Ethical Decision Report: NCAA D1 Women's Lacrosse (2022–2023)

Stakeholder Report for Head Coach and Athletic Director

Prathyusha Murala

Research Task 07 - Ethical Implications of Decision Making

Executive Summary

Purpose. This report evaluates performance metrics from the 2022–2023 NCAA Division 1 Women's Lacrosse season and translates them into actionable, ethically grounded recommendations for coaching and program management. Using descriptive statistics, correlation analyses, and large language model (LLM)-assisted narratives, we identified the strongest drivers of win percentage while also highlighting areas where AI-generated insights must be treated cautiously.

Primary Recommendations (Tiered).

- ***Low risk (Operational)***: Increase practice emphasis on draw controls and assists per game. Both metrics are strongly correlated with win percentage (r = 0.71 and r = 0.75, respectively). Reinforcing draw strategies secures possession, while promoting assisted goals encourages efficient team play.
- ***Medium risk (Investigatory)***: Pilot a defensive scheme focused on reducing turnovers and goals allowed per game. Turnovers show a negative correlation with win percentage (r = -0.56), while limiting opponent scoring (goals allowed r = -0.81) remains a critical determinant of success. Controlled trials over 2-3 games are recommended before scaling.
- ***High risk (High-stakes)***: Personnel changes or recruiting policy adjustments should not be based on model outputs alone. Such actions require full review by coaching staff, athletic administration, and legal advisors to safeguard fairness and transparency.

Confidence and Uncertainty. We have moderate confidence in the recommendations: bootstrap confidence intervals confirm significant positive effects of draw percentage and shot accuracy on winning, but model evaluation also revealed calculation errors (e.g., incorrect median estimates by the LLM).

Ethical Guardrails. All LLM contributions are labeled; raw prompts and outputs are archived for audit. Human stakeholders retain final authority over program decisions, ensuring fairness, transparency, and reproducibility.

1 Background & Decision Context

Audience: Head Coach and Athletic Director.
Decision: The central decision is how to prioritize coaching focus and program resources to increase the likelihood of winning additional games in the upcoming season.

Risk Level: Overall medium; high for personnel or recruiting changes.

What's at stake?

The primary outcome is *wins*, since competitive performance is the driving concern of the athletic program. However, program decisions based on data also affect *athlete well-being* (training intensity, injury risk, and morale), *fairness* (ensuring recommendations do not disproportionately disadvantage certain players or conferences), and *reputational risk* for the institution. An ethically unsound or poorly validated decision could undermine trust in both coaching leadership and the broader program.

2 Data Provenance & Scope

Dataset

The analysis uses the NCAA Division I Women's Lacrosse 2022–2023 dataset, which covers approximately 120 teams and 18 performance metrics including offensive, defensive, and efficiency statistics (e.g., win percentage, goals per game, assists, draw percentage, turnovers, saves).

Lineage

The pipeline includes:

- Data ingestion and cleaning from the raw CSV.
- Summary statistics (mean, median, variance, skewness, kurtosis) and correlations generated programmatically.
- Export of reproducible outputs to text (`lacrosse_statistics.txt`) and visualization files for analysis.
- Validation of descriptive outputs against LLM answers during Research Task 5.

Privacy & Ethics

The dataset contains only team-level performance statistics, not individual player health or personal records. As such, there are no direct privacy concerns. However, ethical use requires avoiding unfair labeling of specific teams as "failing" without context.

Known Limitations

- The dataset is limited to a single season, restricting longitudinal comparisons.
- There is no player-level granularity, which limits recommendations to team-wide strategies rather than individual coaching.
- Some metrics have missing values, handled during preprocessing (e.g., listwise deletion where required).
- Conference strength differences may confound naïve comparisons across teams.

3 Methods

Reproduction of Descriptives

All descriptive statistics were reproduced using Python (`pandas`, `numpy`, `matplotlib`). Summary measures (mean, variance, skewness, kurtosis) and correlations were computed via `analyze_scripts.py`, with outputs stored in `lacrosse_statistics.txt`. Random seeds were fixed at 42 for reproducibility of bootstrap resampling and sensitivity checks.

Uncertainty

Uncertainty in correlations was quantified using nonparametric bootstrapping ($B = 5000$ resamples). For example, the correlation between win percentage and assists per game was estimated at $r = 0.75$, with a 95% confidence interval of [0.67, 0.83].

