Predictive Modeling for Injury Risk and Recovery Time for WNBA Athletes

By: Anahita Hassan *Project Advisor: Prof. Ron Yurko*

Carnegie Mellon University
Statistics & Data Science

Introduction & Research Questions

- The significance of this research lies in: rising WNBA viewership and investment, gender-specific gaps in injury models, and the methodological limitations of current short-term injury prediction methods.
- This study investigates the link between player workload and injury risk in the WNBA using advanced statistical modeling. Specifically, it addresses:
  1. *How does playing time affect injury risk?* → Modeled using **mixed-effects regression** to estimate expected minutes played.
  2. *Can injury risk be predicted between games?* → Framed as a **binary classification** task, using recent workload and performance features.
  3. *Can injury duration be predicted?* → Modeled using **negative binomial regression** for total missed days; **multinomial logistic regression** and **random forest** for categorical recovery periods.

Data Overview & Exploratory Analysis

The analysis uses 2024 WNBA player performance statistics from the wehoop package.[1] The injury data from The Next's WNBA Injury Tracker.[2]

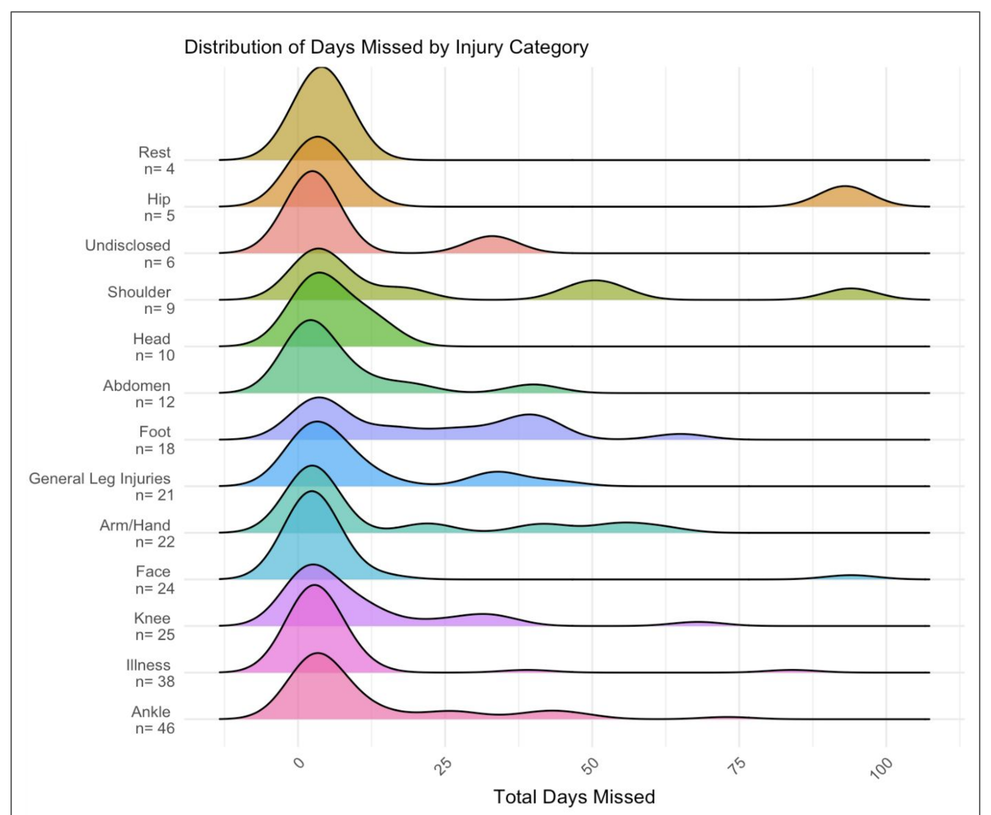

Fig 1. Ridgeline plot of days missed, by injury category.

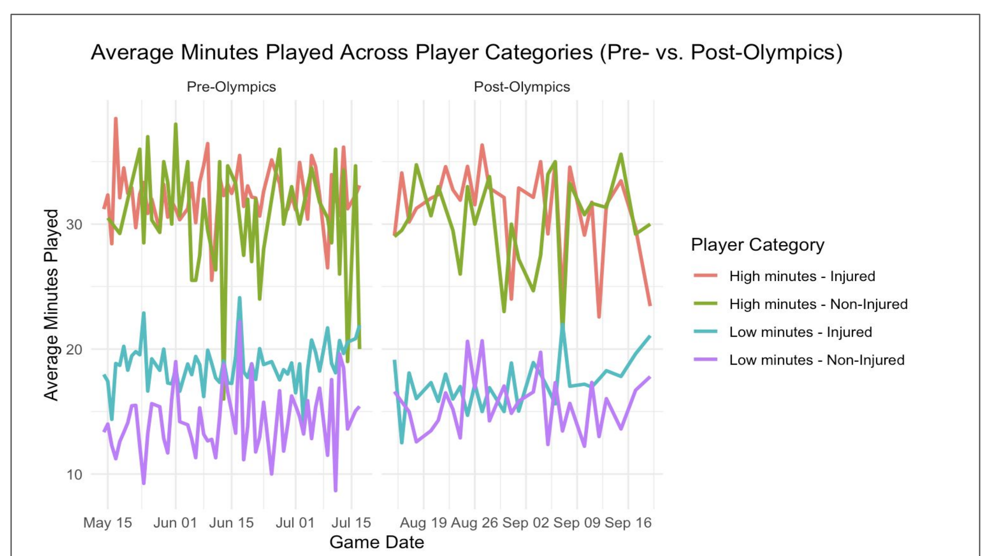

Fig 2. Average minutes played across player subgroups.

