wins for the player pair, statistically significant (with at least 90% confidence) bogey player pairs from the ATP non-Grand Slams dataset, with aggregated Elo ratings-implied probabilities used to compute expected wins. The data used to generate this plot is presented below in Supplementary Table 8.

Supplementary Table 8: Statistically significant (with at least 90% confidence) bogey player pairs from the ATP non-Grand Slams dataset, with aggregated Elo ratings-implied probabilities used to compute expected wins.

Player1 (p1) vs player2 (p2)	*p*-value FET	Expected wins (p1/p2)	Actual wins (p1/p2)	Bogey effect (Y/N)
Garcia-Lopez G. vs Youzhny M.	0.0291	1.9/5.1	6/1	Y
Baghdatis M. vs Garcia-Lopez G.	0.0476	3.28/1.72	0/5	Y
Llodra M. vs Simon G.	0.0476	1.63/3.37	5/0	Y
Almagro N. vs Robredo T.	0.0769	4.47/3.53	8/0	N
Ancic M. vs Blake J.	0.1	0.91/2.09	3/0	Y
Acasuso J. vs Rochus C.	0.1	2.25/0.75	0/3	N
Ramirez-Hidalgo R. vs Verdasco F.	0.1	0.85/2.15	3/0	Y
Benneteau J. vs Wawrinka S.	0.1	0.75/2.25	3/0	Y
Ginepri R. vs Korolev E.	0.1	2.16/0.84	0/3	Y
Almagro N. vs Istomin D.	0.1	2.11/0.89	0/3	Y
Gulbis E. vs Stepanek R.	0.1	0.82/2.18	3/0	Y
Kohlschreiber P. vs Sijsling I.	0.1	2.28/0.72	0/3	Y
Benneteau J. vs Gulbis E.	0.1	0.85/2.15	3/0	N
Querrey S. vs Ramos-Vinolas A.	0.1	2.17/0.83	0/3	Y
Lorenzi P. vs Paire B.	0.1	0.93/2.07	3/0	Y



Supplementary Figure 8: Visualizing, on axes representing the absolute difference in actual wins and expected wins for the player pair, statistically significant (with at least 90% confidence) bogey player pairs from the ATP non-Grand Slams dataset, with aggregated betting odds-implied probabilities used to compute expected wins. This figure was generated using the data presented below in Supplementary Table 9.

Supplementary Table 9: Statistically significant (with at least 90% confidence) bogey player pairs from the ATP non-Grand Slams dataset, with aggregated betting odds-implied probabilities used to compute expected wins.

Player1 (p1) vs player2 (p2)	*p*-value FET	Expected wins (p1/p2)	Actual wins (p1/p2)	Bogey effect (Y/N)
Almagro N. vs Robredo T.	0.0769	4.96/3.04	8/0	N
Ramirez-Hidalgo R. vs Verdasco F.	0.1	0.82/2.18	3/0	Y
Benneteau J. vs Wawrinka S.	0.1	0.86/2.14	3/0	Y
Davydenko N. vs Kiefer N.	0.1	2.06/0.94	0/3	Y
Kohlschreiber P. vs Sijsling I.	0.1	2.09/0.91	0/3	Y
Benneteau J. vs Gulbis E.	0.1	0.95/2.05	3/0	Y
Goffin D. vs Pouille L.	0.1	2.16/0.84	0/3	Y

Supplementary Table 10: Expected win distribution violation quantification for the various datasets. For each type of player pair – whether the player pair contains a bogey player or not – average expected wins and average actual wins were calculated and scaled based on the total number of historical matches between the player pair. Taking the difference between these two values provides a means of quantifying the degree to which the expected win distribution is violated. This data was used to generate the plot depicted in Figure 11.

Dataset / method	Player pair type	Average expected win	Average actual wins	Average difference between Expected and Actual wins
ATP / Elo				
	Pair w/o bogey player	0.6952	0.6665	-0.0073
	Pair w/ bogey player	0.7040	0.0172	0.6868
ATP / Odds				
	Pair w/o bogey player	0.6787	0.7034	-0.0247
	Pair w/ bogey player	0.7080	0	0.7080
ATP Grand Slam / Elo				
	Pair w/o bogey player	0.7333	0.7329	0.00044
	Pair w/ bogey player	-	-	-
ATP Non-Grand Slam / Elo				
	Pair w/o bogey player	0.6886	0.6497	0.0389
	Pair w/ bogey player	0.7150	0.01020	0.7048
ATP Grand Slam / Odds				
	Pair w/o bogey player	0.7358	0.7741	-0.0382
	Pair w/ bogey player	0.6904	0	0.6904
ATP Non-Grand Slam / Odds				
	Pair w/o bogey player	0.6676	0.6841	-0.0166
	Pair w/ bogey player	0.7041	0	0.7041
WTA / Elo				
	Pair w/o bogey player	0.6887	0.6592	0.0295
	Pair w/ bogey player	0.7322	0.0282	0.7040
WTA / Odds				
	Pair w/o bogey player	0.6722	0.6925	-0.0203
	Pair w/ bogey player	0.6929	0	0.6929
WTA Grand Slam / Elo				
	Pair w/o bogey player	0.7244	0.7043	0.0200
	Pair w/ bogey player	-	-	-
WTA Non-Grand Slam / Elo				
	Pair w/o bogey player	0.6807	0.6475	0.0331
	Pair w/ bogey player	0.7217	0.0567	0.6650
WTA Grand Slam / Odds				
	Pair w/o bogey player	0.7133	0.7331	-0.0198
	Pair w/ bogey player	-	-	-
WTA Non-Grand Slam / Odds				
	Pair w/o bogey player	0.6626	0.6814	-0.0188
	Pair w/ bogey player	0.6677	0	0.6677