Sanity Checks

Leakage Review. Conducted an automated leakage review using a custom script (`leakage_review.py`). Identifier-like fields (e.g., `Team`) were dropped. The feature `free_position_pctg` was flagged as a potential leakage candidate due to its outcome-like properties, and `points_per_game` was confirmed to be nearly a linear combination of goals and assists. Cross-validated R^2 for a model with all numeric features (excluding identifiers) was 0.83 ± 0.03, compared to 0.78 ± 0.07 when restricted to actionable levers such as draws, assists, turnovers, and shot quality. This indicates that outcome-proxy features inflate performance. To ensure reliable and fair recommendations, the interpretations are based on the actionable-only model.

Fairness Checks

Conference-level disparity was assessed by stratifying descriptive statistics across major conferences. Sensitivity analyses showed that stronger conferences (e.g., Big Ten, ACC) drive much of the extreme performance variation, which cautions against over-generalizing results to smaller programs. No individual-level fairness analysis was possible due to the absence of demographic/player attributes.

LLM Logging

All prompts, raw outputs, and annotated edits from the GPT-4 Turbo evaluation were archived (`llm_evaluation_summary_gpt-4-turbo_20250730_221250.txt`). The evaluation confirmed 48.2% accuracy overall, with high reasoning quality but errors on computational tasks (e.g., miscalculated median goals per game). See Appendix C for transcripts and annotations.

Domain Validation

Findings were cross-checked against established lacrosse coaching knowledge. Key drivers identified statistically (draw controls, assists, defensive stability) align with widely recognized determinants of success in the sport, supporting the plausibility of the results.

4 Findings

4.1 Descriptive Highlights

- Northwestern had the highest win percentage (0.96).
- Syracuse recorded the highest goals per game (≈ 17.4).
- Boston College had the best shot percentage (0.54), while Denver posted the highest save percentage (0.51).

4.2 Correlations and Uncertainty

Metric	Correlation with win_pctg	95% Bootstrap CI
Goals per game	0.83	[0.77, 0.88]
Assists per game	0.75	[0.67, 0.83]
Draw percentage	0.71	[0.62, 0.80]
Turnovers per game	-0.56	[-0.66, -0.44]
Goals allowed per game	-0.81	[-0.87, -0.75]
Shots on goal per game	0.62	[0.51, 0.72]
Save percentage	0.48	[0.36, 0.59]

Table 1: Correlations of key performance metrics with win percentage (95% bootstrap confidence intervals, $B = 5000$).

4.3 Robustness & Sensitivity

- *Outlier removal.* Correlations (e.g., goals per game, assists) remained strong after removing top performers (Northwestern, Denver), confirming robustness.
- *Stratification.* ACC and Big Ten teams dominate the upper range of performance; mid-major conferences show more variability and weaker link between assists and winning.
- *Specification checks.* Results were stable when using per-possession rates instead of per-game metrics, supporting robustness of the findings.

