1 Executive Summary

The objective of this report is to better understand the contributing factors that explain the outcome of Women's Tennis Association (WTA) tennis matches in 2023. To do so, a Bayesian logistic model will be fit using a Monte Carlo Markov Chain (MCMC) simulation and the Metropolis-Hastings algorithm. Player rankings obtained from the model can be compared with the WTA tennis rankings at the time of the paper for validation and further insight.

Of the top 10 ranked WTA players as of November 8th, 2023, 8 were also among the top 10 players in terms of match winningness based on the fitted Bayesian model. Furthermore, the top 3 ranked WTA players – Iga Swiatek, Aryna Sabalenka, and Coco Gauff – had consistent rankings in the Bayesian model.

While hypothesized to be an important match outcome predictor in the equivalent men's Associate of Tennis Professionals (ATP) tour, inclusion of player interaction effects with different playing surfaces (i.e., hard court) upon inference in the WTA model was not found to improve differentiation of match outcomes.

The differential between fitted player scores in a match had a 79% area under the curve (AUC) when differentiating observed match outcomes. This indicates that simply by linearly ranking each WTA player (using a Bayesian model) and inferring that the player with the superior rank will win, one can expect to get 79% of inferences correct (at least retroactively on 2023 data).

2 Data Collection

WTA match data is publicly available at http://www.tennis-data.co.uk/alldata.php. Datasets include the players, surface, tournament, tier (i.e., semifinal), and the match outcome. 2023 data will be used as of November 8th.

Given the objective of the report is to make inferences over the training data period in contrast to predicting future matches, no hold-out set will be used to evaluate the model. An implicit assumption of this report is that a player's match performance is "consistent" throughout the year. As a result, an auto-regressive model accounting for player trending over time will not be applicable.

3 Data Exploration

Among the 284 women that played in a WTA match in 2023, match win rates by games played can be observed in Chart 1. Win rate is the proportion of a player's matches with a winning outcome.

Chart 1: WTA Player Win Rate by Games Played in 2023

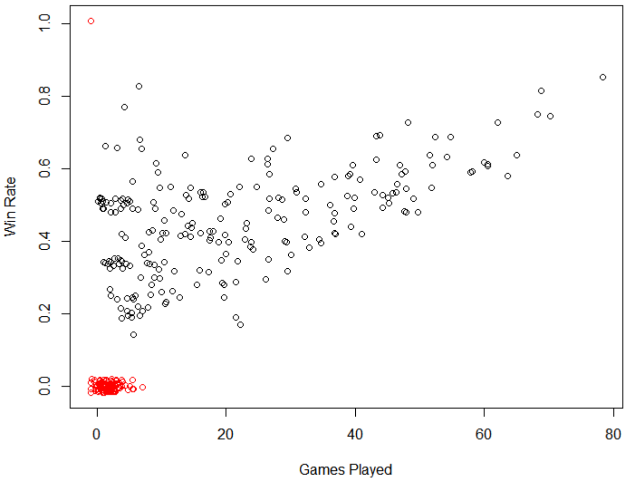
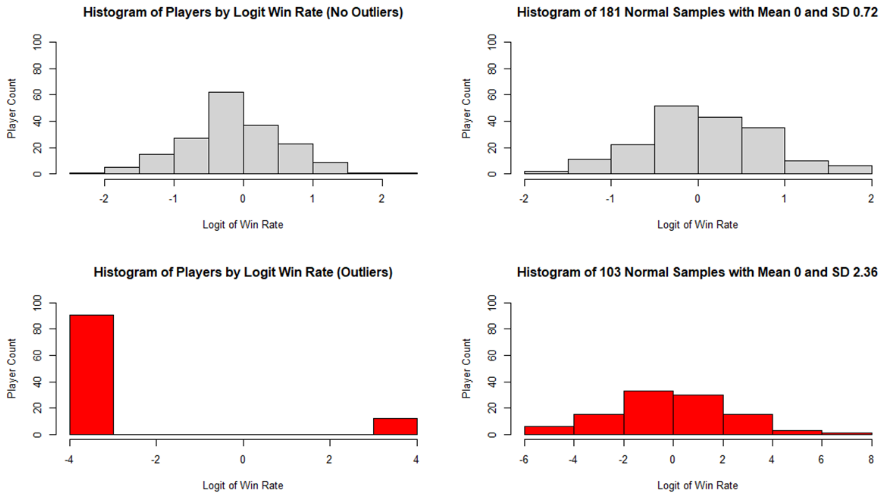