- **Injury severity varies within injury type:** Most injuries (e.g., ankle, knee) led to shorter recovery periods. Shoulder, hip, and face injuries were linked to long absences (often 75–100 days) indicating potential season ending injuries and surgeries.
- **Injury trends peak early in the season**: Injuries were most frequent and severe in May and June, with a drop post-Olympics likely due to the approaching playoffs. These injury patterns influenced player minutes.
- **Minutes played and injury risk**: Injured players averaged more minutes per game than non-injured players. Top performers tend to play more, increasing their exposure and, consequently, their risk of injury.

Modeling and Results

Modeling Player Workload

Goal: develop a model that estimates expected playing time using data from non-injured players. The model is then applied to injured players to analyze residuals and detect deviations from expected playing time.

$$y_{ijg} = \underbrace{\beta_1 \cdot (\text{starter status}) + \beta_2 \cdot (\text{position})}_{\text{Fixed effects}} + \underbrace{b_{ig} \cdot (\text{team name}) + b_{jg} \cdot (\text{player name})}_{\text{Random effects}} + \epsilon_g$$

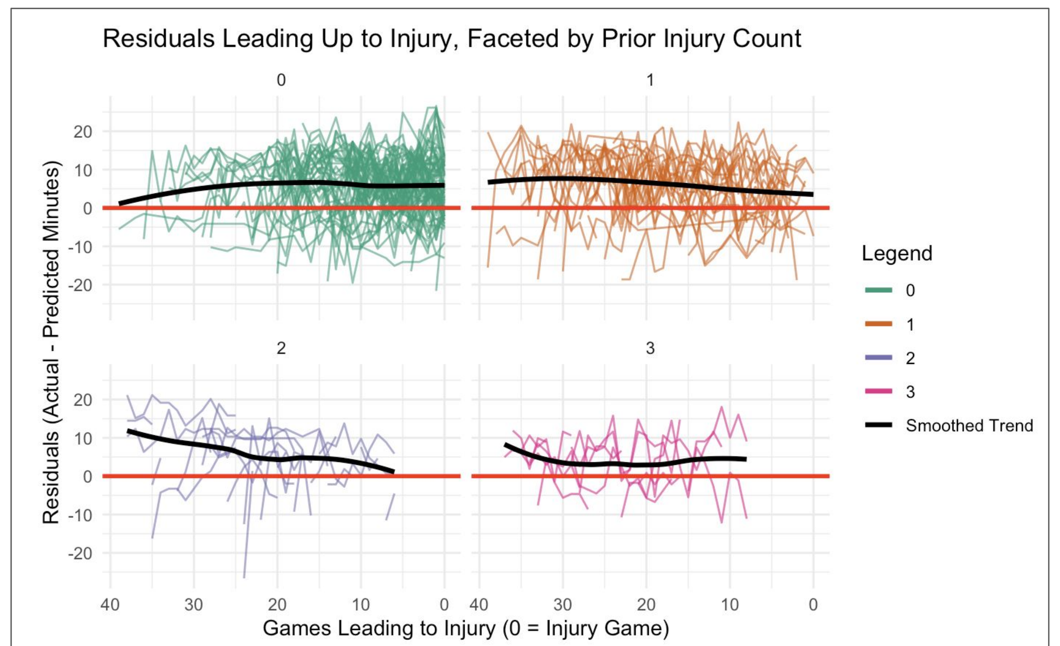

Fig 3. Residuals for games leading up to the injury, faceted by number of prior injuries.

- A three-level linear mixed-effects regression model was used to capture both fixed and random effects (player/team variability).
- The model showed large positive residuals for injured players, especially in games before the injury, suggesting overplaying.