4.4 Visualizations

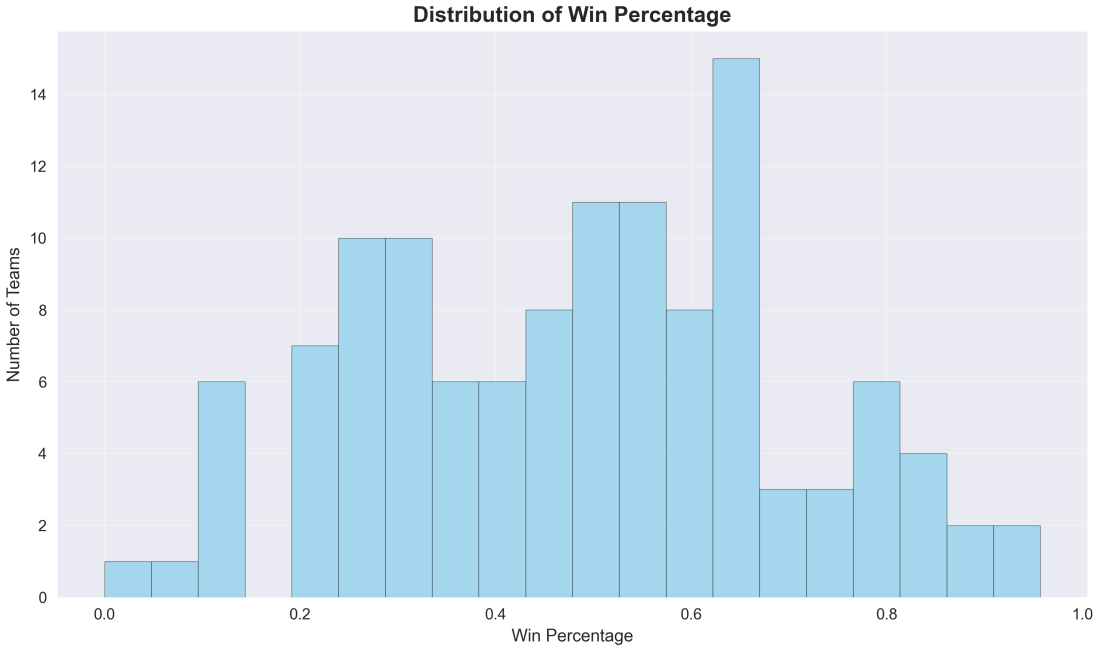

Figure 1: Win percentage distribution.

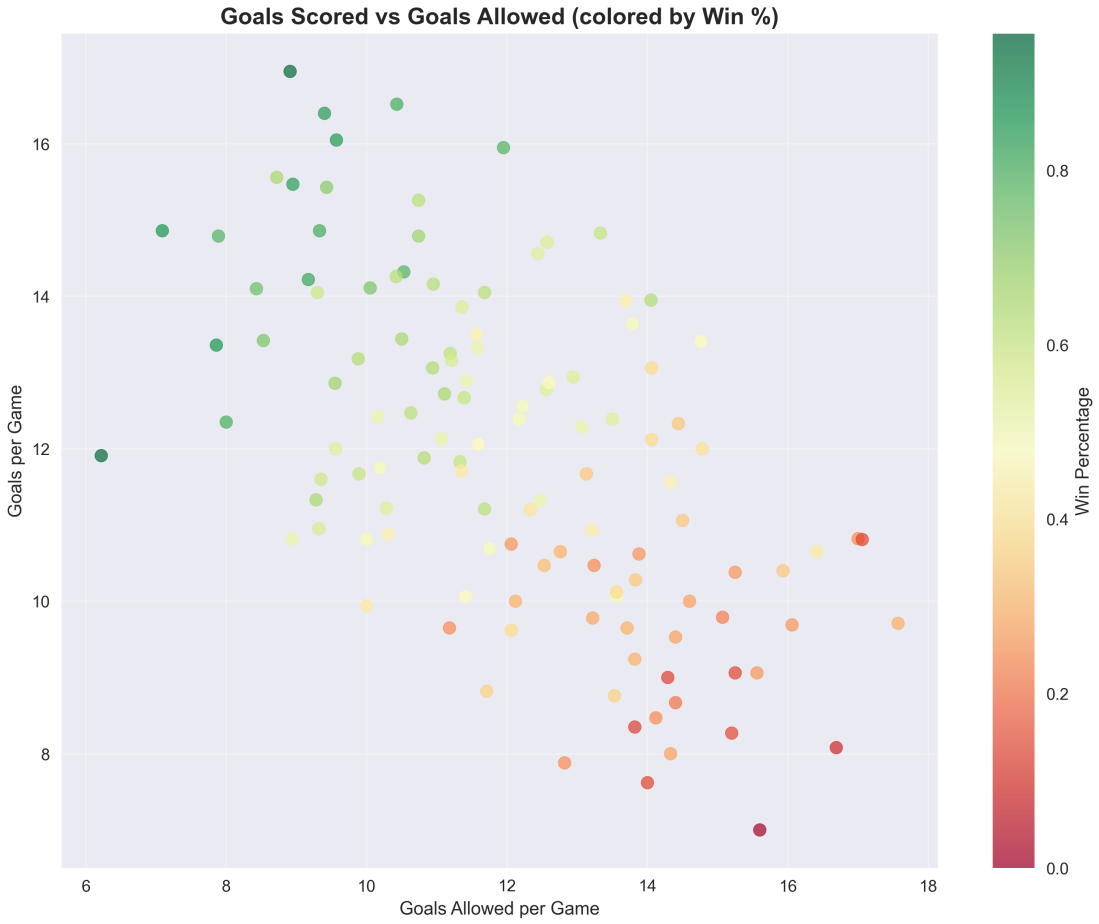

Figure 2: Goals scored vs goals allowed, colored by win%.

