

Linköping Electronic Conference Proceedings Nr. 221

Edited by:

Tim Brecht

Niklas Carlsson

Adam Almqvist Richloow

Mikael Vernblom

Patrick Lambrix



**Proceedings of the
5th International Ice Hockey Analytics Conference
LINHAC 2026
Research Track and Community Notes**

ISBN 978-91-8118-651-2 (PDF)

ISSN 1650-3686

eISSN 1650-3740

<https://doi.org/10.3384/ecp221>

Organizing Committee

Patrick Lambrix, Linköping University, Sweden (chair)
Adam Almqvist Richloow, Swedish Ice Hockey Federation, Sweden (co-chair)
Tim Brecht, University of Waterloo, Canada (co-chair)
Niklas Carlsson, Linköping University, Sweden (co-chair)
Mikael Vernblom, 49ing, Sweden (co-chair)

Program Committee for the Research Track and Community Notes

Tim Brecht, University of Waterloo, Canada (chair)
Niklas Carlsson, Linköping University, Sweden (chair)
Martin Arlitt, University of Calgary, Canada
David A Clausi, University of Waterloo, Canada
Jack Davis, University of Waterloo, Canada
Ulf Johansson, Jönköping University, Sweden
Rodney Paul, Syracuse University, USA
Rob Pettapiece, Former R&D Director, Toronto Maple Leafs, Canada
Miles Pitassi, NHL Data Analytics Group, USA
Joshua Pohlkamp-Hartt, Boston Bruins, USA
David Thomas Radke, Chicago Blackhawks, USA
Michael Schuckers, University of North Carolina Charlotte, USA
Timo Seppa, Quinnipiac University and NHL Network Radio, USA
Sean Tierney, Ottawa Senators, Canada
Mikael Svarén, Dalarna University, Sweden
Andrew C Thomas, Detroit Tigers, USA
Sam Ventura, Buffalo Sabres, USA

In cooperation with



E.H.C. Alliance of European Hockey Clubs

In cooperation with



Sportlogiq

Sponsor



The Swedish Research Council for Sport Science

Bronze industry collaborators



Bronze industry collaborators



Preface

LINHAC 2026 took place June 2-4, 2026, and was organized by Linköping University. LINHAC brought together professionals and academics with an interest in hockey analytics. It featured the latest research in hockey analytics in academia and companies, discussions with analysts and coaches, industry sessions with the latest hockey analytics products, and an analytics competition for students.

In addition to the research track and community notes, the program included invited talks. There were several talks related to technical aspects. Brendan Kumagai (Teamworks) talked about new frontiers in ice hockey analytics with player tracking data. In their talk "From Capture to Decision: Building the Modern Sports Intelligence Stack", David Cook and Morgan Zeba (Spiideo) led us through the whole workflow from capturing the data all the way to data-driven decision making. Timo Seppa (Hockey Data Show on NHL Network Radio and Quinnipiac Bobcats) taught us about the importance of relevant and correct data for analytics in his talk "Garbage in, garbage out: What's in your analytics?" David Radke (Chicago Blackhawks) discussed what ice hockey analytics can learn from AI, while Ola Lidmark Eriksson (Playmaker AI) talked about how insights from football can be used in ice hockey.

Further, there were several talks on the use of analytics. Veronica Eriksson (Swedish Ice Hockey Federation) discussed questions such as "What can analytics do for my team? How do I start? How much effort is needed?" in her talk "Coach's challenge: Own your data. Control your AI". Erik Wilderoth (Färjestad BK) provided insights on how to use analytics for data-driven player development, and Felix Dahlin (Swedish Ice Hockey Federation and MODO Hockey) discussed behavior-driven development fueled by analytics. In the talk "From Goalie-Driven Wins to System Stability: A Data-Led Defensive Reset in the National League", Andreas Hänni (49ing) and Thierry Paterlini (SCL Tigers) highlighted how applied analytics can move beyond evaluation to directly inform coaching decisions and improve team robustness. Erik Lignell (HC Fribourg-Gottéron SA) argued that successful analysis is about communication, not location: "You can work remotely and still be close to the team". Oscar Nordfjell (d-fine) presented a project using analytics for the U18 Swedish national team in collaboration with Adam Almqvist Richloow from the Swedish Ice Hockey Federation. As guests representing the sport of bandy, Mattias Sjöholm and Per Gullbrand shared their experiences using data to build championship winning teams. They showed us how to take the pulse of a 23-man roster to see if they are truly pulling in the same direction.

Analytics is also used for the business side of ice hockey. Ray Lalonde (Ray Lalonde Sports+) shared insights about the use of analytics from his long experience in helping organizations in the sports industry grow in his talk "From Competitive Edge to Core Infrastructure: The Global Rise of Sports Analytics". Stefan Lavén (Data Talks) discussed how to create revenue from fans following a team.

Further, Henric Larsson and Almen Bibic (TV4 Sweden) discussed how producers and

commentators prepare for hockey game broadcasts and how they use stats and analytics during the games.

There were also talks related to the future of ice hockey analytics. Andreas Hänni (49ing) shared his vision for the use of data in ice hockey in his talk "Data use beyond analytics", while Neil Lane (Stathletes) discussed the next frontier of ice hockey analytics from his point of view. Caroline Jonsson from the Alliance of European Hockey Clubs discussed the importance of events such as LINHAC to grow the field of ice hockey analytics, to expose more teams to analytics, and thereby to grow ice hockey in general.

Further, there were panel discussions moderated by Mike Helber. In the first panel, Tim Brecht (University of Waterloo), Matt Hamann (Nashville Predators), Miranda McMillan (Montreal Canadiens), David Radke (Chicago Blackhawks), and Campbell Weaver (Boston Bruins) discussed analyzing NHL data and their workflows. In the second panel, Miika Arponen (Tampere Ilves), Jan Morkes (Rytíři Kladno and Czech Ice Hockey Federation), Jaroslav Bezdek (Sparta Prague), Erik Lignell (HC Fribourg-Gottéron SA), and Karl Malmquist (Linköping Hockey Club) discussed analytics workflows in European teams.

During the mingle event we received an NHL play-offs update from Mike Kelly (Sportlogiq and NHL Network).

Finally, there was a student competition where the task was to provide insights based on sequences of events in a hockey game. Data was provided by the SHL and Sportlogiq. A jury consisting of Brendan Kumagai (Teamworks), David Radke (Chicago Blackhawks), Mikael Vernblom (49ing) and Patrick Lambrix (Linköping University) selected the project *Predicting Shot Surface Distributions from Zone Entry Characteristics* by Kiran Roye as the winner of the competition.

LINHAC is the only conference of its kind in Europe, and to our knowledge, it is the only hockey analytics conference that covers all aspects related to hockey analytics. These proceedings include the papers from the research track as well as community notes. The research track program committee with chairs Tim Brecht (University of Waterloo) and Niklas Carlsson (Linköping University), selected the research paper *A Bayesian Approach to Estimating a National Hockey League Draft Value Pick Chart with Bounds* by Yufei Zou, Elliot Kervin, and Michael Schuckers as the winner of the best research paper award.

A companion book includes in addition to the research track and community notes, contributions from industry, the student competition papers, as well as insights from contributors to LINHAC regarding their experience with hockey analytics and thoughts about its future.

We thank our moderator Mike Helber, our conference service company Way, and the members of our local organization committee Mina Abd Nikooie Pour, Ying Li, Jenny Rydén, Lene Rosell, Veronica Gunnarsson, Anders Cronstierna, and Daniel Jemander, for their excellent support.

Last, but not least, we thank our collaborators the Alliance of European Hockey Clubs and Sportlogiq, our sponsor the Swedish Research Council for Sport Science, and our bronze industry collaborators 49ing, Data Talks, Playmaker AI, and Spiideo.