Chart 2: Distribution Assessment of WTA Player Win Rates in 2023



Jitter is added to both games played and win rate to better observe player volume. From the chart, 3 distinct player groups are illuminated - < 10 games played with a win rate of 0, win rate of 1, and win rates between 0 and 1.

As a qualitative anecdote of the group with a perfect win rate, in 2022, Ashleigh Barty announced her retirement from the sport after winning all 11 of her WTA matches, noting she had accomplished all she had set out for and was excited for new ventures outside the sport.

Chart 2 offers separate histograms for the black and red groups; the latter which we may consider a group of "outliers" under the assumption of a single distribution. Furthermore, a random normal sample of the same number of players from the black group appears to approximately fit the observed data (with consistent conclusions upon resampling).

Based on the data observed so far, the most appropriate model of WTA player win rates may be a Bayesian mixture model that accounts for the different groups at play. For this report where mixture models are out of scope, a normal distribution will be extrapolated across all groups. Resulting bias in modelling the "outlier" groups is noted.

4 Model Formulation

The model aims to infer the outcome of each WTA match in 2023 by estimating the parameters of a Bayesian logistic model. These include a score for each player and across each surface (i.e., clay), tier (i.e., ATP 1000), and round (i.e., semifinal). The score is proportional to the logit of their likelihood of winning a match in a setting. The relative difference between the players across these estimated dimensions is then fed into the logistic model to infer the match outcome.

The model has 284 parameters (1 per play) X 4 play dimensions (overall, surface, tier, and round) as well as an intercept. The acceptance criteria of the model be AUC of ~80% on the training set. Once a credible model is fit, associated player parameters can be analyzed to rank players and understand the relative importance of each dimension in determining match outcomes.

A limitation of this model is that players are linearly scored and compared across the dimensions, as opposed to accounting for non-linearities that may arise from specific player interactions. To fully account for the interaction effect between any two players, around 8e4 parameters (284^2) would be required. Given there are only 2491 matches in the dataset, the resulting observed player interactions matrix would be highly sparse. Parameterizing all player interaction combinations would neither be computationally practical nor lead to the creation of a robust model (given the observed data sparsity). A more favorable approach to account for player interaction effects would be to develop a k-dimensional player embedding that distinguished their play style and characteristics, and then include distance metrics between the obtained "latent" dimensions in the match outcome model. For example, if a hypothetical latent dimension is the frequency at which a player uses a drop shot, the distance between two players in a match on this dimension could be computed and included alongside a dedicated coefficient in the logistic model. While the player interaction effect limitation is noted, incorporation of player embeddings and/or other approaches to account for interaction effects will be left out of scope for this report.

In accordance with the above, the hierarchical specification of the model follows:

$Y_i \mid \varphi, \sigma \sim \text{ind Binomial}(\varphi, n=1)\ i = 1{:}n$
$\Phi = b_0 + {}_{k=1:3}\Sigma\ (b_k^{p1} - b_k^{p2})$
$b_1 \sim N(0, 5)$
$b_k^p \sim N(0, \sigma_k)\ k = 1{:}3,\ p = 1{:}284$
$\sigma_k \sim U(0, 5)\ k = 1{:}3$

A binomial/logistic model is applicable as the dependant variable - match outcome - is binary, with a value of

1 indicating a win for player 1.