Predicting Injury Risk Likelihood

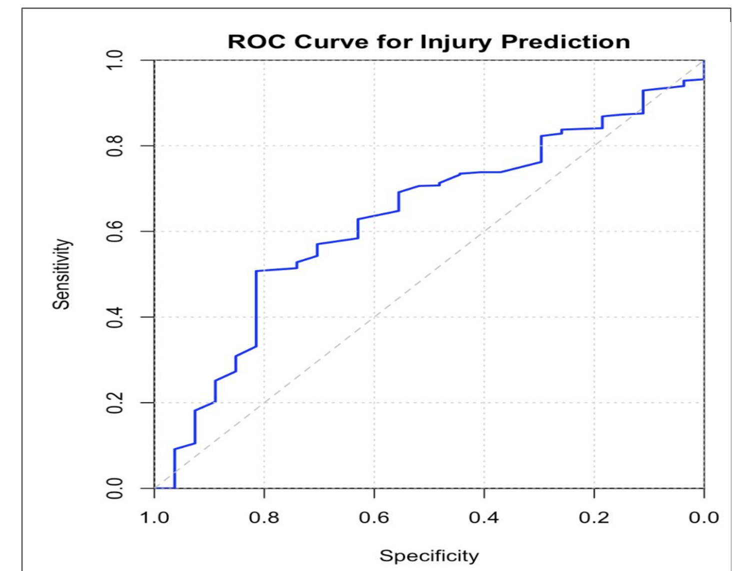

Fig 4. ROC Curve for Logistic Regression Injury Prediction Model (AUC=0.631).

Goal: develop a binary logistic regression model to predict the likelihood of injury between a player's current and next game, using predictors such as minutes and number of prior injuries.

- Model sensitivity = 0.7037, meaning it correctly identifies actual injury events over 70% of the time, which is a promising result given the importance of catching potential injuries.
- Model specificity = 0.5703, meaning it the model falsely flags around 43% of actual non-injuries as potential injuries.

Predicting Recovery Time

- Goal: Predict total days missed based on player stats and injury characteristics.
- Most recoveries are very short: 57% of injured athletes recovered within 0–5 days, and 9% returned with no missed days. Injury severity predictors were stronger predictors of recovery time than player value predictors.
- Negative binomial models outperform linear ones, likely since they handle overdispersion. The multinomial logistic regression (where days missed were categorized into the 4 discrete classes as shown in Fig. 5) performed even better.
- Random forest and multinomial regression models had tradeoffs: Random forest excels at distinguishing Short/Medium recoveries from others, while Multinomial better identifies No/Long recoveries.

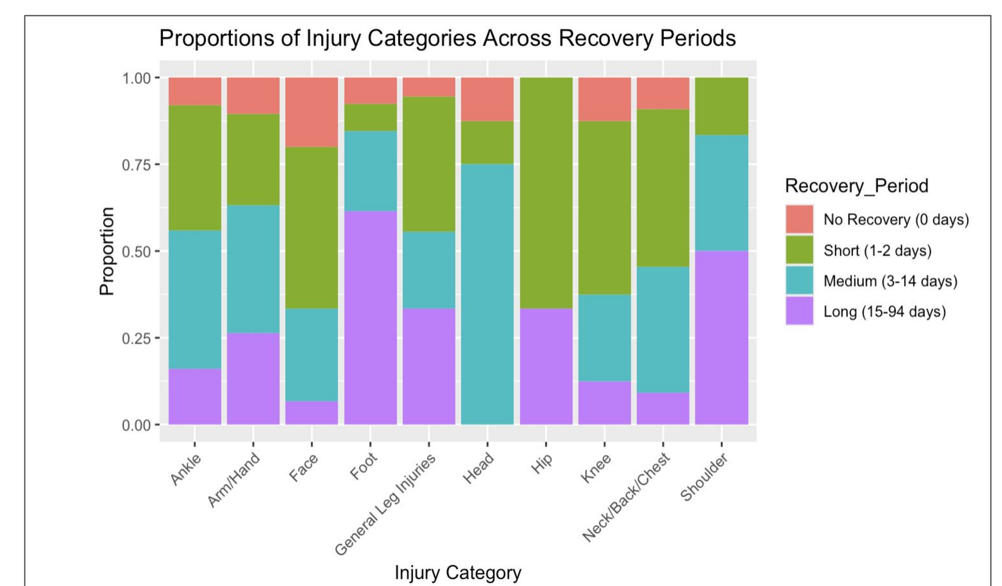

Fig 5. Stacked bar chart depicting proportion of recovery periods based on injury type.

Conclusions

- Minutes played is a strong predictor of injury risk, and the mixed-effects model can be used to predict optimal playing time to potentially prevent injuries due to overuse.
- Likelihood of injury between the current and next game can be predicted using minutes and prior injury history with 57% accuracy, 70% sensitivity, meaning model correctly identifies actual injuries reasonably well.
- Recovery period intervals can also be predicted by multinomial regression and random forest; each model is better at distinguishing different recovery periods.

References

1. Gilani S, Hutchinson G (2024). _wehoop: Access Women's Basketball Play by Play Data_. R package version 2.1.0, <https://CRAN.R-project.org/package=wehoop>.
2. Seehafer, L. (2023, July 5). *WNBA Injury Tracker: Who gets hurt, how often, and why it matters*. The Next. https://www.thenexthoops.com/wnba/wnba-injury-tracker-who-gets-hurt-how-often-and-why-it-matters/

Acknowledgements

Special thanks to Prof. Ron Yurko and Dr. Joseph Devine for their guidance and feedback throughout the research process.