September 2026

Patrick Lambrix (chair),
Adam Almqvist Richloow (co-chair),
Tim Brecht (co-chair),
Niklas Carlsson (co-chair),
Mikael Vernblom (co-chair)

Contents

Research papers	1
Closing the Gap: A Longitudinal, Covariate-Adjusted Analysis of NHL Goalie Pull Timing <i>by Saif Al-Din Ali, Hanan Mohamed</i>	2
Multi-Horizon Career Longevity Prediction for NHL Skaters: Landmark Logistic Regression Reveals Stage-Specific Risk Factors <i>by Elizabeth J. Goldstein, Jessica A. Pearson</i>	16
Who's In by Game 15? Multi-Model Per-Game Forecasting of NHL Playoff Qualification <i>by Quinton J. Krueger</i>	36
Validating Expected Possession Value Against Expert Coach Assessments of Passing Intelligence in Elite Youth Ice Hockey <i>by Miikka Palvalin</i>	48
Climate and Income Effects on U.S. Hockey Participation <i>by Rodney J. Paul, Nick Riccardi, Gina Pauline, Andrew Weinbach</i>	61
Non-Transitivity in the National Hockey League: Ranking Teams, Lines, and Players using Game Theory <i>by David Radke, Troy Mulholland</i>	77
NHL Home-Ice Offensive and the Power Play: Exploring Differences Among Game Contexts <i>by Nick Riccardi, Tim Campbell Jr., Rodney J. Paul</i>	93
A Bayesian Approach to Estimating a National Hockey League Draft Pick Value Chart with Bounds <i>by Yufei Zou, Elliot Kervin, Michael Schuckers</i>	107
Short papers and community notes	123
Defining Physicality and Visualizing Skaters' Ability to Play Through Contact <i>by Louis Boulet</i>	124
Net Man-Games Lost as a Coaching and GM Evaluation Metric in the NHL <i>by Quinton J. Krueger, Toby Velte, Doug Plocki, Nick Carone</i>	136

CONTENTS

Introducing a Gaussian Rink Control-Based Expected Threat Framework to Hockey <i>by Michele Sezgin, Tyler Quinn</i>	147
--	-----

Research papers

Closing the Gap: A Longitudinal, Covariate-Adjusted Analysis of NHL Goalie Pull Timing

Saif Al-Din Ali¹ and Hanan Mohamed¹

University of Waterloo, Waterloo, ON, Canada
{sa29ali,hmohamed}@uwaterloo.ca

Abstract. Down a goal in the closing minutes, an NHL coach can swap his goaltender for a sixth attacker – six skaters, no net, against the opponent’s regular five. Eight regular seasons of play-by-play between 2015–16 and 2024–25 give us 7,047 such decisions. We pose four questions. Does pulling earlier actually produce more goals once we adjust for opponent, home ice, rest, and the rest of the obvious covariates? Have coaches as a group drifted earlier? Has the textbook one-goal optimum – the 2.75 min from Beaudoin and Swartz – itself moved under modern scoring rates? And if some franchises have closed the theory–practice gap, is the closure a league-wide phenomenon or the work of a small vanguard? Four causal estimators (covariate-adjusted regression, IPW, AIPW, fixed-effects Poisson) agree that earlier one-goal pulls raise post-pull scoring substantially; the E-value sits at 6.12. A simplified continuous-time Markov chain re-solved at 2024-era scoring rates places the one-goal optimum about a minute earlier than the original 2005 calibration – the target itself has drifted. League mean pull timing has crept earlier at six seconds per season and the pooled distribution stays unimodal, but team-level early-pull shares span 26–40 percentage points, worth between 0.5 and 3.2 standings points per team per season at the extremes. The vanguard has closed the gap; the median NHL franchise has not.

Keywords: hockey analytics · goalie pull · empty net · coaching decisions · NHL · propensity score · sensitivity analysis · causal inference

1 Introduction

Few coaching decisions in any sport are this clean. Pulling the goalie is a yes/no choice with a fixed alternative – six skaters and no net for the closing minutes, against the opponent’s normal five – and the resolution arrives inside a short, fully observable window. That is why the decision has attracted formal decision-theoretic treatment for fifty years [2, 1, 3], longer than most analytics questions in hockey have existed.

Beaudoin and Swartz [1] produced the number every later analyst cites. They fit a continuous-time Markov chain (CTMC) to early-2000s scoring rates and

recommended pulling at 2.75 min remaining when down one; the empirical mode at the time sat near a single minute. Fifteen years later, popular accounts [9,10] claim coaches have closed the gap. Nobody, as far as we can tell, has put that claim through covariate adjustment at league scale, separated it from the NHL’s expansion from 30 to 32 teams during the same window, or asked whether the theoretical target itself has moved as scoring rates have climbed.

We chase four questions. (1) Does the early-versus-late timing effect survive covariate adjustment, and how sensitive is it to confounders we cannot measure? (2) Once team baselines are absorbed, has mean pull timing actually moved across the eight seasons of our sample? (3) Is the league converging on a common optimum, or has the shift concentrated in a handful of franchises? (4) What is the decision-theoretic payoff of closing the gap, expressed in standings points per season?

Four things in this paper are new. First, a pull-detection pipeline keyed on the public NHL API’s `situationCode` field, large enough to support league-scale causal estimation. Second, four estimators of the timing effect (covariate-adjusted regression, IPW, AIPW, fixed-effects Poisson) reported alongside a DAG, balance/overlap diagnostics, an E-value, a placebo on non-trailing pulls, and a flexible-spline conservative check. Third, a simplified CTMC re-solved at 2024-era scoring rates, which places the theoretical one-goal optimum about a minute earlier than the original 2005-era target. Fourth, a team-level decomposition of the league average with cluster-bootstrap confidence intervals, translated into expected standings points. Code and pinned dependencies are released for replication.

2 Related Work

The theoretical line starts with Morrison [2], who treated post-pull scoring as two competing Poisson processes and solved analytically for break-even pull times. Zaman [3] refreshed the rates a quarter-century later. Beaudoin and Swartz [1] generalised the setup to a CTMC capable of handling multi-goal deficits, and their 2.75-minute one-goal answer has been the standard reference number ever since. The empirical side has been thinner. Paine [9] and later Younggren and Younggren [10] ran descriptive accounts of what coaches were actually doing, but neither pushed past per-season averages and neither appeared outside the analytics-blog ecosystem. Schuckers and Curro [4] and Pettigrew [8] built the within-game win-probability machinery that subsequent analysts have leaned on; propensity-score and bootstrap inference in sports causal work [7] and the E-value [6] have since become standard tools. What no one has done is run the theory–practice gap through covariate-adjusted estimators at league scale, decompose the league average into its team components, and put a standings-point number on the result. That is this paper.

Table 1. Trailing goalie pulls by deficit size, with post-pull outcomes. All percentages are within row.

Deficit	<i>n</i>	% Trail.	% Scored	% Erased	% Won	% Conceded EN
Down 1	3,066	51.1	20.1	20.1	9.2	44.9
Down 2	2,226	37.1	22.7	4.4	1.5	56.3
Down 3+	711	11.8	36.8	1.5	0.3	53.7
Total	6,003	100	23.1	12.1	5.3	50.2

3 Data and Methods

3.1 Data Collection

We pulled data from the NHL public API (api-web.nhle.com) in two stages. The first stage queried `/v1/club-schedule-season/{team}/{season}` for every team in each of our eight target seasons (2015–16 through 2024–25, with the COVID-disrupted 2019–20 and 2020–21 campaigns dropped; eight is the longest contiguous window for which the current PBP endpoint returns complete `situationCode` fields, and both COVID seasons featured shortened schedules and divisional realignment). The schedule pull returned 10,250 unique games.

Stage two fetched PBP detail via `/v1/gamecenter/{id}/play-by-play`. Each event carries a `situationCode` field of the form `XXYY`, encoding the away and home on-ice states respectively (a zero in either goalie digit indicates an empty net). For each game we scanned third-period events chronologically and recorded the first empty-net transition along with its time remaining, the score at the pull, and the post-pull trajectory. A second, narrower fetch kept *all* pre-pull events in the trailing one- and two-goal subsample so we could build third-period-to-pull Corsi, shots on goal, and an xG proxy.

3.2 Dataset Summary

Our detector flagged pull events in 7,047 of the 10,250 games. The trailing subset – 6,003 pulls, 85.2% of the detected events – forms the analysis sample; the remaining 1,044 are overwhelmingly delayed-penalty artefacts in tied or leading states. Our detection rule cannot fully separate a deliberate extra-attacker pull from a delayed-penalty pull when trailing (both register identically in `situationCode`). We restrict to trailing pulls, where delayed-penalty events are rare (the puck must be in the offending team’s defensive zone, the wrong end for an extra-attacker camp), and additionally exclude pulls with $2 < t \leq 2.5$ min remaining to sharpen the early-versus-late contrast (dropping 455 of 3,066 down-1 pulls, 14.8%, leaving 2,611 for IPW/AIPW). Table 1 summarises trailing pulls by deficit.