Player dimensions are sampled from a common normal distribution with a mean of 0 and standard deviation uniformly distributed between 0 and 5. A mean of 0 can be considered as the typical player with a win rate of 0.5 as LOGIT(0.5) = 0, as player score is proportional to the logit of their inferred win rate.

If the sigmoid function is applied to a dimension score above (or below 5), the result is an inferred win rate above 99% (or below 1%), of which no players that played more than 10 games had accomplished in 2023. As such, uniform sampling of a standard deviation with a maximum value of 5 is a simple and reasonable choice.

An alternative hierarchical model with another layer will be evaluated against the baseline model. In the more complex model, each players win rate is now drawn from a unique distribution normal distribution centred around the logit of their win rate in the prior year, and with a standard deviation drawn from an inverse gamma distribution with an initial guess 1/3rd with and uninformative sample size of 10. To account for the additional parameters required for such a model, only the overall player scores and an intercept will be considered in the logistic model. The notable change is that line 4 of the baseline model formulation is expanded into the following:

$b_k^p \sim N(u_1^p, 1/3)$ p = 1:410
$u_1^p \sim N(pu^p, \sigma_1^p)$ p = 1:410
$\sigma_1 \sim \Gamma^{-1} (0.333/2, 0.333*10/2)$

5 Model Fitting and Assessment

Three MCMC chains were run for 5e3 iterations with an additional burn-in of 1e3 iterations each. The aggregate model diagnostics are discussed and presented where applicable below. Model predictions and residuals are calculated using the mean coefficient values across the 15e3 iterations of the simulation.

The assessment of the autocorrelation plots of the player score coefficients for the first 3 players (ordered alphabetically) indicate that the model reached a stationary state without noticeable global trends. Across all player scores, autocorrelation dropped to ~ 0 at lag of 5 and the effective sample size averaged 8812. Assessment of the training observation inference performance across the models can be found in Table 1.

Table 1: Model Evaluation Statistics

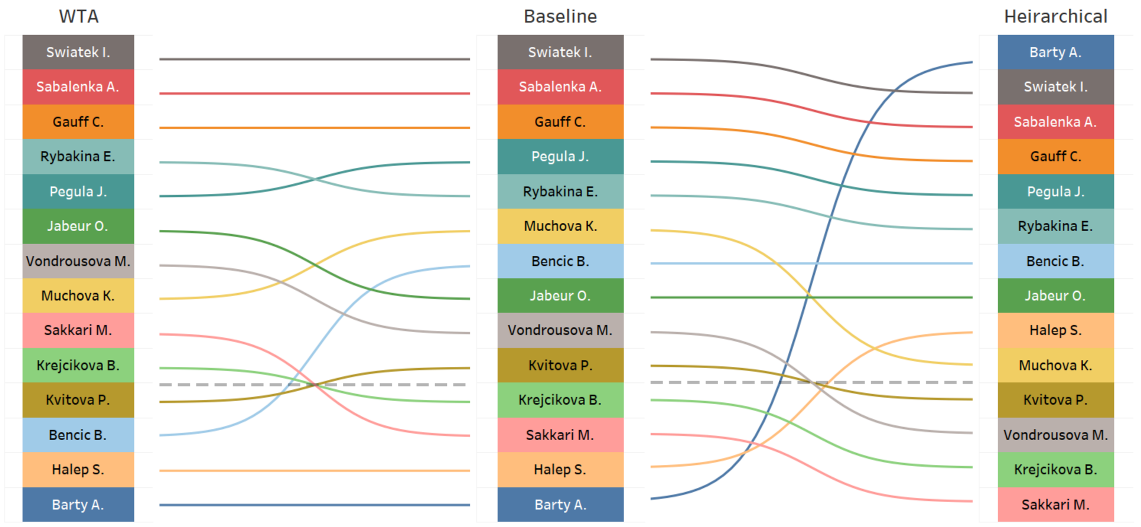
Statistic	Baseline Model	Alternate Hierarchical Model
Accuracy at 50% Threshold	0.7133	0.6860
AUC	0.7899	0.7484
AUC Player Score Only	0.7917	0.7484
DIC	3121 (228.3)	3277 (111)

AUC of the baseline model is ~ 80% which satisfies the acceptance criteria. The model inferences should be interpreted as the expected outcome if both players play at their average given the match specifications (i.e., surface). As a result, a more stringent model acceptance criteria may not be warranted.