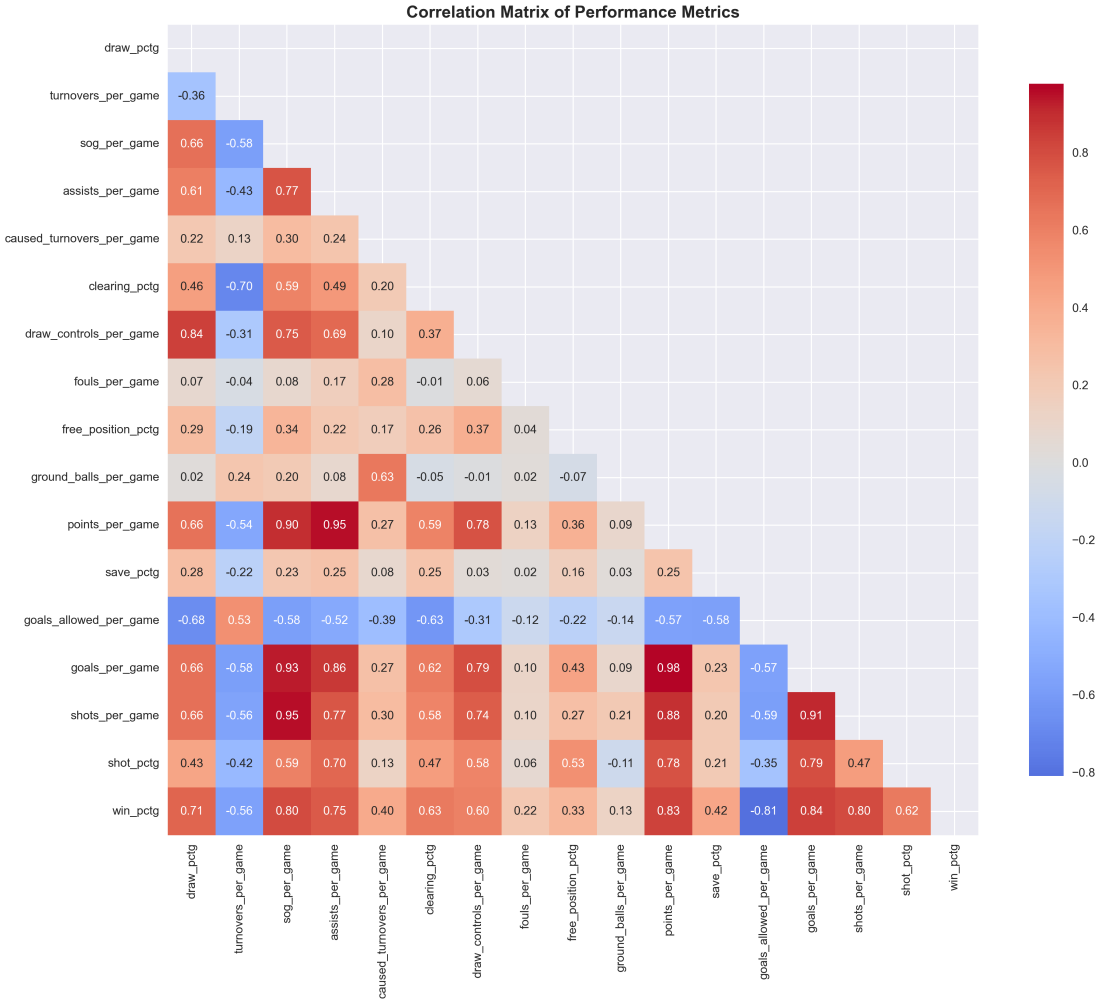

Figure 3: Correlation matrix of performance metrics.