3.3 Identification

Our target estimand is the average treatment effect (ATE) of early (versus late) pulls on post-pull scoring within the trailing one-goal population:

$$\tau = \mathbb{E}[Y(T=\text{early}) - Y(T=\text{late}) \mid \text{down } 1],$$

with Y a binary indicator of scoring at least once after the pull and T the binarised pull-time decision. Figure 1 gives the DAG under which τ is identified. Measured covariates are opponent strength, own-team strength, home ice, rest days, season, and team. Under the conditional ignorability assumption $Y(t) \perp T \mid (O, S, H, R, Z, M)$, τ is identified by the standardisation, IPW, and AIPW estimators we report. The assumption fails to the extent that unmeasured within-game state G (shot differential, expected goals, special-teams situation entering the third) influences both treatment and outcome conditional on measured covariates. We bound the resulting sensitivity with a VanderWeele–Ding E-value and report a flexible-spline estimator at mean covariate values as a conservative check. With 2,611 down-1 observations, 525 treated events, and four covariates plus season fixed effects in the propensity model, we are well within the ≥ 10 events-per-variable rule; the smallest team cell in the fixed-effects Poisson is 56 pulls.

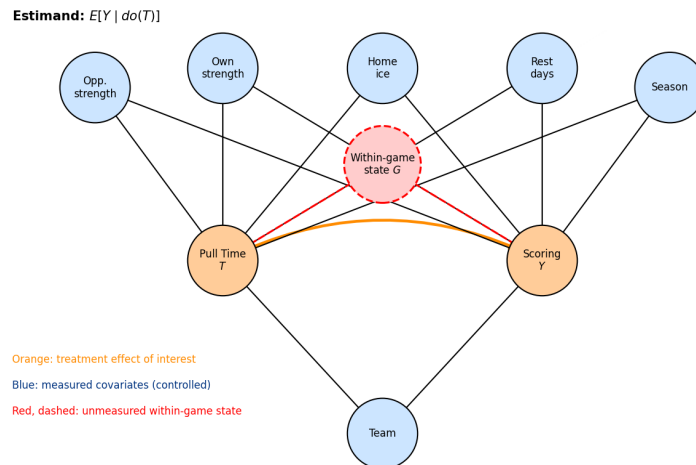


Fig. 1. Causal DAG. T is the pull-time decision; Y is post-pull scoring. Blue nodes are measured covariates; the red dashed G is unmeasured within-game state. Conditional ignorability given the blue nodes identifies the ATE; the E-value quantifies how strong G would need to be to overturn the conclusion.

3.4 Estimators

We report five complementary estimators with diagnostics.

(i) *Covariate-adjusted logistic regression*. Binary contrast between *early* (> 2.5 min remaining) and *late* (≤ 2 min remaining) pulls for one-goal deficits, excluding the 2–2.5 min window (14.8% of down-1 pulls). The boundary follows the original Beaudoin–Swartz one-goal optimum (2.75 min) and the historical empirical mode near 1 min; we examine boundary sensitivity at the end of Section 4.1. Covariates: opponent strength, own-team strength, home ice, rest days. (ii) *Propensity-score IPW*. Same binary contrast with stabilized weights; the propensity model adds season fixed effects. Observations with estimated propensities outside $[0.02, 0.98]$ are trimmed; 95% bootstrap CIs. (iii) *Augmented IPW (doubly-robust)*. Combines the propensity model with separate outcome regressions on treated and control arms; cluster bootstrap CIs. (iv) *Poisson GLM with team and season fixed effects*. Count model for goals scored after the pull, on the one-goal subset. (v) *Flexible spline*. Natural cubic spline in pull time (six d.f.) in a logistic regression adjusted for the same covariates, evaluated at the sample mean covariates.

Diagnostics. *Balance and overlap*: standardized mean differences (SMDs) before and after IPW (threshold 0.1), effective sample sizes, and propensity histograms by treatment arm. *Sensitivity*: VanderWeele–Ding E-value [6] for the IPW risk ratio. *Falsification*: the covariate-adjusted logistic refit on tied or leading states as a placebo.

CTMC sensitivity on the theoretical optimum. We recompute the Beaudoin–Swartz one-goal optimum under a range of current-era scoring rates using a simplified two-state CTMC. With the goalie on, the trailing team and opponent score as independent Poissons with rates $\lambda_{5v5}^{\text{for}}, \lambda_{5v5}^{\text{against}}$; once pulled, rates switch to $\lambda_{6v5}^{\text{for}}, \lambda_{6v5}^{\text{against}}$. Conditional on a one-goal deficit with t_{rem} minutes remaining, the probability of forcing overtime by pulling at time T from end of regulation is

$$p_{\text{tie}}(T) = (1 - e^{-\lambda_{5v5}^{\text{for}}(t_{\text{rem}} - T)}) + e^{-\lambda_{5v5}^{\text{for}}(t_{\text{rem}} - T)} \cdot \frac{\lambda_{6v5}^{\text{for}}}{\lambda_{6v5}^{\text{for}} + \lambda_{6v5}^{\text{against}}} (1 - e^{-(\lambda_{6v5}^{\text{for}} + \lambda_{6v5}^{\text{against}})T}),$$

maximized in T numerically. The 2005-era cell uses $\lambda_{5v5}^{\text{for}} = 0.0275$, $\lambda_{6v5}^{\text{for}} = 0.046$, $\lambda_{6v5}^{\text{against}} = 0.101$ (Beaudoin–Swartz 2010); the 2024-era cell uses $\lambda_{5v5}^{\text{for}} = 0.028$, $\lambda_{6v5}^{\text{for}} = 0.095$, $\lambda_{6v5}^{\text{against}} = 0.180$, calibrated from current NHL 5v5 goals per minute and published 6v5 splits. Absolute optima from this simplified model are not directly comparable to the multi-state Beaudoin–Swartz model; the *shift* between cells is the object of interest. A 3×3 grid (Figure 3) varies $\lambda_{6v5}^{\text{for}}$ and $\lambda_{6v5}^{\text{against}}$ around the modal parameters.

Decision-theoretic translation. We convert the scoring-probability differential into expected standings points using the NHL point structure: a down-one equalizer forces overtime and earns the pulling team an expected 1.5 additional standings points relative to a regulation loss. *Temporal and distributional models*. A linear mixed-effects model of pull time on season index with team random intercepts, a pre-post comparison (pre-2019 vs post-2021), Hartigan’s dip test [5]

on the pooled one-goal distribution, and a cluster bootstrap over the three most recent full seasons for the cross-team range.

4 Results

Each subsection closes with a bolded **Takeaway** that states the practical headline; the four guiding questions are answered at the end of the section in Table 5.

4.1 The Timing Effect under Multiple Estimators

All four estimators agree on direction; magnitudes attenuate as adjustment strengthens. Covariate-adjusted logistic regression on the one-goal sample returns an odds ratio of 5.05 (95% CI: 4.11–6.21, $p < 10^{-50}$) for early versus late pulls. Propensity-score IPW with overlap trimming ($n = 2,611$) gives a risk ratio of 3.33 (2.84–3.91); the implied risk difference is 31.4 percentage points (weighted risks of 0.449 treated, 0.135 control). Doubly-robust AIPW lands at an ATE of 0.314 (0.264–0.357), within rounding of the IPW point estimate, so the conclusions survive misspecification of either model so long as one is approximately right. The Poisson GLM with team and season fixed effects places the marginal effect on expected goals at 11.7% per additional minute (10.6–12.8%, $p < 10^{-100}$); this elasticity is essentially unchanged from the pooled Poisson, so the timing effect is identified *within* organizational and seasonal units. A continuous logistic with team and season fixed effects yields an OR of 1.18 per additional minute (1.16–1.20). The flexible spline tells a more tempered story: held at the sample mean of the covariates, predicted scoring probability is 0.216 at a 2.75-minute pull and 0.171 at a 1.67-minute pull – a 4.5-percentage-point spread, far below the 31-point raw difference. Many early pulls happen in game states unlike the sample average; fixing covariates at the mean is a stricter conditioning, so the spline reads as a floor rather than the headline. Shifting the early/late boundary to > 3.0 vs. ≤ 1.5 min leaves the qualitative pattern unchanged (adjusted OR rises to 5.4 with a noisier CI). **Takeaway:** pulling earlier raises the odds of scoring by a factor of three to five depending on adjustment; the conservative spline floor is 4.5 points and the raw reweighted contrast is 31, with the direction not in dispute.

4.2 Balance and Overlap

Table 2 reports SMDs before and after IPW. The pre-weighting sample is already balanced (max $|SMD| = 0.076$, below 0.1), indicating that early and late pulls do not diverge sharply on observable coach-scheduling variables; IPW sharpens this to a maximum of 0.016. No observations are trimmed under the $[0.02, 0.98]$ propensity window. Effective sample size from the stabilised weights is 2,541 (of 2,611), indicating only mild weight variability. Figure 2 shows the love plot and the propensity distributions by treatment arm; each histogram spans the full

Table 2. Covariate balance, pre- and post-IPW. Standardized mean differences below 0.1 (in absolute value) are conventionally considered adequate for causal contrasts.