Interestingly, although the baseline model incorporates additional dimensions that encode a player's performance across each surface, tier, and round (leading to a the large DIC penalty of 228.3), model predictions obtained by using only the aggregate player scores and intercept performed best as measured by AUC. Based on this and further descriptive analysis of residuals that surfaced a lack of correlation with these match dimensions, the conclusion of this report is that playing surface, tier, and round didn't play a substantive role in the determining the outcome of 2023 WTA matches.

However, it is possible that the inclusion of these features could help distinguish match outcome uncertainty, and/or act as controls to obtain the most accurate player scores. Based on the objective of understanding the most important factors that contribute to match outcomes, the remainder of this paper will focus on analyzing the aggregate player scores obtained from the baseline.

Chart 3: WTA Player Rankings by Model



Highest ranked players listed at the top. For players below the top 10 as indicated by the dashed line, rank differences are only relative.

Chart 3 displays the official top 10 WTA rankings as of November 8th, 2023, alongside the divergence in rankings observed in the baseline and hierarchical Bayesian models (based on the ranking of player scores). Naturally, actual WTA rankings account for more factors than just player win rates in a year, so a one-to-one comparison isn't expected. Starting with the WTA vs. baseline model ranks, the top 3 players have WTA rankings that coincide with the rankings from their inferred win rates. The baseline model reveals a collection of players with inferred win rates that are greater than their official WTA rank, specifically Bencic B. and Muchova K jump by more than two positions. Matches between players whose lines on the chart intersect may offer lucrative betting opportunities, given the predicted winner changes based on whether their WTA ranking, or inferred win rate is compared. As an anecdote, Bencic B. and Sakkari M. played one match since 2022 of which Bencic B. won.

Divergence between the baseline and hierarchical model are indicative of a unique property in the hierarchical model, illuminated by the inclusion of an additional set of players outside the 284 included in the baseline model. Specifically, since the hierarchical model drew a players (mean) win rate from a normal distribution centred around their win rate in 2022, to fully observe the workings of this model., any players that played a match (and thus had a prior mean) in 2022 were considered. As a result, Barty A. who won all 11 of her matches in 2022 but played no matches in 2023 somewhat paradoxically has the highest inferred win rate in 2023. A similar situation holds for Halep S. While including only players with 2023 match data may in a sense lead to the most meaningful model, observing such anomalous results are helpful to interpret the workings of the hierarchical model.

A a useful property of the fit Bayesian model is the ability to infer the probability that a given player's score is higher than another's. The standard deviation of player scores observed in the simulation were widely dispersed between 0.3 to 0.9, indicating that in certain matches, actual differences between player scores may be less likely and hence model inferences may be less accurate. In the 595 matches where player score differential surpassed 1, there was over a 90% inference accuracy, which could be improved when accounting for standard deviations.

6 Conclusion

WTA matches offer a thrilling, competitive battle among top female athletes in the sport. WTA rankings, updated regularly based on match and tournament results, are the standard benchmark used to delineate a player's talent in the game at a point in time. Player talent rankings can be inferred by fitting a Bayesian model to match outcome data, providing a more direct link to relative win likelihoods, and pointing to exciting matchups where relative ranks flip. Interestingly, player talent defined in such a way appears to be agnostic of play parameters. As such, future improvements can be expected from better techniques to model the most important factor in the sport – the player.