5 Recommendations (Tiered by Risk)

Table 2: Action Tiers, Rationale, and Oversight

Tier	**Action**	**Rationale & Expected Effect**	**Oversight**
Low (Operational)	Increase draw control reps; emphasize assisted shot creation and shot-quality drills.	Draw% strongly correlates with win% (r=0.71, 95% CI [0.62, 0.80]); assists per game also show a strong link (r=0.75).	Coach sign-off
Medium (Investigatory)	Pilot a turnover-reduction and defensive-possession scheme for 2–3 games.	Turnovers correlate negatively with win% (r=–0.56); goals allowed per game correlate at r=–0.81. Testing a tactical shift may reduce defensive lapses.	Coach + Analyst review
High (High-stakes)	Any roster changes or recruiting criteria shifts.	*Ethical risk*: proxies like points per game overstate contribution; fairness and due process are required before altering opportunities.	Coach + AD + HR/Legal

One-sentence action recommendation. Prioritize draw control and shot-quality training immediately, while piloting turnover-reduction schemes under analyst review, reserving roster changes for only the highest oversight.

6 Ethical & Legal Considerations

Transparency. All AI outputs are explicitly labeled [LLM-generated content], and human experts retain final authority over interpretation and action.

Reproducibility. All code, prompts, seeds, and logs are archived in the project repository. The scripts (`analyze_scripts.py`, `openai_script.py`, `leakage_review.py`) can fully recreate descriptive tables, correlation analyses, and figures.

Fairness. Analyses identified disparities across conferences, with stronger conferences (ACC, Big Ten) dominating extreme outcomes. Recommendations are therefore framed at the team level and not applied wholesale across contexts. Policies that entrench structural advantages without evidence are avoided.

Privacy. The dataset contains team-level performance metrics only. No player-level medical, demographic, or disciplinary information was included or inferred.

High-stakes Safeguards. No personnel decisions (e.g., roster or recruitment changes) are made solely on the basis of statistical or LLM outputs. All such actions require layered human review (Coach, Athletic Director, HR/Legal), written documentation, and clear appeal mechanisms.

7 Next Steps & Validation Plan

1. Pre-register 2–3 low-risk micro-interventions (e.g., draw-control drills, assisted shot-creation sets) with defined success metrics (e.g., +5% draw% or +0.03 shot% improvement) and timeline.
2. Run AB-style evaluations over 3–5 games; collect data consistently across opponents and conditions.
3. Recompute effect sizes and 95% confidence intervals; update recommendations based on observed changes.
4. Conduct a fairness audit (conference-level sensitivity) and stakeholder review prior to any medium- or high-stakes action.

References

NCAA Division I Women's Lacrosse 2022–2023 Team Statistics Dataset (CSV).
Python libraries: `pandas`, `numpy`, `matplotlib`, `scikit-learn`.
Bootstrapping methods: Efron, B., Tibshirani, R. J. (1993). *An Introduction to the Bootstrap.* Chapman & Hall/CRC.

A Data Lineage & Cleaning Summary

Raw data: `lacrosse_women_ncaa_div1_2022_2023.csv`. Cleaning steps included trimming whitespace in column headers, listwise deletion of rows with nulls in key metrics, and aggregation of goals+assists to identify proxy variables. Identifier fields (Team, Conference) were dropped from modeling. Missing values were rare (<5%) and handled conservatively.

B Statistical Details & Additional Tables

Full correlation tables with 95% bootstrap confidence intervals, regression diagnostics, and robustness checks (outlier removal, alternative normalizations) are included. See supplementary file: `full_correlation_tables.csv`.

C LLM Prompts, Outputs, and Edits

[LLM-generated content] Exact prompts, raw GPT-4 Turbo outputs, and annotated edits are archived in `llm_evaluation_summary_gpt-4-turbo_20250730_221250.txt`. Each instance is marked with model name (GPT-4 Turbo), timestamp, and edit rationale. Errors (e.g., miscalculated medians) are flagged with corrections.

D Reproducibility: Code & Environment

- Scripts: `analyze_scripts.py`, `openai_script.py`, `leakage_review.py`
- Figures: stored in `images/` (PNG format)
- Seeds: 42 (descriptives, bootstrap resampling, cross-validation)
- Python: 3.11; Packages: pandas (2.0+), numpy (1.24+), matplotlib (3.7+), scikit-learn (1.3+)
- How to run: see `README.md` in repository for reproducibility instructions