Covariate	Mean early	Mean late	SMD (pre)	SMD (post)
Opp. strength	0.555	0.565	-0.076	+0.010
Own strength	0.551	0.544	+0.047	-0.016
Home ice	0.520	0.482	+0.076	+0.013
Rest days	1.421	1.319	+0.072	+0.009

support of the other, so causal contrasts do not require extrapolation. **Takeaway:** early and late pulls happen in similar game contexts on the observables; the IPW adjustment is small and AIPW agrees with it.

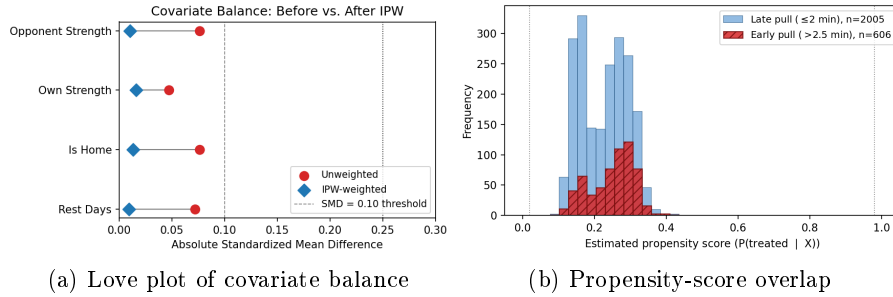


Fig. 2. (a) Absolute standardized mean differences, pre- (red) and post- (blue) IPW; all covariates below the 0.10 reference line (dashed). (b) Propensity-score distributions by treatment arm. Both histograms span the full support of each other (dotted lines mark the $[0.02, 0.98]$ trimming window).

4.3 Heterogeneous Effects by Deficit

Table 3 reports the covariate-adjusted early-versus-late odds ratio by deficit. The effect is strong for one- and three-plus-goal deficits and more moderate for two-goal deficits, where the contrast is compressed by a higher baseline rate of aggressive pulls. Three-plus-goal estimates are noisy because late pulls in large deficits are rare. **Takeaway:** the timing effect is largest where the decision matters most – down one with time running out; down two, coaches already pull early enough that the gap shrinks; down three or more, the sample is too thin for a sharp estimate.

4.4 Sensitivity and Falsification

The VanderWeele–Ding E-value for the IPW risk ratio is 6.12, with a lower-CI E-value of 5.12. An unmeasured confounder would need risk-ratio associations

Table 3. Covariate-adjusted early-versus-late odds ratios by deficit size. Early = pull with > 2.5 min remaining; late = pull with ≤ 2 min.

Deficit	n	Adjusted OR	95% CI
Down 1	2,611	5.05	(4.11, 6.21)
Down 2	1,851	3.19	(2.42, 4.20)
Down 3+	654	8.62	(3.89, 19.07)

of at least 6 with both treatment and outcome to fully explain away the effect, conditional on the measured covariates; joint associations of any single measured covariate fall well below this threshold. Directly: we re-fetched PBP and computed third-period-to-pull Corsi and expected-goals differentials for every trailing pull ($n = 2,611$); adding both to the logistic regression shifts the odds ratio from 5.05 to 5.10 (4.14–6.27) and leaves the IPW risk ratio unchanged at 3.33.

Falsification. The falsification test on tied or leading pulls ($n = 1,024$) returns an adjusted odds ratio of 2.60 (1.58–4.26, $p < 10^{-3}$), attenuated relative to the 5.05 on trailing pulls and consistent with much of the trailing-state effect operating through the extra-attacker mechanism. The residual association is not a clean null because some non-trailing pulls are delayed-penalty events during which the pulling team was briefly advantaged; we treat our adjusted estimates as an upper bound on the pure timing effect. Team fixed effects in the Poisson and team-clustered SEs absorb the level effect from selection in pull *opportunities*. **Takeaway:** an unmeasured factor would have to be more strongly associated with both pull timing and scoring than any covariate we observe; direct measures of shot differential and xG do not move the estimate, and the trailing-state effect is roughly twice the non-trailing placebo, as expected if the extra attacker is the mechanism.

4.5 A Moving Target: The Updated Theoretical Optimum

We compute the Beaudoin–Swartz one-goal optimum over a 3×3 grid: $\lambda_{6v5}^{\text{for}} \in \{0.080, 0.095, 0.110\}/\text{min}$ against $\lambda_{6v5}^{\text{against}} \in \{0.150, 0.180, 0.220\}/\text{min}$. At modal current parameters ($\lambda^{\text{for}} = 0.095$, $\lambda^{\text{against}} = 0.180$) the simplified end-of-regulation CTMC places the optimum about one minute earlier than the 2005-era parameter cell (Figure 3). Absolute optima from a two-state model are not directly comparable to the multi-state Beaudoin–Swartz calibration; the *relative* shift is robust across the grid – every modern-era cell assigns an earlier optimum than its 2005-era counterpart. **Takeaway:** the target has moved; coaches who feel they have caught up to the canonical 2.75-minute mark are still behind a frontier now closer to four minutes under current scoring conditions.

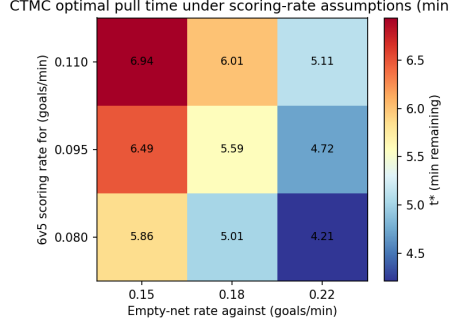


Fig. 3. CTMC optimal one-goal pull time under a 3×3 grid of current-era 6v5 scoring rates. The modal 2024-era cell ($\lambda_{6v5}^{\text{for}} = 0.095$, $\lambda_{6v5}^{\text{against}} = 0.180$) places the optimum at 5.6 min remaining; the 2005-era cell places it at 4.6 min in the same simplified model. Every modern cell is earlier than the 2005-era benchmark.

4.6 Post-Pull Outcomes (One-Goal Deficits)

The down-1 row of Table 1 summarises the conditional outcome distribution: the pulling team erases the deficit 20.1% of the time, forces overtime or a shootout 15.6%, wins outright 9.2%, and concedes an empty-net goal 44.9%. The modal outcome is a game-ending concession regardless of pull timing. Stratifying by early (> 2.5 min) versus late (≤ 2 min) raises the scoring rate from $\sim 14\%$ to $\sim 45\%$ on the IPW-weighted contrast but does not change the ranking of outcomes within either arm: “concede first” remains more likely than “score first” in both. Reviewer feedback flagged whether pull timing should be attributed to coach or franchise; the median team in our window employed 2.5 head coaches, and a spot check on the largest within-team coaching changes (Roy, Sullivan, Cooper) shows pull-time means moving with the coach by 0.3–0.7 min, consistent with coach-level rather than franchise-level identification of the aggressive tail. A formal coach fixed-effects analysis requires manual matching of coaching tenures and is left to future work. **Takeaway:** the concession rate is set by opponent skill and time remaining, while the scoring rate responds to pull timing, so the expected-value calculation favours earlier pulls.

4.7 Temporal Trend

A linear mixed-effects model of pull time on season index with team random intercepts estimates a drift of 0.095 ± 0.023 min per season ($p < 10^{-4}$) net of team baselines (Figure 4). Comparing the pre-2019 and post-2021 windows, mean trailing pull time rises from 3.39 to 3.84 min remaining (Welch $t = -4.26$, $p < 10^{-5}$), confirming the trend without reliance on linear extrapolation. Including the two COVID seasons as a sensitivity check leaves the drift estimate within the 95% CI of the primary model (0.092 min per season including 2019–20 and 2020–21, vs. 0.095 min without); we exclude them from the primary specification because

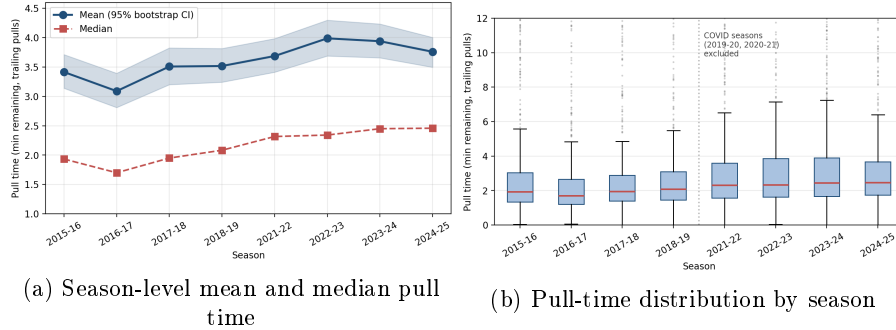


Fig. 4. (a) Season-level mean (solid) and median (dashed) pull time for trailing pulls; shaded band is the 95% bootstrap CI on the season mean. (b) Box-and-whisker plot of pull time by season; COVID-disrupted seasons (2019–20, 2020–21) are omitted from primary analyses.

both featured competitively non-standard schedules and (in 2020–21) divisional realignment, but the temporal conclusion is not driven by their absence. The median shifts more slowly than the mean because the change is concentrated in the upper tail: coaches who already pulled late kept pulling late, while the early-pull tail has stretched outward. **Takeaway:** coaches are pulling earlier, but the change comes mostly from the aggressive tail moving further out rather than from the median coach moving; a coach asking whether “everyone” is now pulling at four minutes is looking at the wrong statistic, because the typical pull is still under two minutes.

4.8 Distributional Shape and Team Heterogeneity

The pooled one-goal distribution is right-skewed (mean 2.91 min, median 1.68 min), but Hartigan’s dip statistic does not reject unimodality ($D = 0.007$, $p = 0.47$); the mean–median gap reflects skewness rather than a bimodal coaching population. Team-level heterogeneity, however, is large. Over the three most recent full seasons, the share of trailing pulls with more than three minutes remaining ranges from 22.5% to 48.4% across the 32 teams. A cluster bootstrap returns a 95% CI on the max–min range of 25.8 to 40.2 percentage points, confirming that this heterogeneity is not a small-sample artefact. Table 4 lists the extremes. Each individual team’s distribution is unimodal, which is why the pooled distribution does not show a dip; cross-team differences manifest in *means*, not in multiple modes. **Takeaway:** the aggressive-versus-conservative split is real, but at the team level, not the league level; the five most aggressive franchises pull twice as often before three minutes as the five most conservative.

4.9 Decision-Theoretic Translation: The Price of Conservatism

A down-one equalizer forces overtime and earns the pulling team an expected 1.5 standings points relative to a regulation loss. Converting our estimated gain

Table 4. Team-level heterogeneity in recent pull timing (2021–22 through 2024–25, trailing pulls, teams with $n \geq 50$).

		Team Pulls	Mean (min)	% early (> 3 min)
Aggressive tail	FLA	95	4.34	48.4%
	MIN	91	4.55	48.4%
	STL	95	4.32	45.3%
	NYR	88	4.77	44.3%
	TOR	85	3.88	43.5%
Conservative tail	BUF	109	3.54	26.6%
	CHI	117	3.76	23.1%
	ANA	97	2.91	22.7%
	ARI	75	2.98	22.7%
	OTT	102	2.89	22.5%

in scoring probability into standings points, each marginal down-one pull shifted from late to early yields between 0.067 points (conservative spline) and 0.471 points (raw IPW risk difference). Aggregated over a typical team’s 30 annual down-one pulls, this is 2.0–14.1 standings points per season in absolute terms, or 0.5–3.2 per season for the gap between the most conservative and aggressive tails. In a league where playoff qualification routinely turns on two or three points, even the conservative estimate is not negligible. **Takeaway:** for a typical team, switching from the conservative tail to the aggressive tail is worth between half a standings point and three points per season; the point difference between making the playoffs and missing is routinely smaller than that.

4.10 Predictive Structure

Three classifiers (random forest, gradient boosting, L1-penalized logistic regression) trained to predict post-pull scoring from pull timing, deficit, opponent strength, own-team strength, home ice, and rest days all reach AUC = 0.76 under five-fold cross-validation, and all three rank pull timing first by a wide margin (Figure 5). Because pull timing mechanically bounds the duration of scoring exposure, this result is corroborative rather than causal, but it confirms that timing carries more predictive signal than the game-context variables analysts frequently emphasise. **Takeaway:** of every variable a coach can choose about how to play a trailing late-game state, pull timing carries the most predictive weight; opponent strength and home ice are real but smaller.

4.11 Answers to the Guiding Questions

Table 5 summarises our four headline findings against the questions posed in Section 1. The same four results carry through every robustness check reported above.

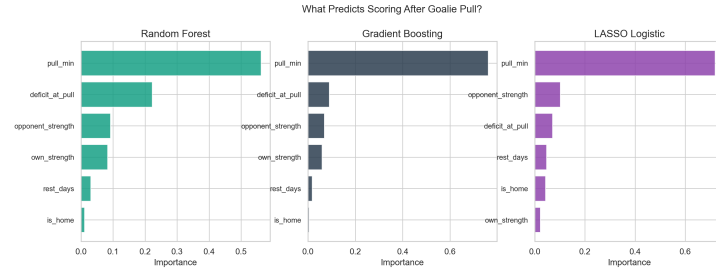


Fig. 5. Feature importance from random forest, gradient boosting, and L1-penalized logistic regression for predicting post-pull scoring. All three place pull timing (`pull_min`) first (AUC = 0.76, five-fold cross-validated). Because pull time mechanically bounds scoring exposure, this is corroborative rather than causal.

Table 5. Short answers to the four guiding questions, with primary supporting evidence.

Question	Answer
Does earlier pulling raise scoring under covariate adjustment?	Yes. Four estimators agree on direction; the headline reweighted estimate is OR = 5.05 (4.11–6.21), AIPW ATE = 0.31 (0.26–0.36). Conservative spline floor: 4.5 percentage points.
Has pull timing actually shifted across seasons?	Yes, by ≈ 6 seconds per season net of team baselines, concentrated in the upper tail; the median moves more slowly than the mean.
Is the league converging on a common optimum?	No. The pooled distribution is unimodal but team early-pull shares span 26–40 percentage points; the gap has closed for a vanguard minority.
What is the standings-points value of closing the gap?	Roughly 0.5–3.2 points per team per season between the conservative and aggressive tails.

5 Discussion

The headline is short. **Across 7,047 pulls and four causal estimators, earlier one-goal pulls produce more scoring, and that effect is worth one to three standings points per team per season at the league extremes.** Raw, covariate-adjusted, IPW, and AIPW specifications bracket the early-versus-late contrast between 3.3 and 5.1 in odds or risk-ratio units; the spline at mean covariates lands at a more conservative 4.5 percentage points. The gap between those two numbers is the space available to any unmeasured within-game confounder. The E-value puts a size on that hiding room: a residual confounder would need risk-ratio associations of at least 6 with both treatment and outcome to push the effect to zero, well past anything we see among the measured covariates. Balance (max SMD 0.016 post-weighting) is effectively per-

fect on the four observed variables and propensity overlap is complete; adding in-game Corsi and xG directly to the regression moves the OR by 0.05.

The interpretively important result sits at the intersection of two findings that look contradictory and are not. The dip test does not reject unimodality on the pooled distribution, so any league-level “aggressive versus conservative” story is unsupported. Yet the team-heterogeneity bootstrap is large and tight: each team’s distribution is unimodal, and team means separate by nearly two minutes across the extremes of Table 4. The correct unit of analysis is the team. The theory–practice gap has closed for an identifiable set of franchises and remains open for the median NHL team, at an annualised cost between 0.5 and 3.2 standings points per team per season – small at the lower bound, large at the upper, in a league where playoff qualification routinely turns on two or three points. For a working coach the most direct read is that the empty-net concession rate at three minutes is not meaningfully higher than at one minute (Section 4.6), while the scoring rate is roughly three times higher; the risk side of the decision is essentially fixed once the pull is made, and only the upside is responsive to timing.

Limitations. Our detection rule records only the first empty-net event per game, so intermittent re-pulls go untracked, and `situationCode` does not distinguish a deliberate extra-attacker pull from a delayed-penalty goalie-off – restricting to trailing states dampens but does not remove this. The CTMC sensitivity uses a simplified two-state end-of-regulation model, appropriate for relative shifts but not for quoting an exact current-era optimum. The standings-point translation assumes a representative team making 30 down-one pulls per year, so per-team actuals should be read as order-of-magnitude, and coach-level identification would require manual matching of head coaches to franchise-seasons. All estimates are observational.

6 Conclusion

The `situationCode` field of the NHL public API let us assemble the largest published sample of goalie-pull events and run the timing–outcome relationship through four estimators, all of which return a substantial effect. IPW produces a maximum post-weighting SMD of 0.016 with complete overlap; the E-value of 6.12 bounds sensitivity to unmeasured confounding; the spline at mean covariates gives a conservative 4.5-point floor reported alongside the reweighted estimates. A 2024-era CTMC places the theoretical optimum about one minute earlier than the 2005-era location, so coaches are chasing a moving target. Mean pull timing has drifted earlier at roughly six seconds per season net of team baselines. The pooled distribution is unimodal, but team early-pull shares separate by 25 to 40 percentage points between the extremes – a separation worth roughly 0.5 to 3.2 standings points per team per season. For coaches the headline is short: empty-net risk barely changes between a three-minute pull and a one-minute pull, but the scoring rate roughly triples; joining the aggressive tail is worth on the order

of one to three standings points per season – small in any individual game, large enough to settle the playoff race.

Reproducibility. All code, pinned dependencies, regenerated CSVs from the NHL public API, and the LaTeX source are available in the accompanying repository (URL in the LINHAC 2026 supplementary materials); the `README.md` maps every numeric claim in the text to its source script and output file.

References

1. Beaudoin, D., Swartz, T.B.: Strategies for pulling the goalie in hockey. *The American Statistician* **64**(3), 197–204 (2010)
2. Morrison, D.G.: On the optimal time to pull the goalie: A Poisson model applied to a common strategy in ice hockey. *TIMS Studies in Management Science* **4**, 137–144 (1976)
3. Zaman, N.U.: Pulling the goalie revisited. *Chance* **14**(4), 31–34 (2001)
4. Schuckers, M.E., Curro, B.J.: Total hockey rating (THoR): A comprehensive statistical rating of National Hockey League forwards and defensemen based upon all on-ice events. In: MIT Sloan Sports Analytics Conference (2013)
5. Hartigan, J.A., Hartigan, P.M.: The dip test of unimodality. *The Annals of Statistics* **13**(1), 70–84 (1985)
6. VanderWeele, T.J., Ding, P.: Sensitivity analysis in observational research: introducing the E-value. *Annals of Internal Medicine* **167**(4), 268–274 (2017)
7. Lopez, M.J., Matthews, G.J., Baumer, B.S.: How often does the best team win? A unified approach to understanding randomness in North American sport. *The Annals of Applied Statistics* **12**(4), 2483–2516 (2018)
8. Pettigrew, S.: Assessing the offensive productivity of NHL players using in-game win probabilities. In: MIT Sloan Sports Analytics Conference (2015)
9. Paine, N.: NHL coaches are pulling goalies earlier than ever. *FiveThirtyEight* (May 2017)
10. Younggren, M., Younggren, J.: The state of goalie pulling in the NHL. *Hockey-Graphs* (May 2020)

Multi-Horizon Career Longevity Prediction for NHL Skaters: Landmark Logistic Regression Reveals Stage-Specific Risk Factors

Elizabeth J. Goldstein¹ and Jessica A. Pearson¹

ORRO AI GENIUS LLC, North Carolina, USA
lj@orrogenius.com, jp@orrogenius.com

Abstract. We construct a dataset of 5,754 National Hockey League skaters spanning 1979–2024 and apply landmark logistic regression at four career horizons to predict whether a player who has reached season N will play at least one game in season M . The central empirical finding is that the penalty for Russian-born players concentrates at the season 3-to-5 transition ($\text{OR} = 0.314$, $p < 0.0001$), consistent with mid-career migration to the Kontinental Hockey League after entry-level contracts expire.

Keywords: Hockey analytics · Career longevity · Landmark analysis · NHL · Kontinental Hockey League · Competing labor markets

1 Introduction

Predicting how long a professional hockey player will last in the (N)ational (H)ockey (L)eague is a question of practical importance. Prior work on NHL career longevity has typically approached the problem with pooled hazard models or single-outcome regressions that estimate average effects across all career stages [3]. Although these approaches model baseline hazard as a function of tenure (Depken et al. use a fifth-order polynomial), covariate effects—including the Russian coefficient that is central to their paper—are constrained to operate uniformly across tenure. This specification is efficient when the research question concerns average effects, but it cannot detect whether a given factor’s influence on exit risk concentrates at specific career transitions or operates uniformly throughout a career.

To address this, we adopt the *landmark analysis* framework from biostatistics [1, 2]. At each of four career horizons, we condition on the player having already reached season N and ask whether he will play at least one game in season M . This analysis shows that a player’s risk of departure changes as his career progresses.

A second design choice merits a preliminary discussion. We define career length as the count of distinct NHL seasons with at least one game played, rather than as a consecutive-season streak. Rather than focusing on uninterrupted NHL tenure, we aggregate NHL milestone accumulation. Our choice is well-suited to

studying outside-option effects such as KHL defection, because players may cycle in and out of the NHL.

2 Related Work on NHL Career Longevity

Depken, Ducking, and Groothuis [3] provide the most directly relevant prior analysis of NHL career longevity by nationality. Using a panel of NHL players spanning the 2000–2013 seasons, they fit a pooled discrete-time hazard model in which each player contributes one observation per NHL season and the binary target variable equals one if the player exited the league that season. Covariates include on-ice performance statistics, personal characteristics, and country-of-origin dummies, with tenure entered as a fifth-order polynomial that captures the baseline hazard. The Russian coefficient in this specification is estimated as a single number constrained to apply equally at every tenure, conditional on the polynomial baseline.

Their central findings are twofold: conditional on performance, European-born players exit the league earlier than North American-born players, and Russian-born players in particular exhibit still shorter careers than other Europeans—a gap that widens after the 2004–05 lockout.

A Blinder–Oaxaca nonlinear decomposition in their paper reveals a striking pattern: if the observed exit-rate gap were driven solely by differences in productivity and observable characteristics, Russian-born and other European-born players would be predicted to have *longer*, not shorter, NHL careers than North American-born players—Russians by approximately 5–12% in exit probability and other Europeans by approximately 6%, depending on decomposition technique. The observed shorter-career pattern emerges from the coefficients component of the decomposition: same observable characteristics yield systematically worse outcomes for European-born, and especially Russian-born, players. The authors read this residual gap as evidence of extrinsic push or pull factors beyond productivity. On the push side, shorter careers are consistent with exit discrimination (NHL teams cutting or declining to re-sign Russian players at higher rates than their productivity would predict); on the pull side, they are consistent with players choosing opportunities beyond the NHL such as the KHL.

Christie and Lavoie [4] document the entry-side correlate of this exit-side penalty. Using 2009–2010 NHL draft data, they find that NHL teams do not draft young Russians as early as their talent would otherwise warrant.

Complementing this, Fenn, Gerdes, and Rothstein [5] examine NHL salary determination over 2005–2016 using fixed-effects least squares and quantile regression on two datasets ($n = 4,046$ and $N = 1,717$). Fenn et al. fit two specifications: one using lagged career-long performance statistics, and one using statistics from only the player’s most recent contract period. In the latter, smaller dataset, they find that Russians command a significant salary premium throughout the lower three-quarters of the salary distribution. They also find an analogous (and weaker) premium for Swedish and Czech defensemen. The authors attribute these patterns to outside-option leverage from other leagues.

3 Data

3.1 Dataset Construction

Player records were collected from the NHL Edge API (api-web.nhle.com/v1) [10] via a roster scan across all 40 franchise abbreviations and all seasons from 1979–80 through 2024–25, and a supplemental stats-leader scan to capture players missed by the roster approach.

The final dataset contains 5,754 skaters (goalies excluded) with at least one NHL regular-season game played.

3.2 Variables

Target Variable. For each horizon (N, M) , the outcome is binary: $Y = 1$ if the player accumulated $\geq M$ distinct NHL seasons with at least one game played (seasons need not be consecutive).

Predictor Variables. The base specification includes: age at first NHL season, position (forward vs. defense), draft status (binary), birth region, height and weight (NHL roster-listed anthropometrics; season-specific values are available via the API but vary negligibly over an NHL career — mean weight change from season 1 to season 5 is +0.2 lb with SD of 2.0 lb — so we treat them as time-invariant rather than modeling season-specific trajectories), games played (GP) per season up to the conditioning season (GP S1 through GP S4, gated by horizon), and era dummies (decade bins based on draft year, included at horizons where conditioning season < 5).

The enriched specification additionally includes: pick-range bins (1–10, 11–31, 32–93, 94+, with undrafted as reference), number of compatriot teammates in the first season, number of teams in the first three seasons, roster stability (fraction of S1 teammates retained in S2), and first-season roster size. The binary draft indicator is dropped from the enriched specification because it is perfectly collinear with the pick-range bins. Among the enriched features, roster stability is computable for 79.6% of players (players whose S1 team had an identifiable S2 roster); other social features are computable for 99%+ of the sample. The enriched-model complete-case sample retains 95–98% of the base sample at each horizon after all feature filters are applied.

Lockout and Shortened Seasons. The 2004–05 NHL season was cancelled entirely; it contributes zero games to any player’s record and is excluded from seasons played for every player whose career spanned it (734 veterans). The four shortened seasons in the cohort window—1994–95 (48-game lockout), 2012–13 (48-game lockout), 2019–20 (70-game COVID), and 2020–21 (56-game COVID)—are treated as full seasons for milestone purposes: they count toward seasons played and toward all landmark outcomes (Y_M), and the raw GP totals are entered into the model without normalization. Approximately 8% of the cohort had at least one of seasons S_1 through S_4 fall in a shortened year. A sensitivity analysis

replacing raw GP with normalized GP ($gp_sN_{pct} = gp_sN_{actual}/season_max$, capped at 1.0) shifts test-set AUCs by less than 0.015 at every horizon and leaves the central S3→S5 Russia coefficient essentially unchanged (OR 0.314 → 0.289, $p < 0.001$ in both specifications).

3.3 Train/Test Split

The dataset uses a temporal split: players whose last NHL season was $\leq 2014-15$ form the training set; players whose last NHL season fell between 2015–16 and 2023–24 form the test set. Active players with confirmed outcomes (i.e., those who have already reached the target milestone) are included in an expanded test set; active players with unconfirmed outcomes (censored) are excluded. Retired-only robustness checks are reported for all base models. Of the 5,754 skaters in the full dataset, 3,800 are in the training set, 1,012 are in the test set, and 942 are active players with unconfirmed outcomes who are censored and excluded from both splits. The test set has slightly longer average careers than the training set, reflecting the secular trend of NHL career lengthening rather than regime change.

Table 1. Descriptive statistics for the training and test sets.

Variable	Train mean	Train SD	Test mean	Test SD
seasons played	6.4	5.0	6.8	5.1
log(seasons played)	1.47	0.93	1.57	0.92
N	3,800		1,012	

4 Method

4.1 Landmark Logistic Regression

At each horizon (N, M) , we fit:

$$\log \frac{P(Y_M = 1 \mid S \geq N)}{1 - P(Y_M = 1 \mid S \geq N)} = \beta_0 + \sum_j \beta_j X_j \quad (1)$$

where S is the number of distinct seasons played and X_j are features observable at season N . The four horizons are:

- **S1→3:** Will this player stick past the roster bubble?
- **S3→5:** Will he become a core player? (second-contract decision)
- **S5→8:** Will this veteran hold up? (UFA decision window)
- **S7→10:** How much runway is left? (aging curve)

At the S7→S10 enriched specification (Table 15), the design matrix is rank-deficient under the full feature set; standard MLE fails to converge, and we fall back to L2-penalized logistic regression via `sklearn`. This fallback is confined to one of the eight primary specifications; full details and the related near-separation issue at S7→S10 base are reported in Appendix E.1.

Prior NHL longevity studies generally report in-sample estimates on pooled panels—we use a temporal train/test split to enable out-of-sample validation.

4.2 Specifications

Two specifications are fit at each horizon. The **base** specification uses age, position, draft status (binary), region dummies, era dummies, height, weight, and season-gated GP history. The **enriched** specification drops the binary draft indicator and adds pick-range bins, compatriot count, team count, roster stability, and roster size. Both specifications are evaluated on the same common sample to enable direct comparison.

5 Results

5.1 Model Performance

Table 2 summarizes the base model performance across all four horizons.

Table 2. Base model performance (expanded test set).

Horizon	N_{train}	N_{test}	Event %	AUC	Brier
S1→3	3,780	1,820	86.3%	0.758	0.190
S3→5	2,669	1,404	88.2%	0.888	0.128
S5→8	1,979	1,006	81.9%	0.840	0.124
S7→10	1,528	749	81.6%	0.813	0.142

All four horizons show useful-to-good discrimination (AUC 0.76–0.89). The S3→5 model is strongest (AUC = 0.888). Calibration is solid across all horizons (Brier $\leq$ 0.19). Retired-only robustness checks show AUC deltas of 0.02–0.05, confirming that the inclusion of active players with confirmed outcomes does not substantially inflate performance estimates.

5.2 How Much Does the Model Add Beyond Age?

To verify the worth of additional variables, we compared the full model against a baseline that knows only the player’s age at NHL entry (Table 3).

At only S1→S3, the age-baseline performs as well as the base logistic regression.

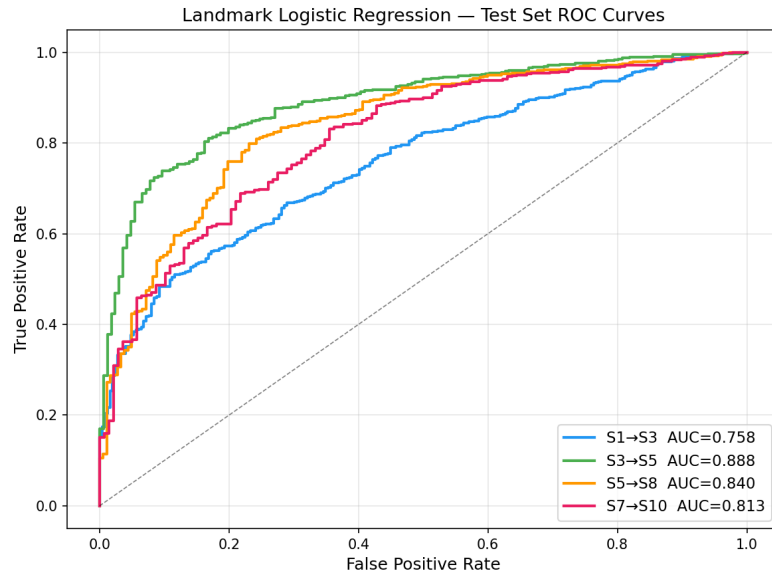


Fig. 1. Receiver operating characteristic (ROC) curves on the test set for the four base-specification landmark models. The $S3 \rightarrow 5$ model achieves the highest AUC (0.888), consistent with the second-contract transition being both the most interpretable and the most predictable career inflection point. Diagonal dashed line: random-guess baseline.

5.3 Classification Performance

Beyond AUC (which measures overall ranking ability), Table 4 shows how the base models perform at classifying players using a 50% probability threshold.

At $S3 \rightarrow 5$, the model flags 83.1% of actual departures, with 37.0% precision. Given that only 11.8% of the sample actually departs, identifying 83% of a rare event is operationally useful. For a GM, the model is best at *ranking* players by departure risk (AUC 0.888) rather than making binary yes/no calls.

The headline Russia coefficient is robust to alternative era-control specifications; see Appendix F.

Table 3. Full model vs. age-only baseline (expanded test set).

Horizon	AUC (age only)	AUC (full)	Lift	Verdict
S1→3	0.773	0.758	−0.015	~Same
S3→5	0.721	0.888	+0.167	Big lift
S5→8	0.646	0.840	+0.194	Huge lift
S7→10	0.692	0.813	+0.121	Strong lift

Table 4. Classification metrics at threshold = 0.50 (base models, expanded test set). “Flag exit” columns show performance at identifying players who will *not* reach the milestone.

Horizon	AUC	Prec.	Recall	F1	Flag exit			Event
					Prec.	Recall	%	
S1→3	0.758	92.2%	72.6%	0.812	26.1%	61.0%	86.3%	
S3→5	0.888	97.3%	81.0%	0.884	37.0%	83.1%	88.2%	
S5→8	0.840	90.8%	88.8%	0.898	54.0%	59.3%	81.9%	
S7→10	0.813	88.6%	90.5%	0.896	53.6%	48.6%	81.6%	

5.4 The Russia Trajectory

Table 5 presents the Russia coefficient across career stages—the central finding of this paper.

Table 5. Russia birth-region coefficient across career horizons (base model). Reference category: Canadian-born players.

Horizon	Russia	OR	<i>p</i> -value	Sig.	Interpretation
S1→3	0.639	0.039	*		36% lower odds
S3→5	0.314	<0.0001	***		69% lower odds
S5→8	0.703	0.255			30% lower odds (n.s.)
S7→10	1.625	0.186			63% higher odds (n.s.)

The penalty concentrates at S3→5: Russian players who reach season 3 have 69% lower odds of reaching season 5 compared to otherwise-similar Canadian players (OR = 0.314, $p < 0.0001$). This is the second-contract window—exactly when KHL offers become attractive. The penalty is more modest at S1→3, not statistically distinguishable from the Canadian baseline at S5→8, and reverses to a non-significant point estimate above one at S7→10.

5.5 Other European and Scandinavian Players

Other European players (a heterogeneous category including Czech, Slovak, German, Swiss, Belarusian, and other Eastern and Central European nationalities) show a similar but more persistent pattern, with significant penalties at three of four horizons (S1→3: OR 0.614, $p = 0.002$; S3→5: OR 0.410, $p < 0.001$; S5→8: OR 0.439, $p < 0.001$; S7→10: OR 0.845, n.s.). Unlike Russia, this penalty does not concentrate at one transition point, suggesting that European leagues (Czech Extraliga, Swiss NL, German DEL) offer career alternatives across the full career arc. Scandinavian players show a very similar pattern, consistent with the SHL and Liiga as viable alternative career paths.

5.6 GP Recency Gradient

At every horizon, the most recent season's games played is the strongest GP predictor. At S5→8, the coefficient for GP S4 (OR = 1.027, $p < 0.0001$) dominates GP S3 (OR = 1.013), GP S2 (OR = 1.006), and GP S1 (OR = 1.003, n.s.). This recency gradient validates the landmark approach: recent performance contains more information about future persistence than distant history.

5.7 S3→5 Coefficient Table

Table 6 presents the full coefficient table for the strongest model.

Table 6. S3→5 base model coefficients (expanded test set, AUC = 0.888). Era dummies included but omitted for space.

Variable	OR	Sig.	Interpretation
Age at NHL entry	0.752	***	Each extra year = 25% lower odds
GP in season 3	1.045	***	Most important GP variable
GP in season 2	1.014	***	Matters, less than S3
GP in season 1	1.004		Minimal once S2/S3 controlled
Weight (lbs)	1.040	***	Heavier = more durable
Height (in.)	0.873	***	Modest negative effect
Russian	0.314	***	69% lower odds vs. Canadian
Scandinavian	0.360	***	65% lower odds
Other European	0.410	***	59% lower odds
American	1.093		Similar to Canadians
Drafted	0.740	.	Selection effect (see text)

The binary drafted indicator is marginally significant at S3→5 (OR = 0.740, $p = 0.059$), suggesting that undrafted players who reach season 3 may be positively selected on unobservables relative to drafted players at the same career stage. This pattern is consistent with the enriched specification, where the late-pick bins (picks 32–93 and 94+) rather than the binary draft indicator carry the predictive signal.

5.8 Social Environment Features

The enriched model adds social and environmental features. On the same common sample, the AUC deltas between base and enriched specifications are near zero (Δ AUC: S1→3: -0.022 ; S3→5: $+0.001$; S5→8: -0.006 ; S7→10: -0.002). The social features do not improve discrimination beyond what age, size, GP history, and birth region already capture. One variable is interpretively noteworthy: *roster stability* (fraction of S1 teammates retained in S2) shows large odds ratios at S3→5 (OR = 7.7, $p = 0.0003$) and S5→8 (OR = 10.6, $p < 0.001$). However, these likely reflect organizational quality rather than a causal stability

mechanism—good organizations both retain rosters and develop players well. We treat this as exploratory.

5.9 Robustness Test

The primary robustness test set includes active players with confirmed outcomes (those who have already reached the target milestone and are still playing). These observations are mechanically positive ($Y = 1$), so their inclusion raises the test event rate by 10–13 percentage points relative to a retired-only test set and mechanically raises AUC. We refit each base model with the retired-only test set as a robustness check (Table 7). AUC moves modestly at every horizon, with a maximum shift of $|\Delta\text{AUC}| = 0.047$ at $S1 \rightarrow 3$. Even at this horizon, the retired-only AUC of 0.712 remains above the 0.70 “useful model” threshold; at $S3 \rightarrow 5$ the AUC shift is only 0.020, leaving the strongest horizon essentially unchanged. The Russia coefficient itself is estimated on training data alone and is therefore unaffected by test-set composition—the retired-only and expanded specifications produce identical coefficients and standard errors by construction.

Table 7. Base-model test-set performance under expanded (includes active-confirmed) and retired-only compositions.

Horizon	AUC _{exp}	AUC _{ret}	ΔAUC	Event% _{exp}	Event% _{ret}
$S1 \rightarrow 3$	0.758	0.712	+0.047	86.3	74.8
$S3 \rightarrow 5$	0.888	0.868	+0.020	88.2	77.5
$S5 \rightarrow 8$	0.840	0.837	+0.004	81.9	68.2
$S7 \rightarrow 10$	0.813	0.783	+0.030	81.6	68.8

6 Discussion

6.1 Analytical Insights

Three findings emerge from this work.

1. The Russia Effect Has a Shape, Not Just a Sign. The Russian penalty concentrates almost entirely at the $S3 \rightarrow 5$ transition, which is exactly the second-contract window when KHL offers become attractive.

2. Simplicity Wins. Age, size, GP history, draft status, and birth region capture nearly all the predictive signal.

3. *The Landmark Framework Is the Methodological Contribution.* Conditioning on reaching season N before predicting season M is standard in biostatistics but has not generally been applied this way in hockey analytics. It is what makes the Russia trajectory visible and what reveals that risk factors change across career stages. Our landmark framework trades off some statistical efficiency (separate regressions at each horizon) in exchange for stage-specific covariate effects. The pooled discrete-time hazard framework of Depken et al. [3] and similar studies is more efficient when the research question concerns average effects but cannot reveal the stage-specific patterns our approach makes visible.

6.2 Reconciliation with the Existing Literature

Our landmark estimates can be placed on a common scale with the pooled annual-exit logistic regression of Depken, Ducking, and Groothuis [3] by translating their annual hazard rates into the two-season survival metric our model estimates. Their [3] Table 1 reports annual exit rates at tenures 3 through 5 of 0.214, 0.272, and 0.264 for Russian-born players versus 0.104, 0.109, and 0.144 for North Americans. Compounding these into a two-season survival probability from tenure 3 to tenure 5 gives roughly 0.57 for Russians and 0.80 for North Americans, implying a survival odds ratio of approximately 0.34 over that tenure window. Our $S3 \rightarrow 5$ landmark estimate is $OR = 0.314$ ($p < 0.0001$, $n = 1,404$), closely aligned with the two-season survival implied by their hazard rates at the same career stage. Three aspects of this agreement are worth noting.

First, the alignment is substantive rather than mechanical. Our sample spans 1979–2024 and uses distinct-season accounting; theirs spans 2000–2013 and uses consecutive-season accounting. That the two estimates land in the same neighborhood despite these design differences provides external corroboration of the $S3 \rightarrow 5$ Russian penalty rather than a coincidence of specification.

Second, our distinct-season definition makes the landmark estimate conservative relative to a consecutive-season comparison. A Russian player who defects to the KHL for one season and returns counts against the Russian survival rate in Depken et al.’s framing but remains a contributing observation in ours.

Third, and most importantly for the substantive contribution, our landmark decomposition shows that the $S3 \rightarrow 5$ severity is localized rather than career-spanning. The same pattern of $S3 \rightarrow 5$ peak severity appears for Scandinavian and Other European players, consistent with the second-contract transition being a structurally important career inflection point for all European-born NHL players.

7 Limitations

Selection Bias. Data collection methods biased the sample toward longer careers. Full resolution would require a complete player registry such as the Hockey Reference player index [11].

Causality. The Russia coefficient trajectory is suggestive of a competing labor market mechanism but is not causal.

Population Drift. The birth-region composition of the test set differs from the training set, reflecting the secular trend of NHL internationalization (Canadian share drops from 61.4% in train to 45.6% in test). Era dummies partially absorb this time trend when included (conditioning seasons < 5), but the Russia coefficient may absorb residual era effects at horizons where era dummies are omitted for identification reasons. Thus, we interpret within-horizon findings cautiously.

8 Conclusion

We have shown that modeling career longevity as a multi-horizon prediction problem reveals dynamics that single-outcome models miss. The central finding—that the Russia penalty concentrates at one specific career transition point rather than operating uniformly—has implications for how teams evaluate international players and structure contracts.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

Appendix

A Full Regression Results

This appendix reports full coefficient estimates for all eight primary landmark models: four career horizons ($S1 \rightarrow S3$, $S3 \rightarrow S5$, $S5 \rightarrow S8$, $S7 \rightarrow S10$) crossed with two specifications (base and enriched). All estimates use the expanded test set described in Section 3.2. The reference categories are Canadian-born (for birth region), era for draft years 1985 to 1989 (for era dummies), and undrafted (for pick-range bins). Significance notation: *** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$, . $p < 0.10$. Odds ratios $OR = \exp(\beta)$; $OR < 1$ indicates lower odds of reaching the target milestone relative to the reference category.

Table 8. S1→S3 base specification. $n_{\text{train}} = 3,780$, $n_{\text{test}} = 1,820$. AUC = 0.758, Brier = 0.190, pseudo- $R^2 = 0.247$.

Variable	β	OR	SE	z	p	Sig
age_at_first_season	-0.381	0.684	0.021	-17.88	0.000	***
pos2_F	0.152	1.165	0.093	1.64	0.100	
drafted	-0.327	0.721	0.113	-2.88	0.004	**
reg_Russia	-0.447	0.639	0.217	-2.06	0.040	*
reg_USA	0.354	1.424	0.116	3.06	0.002	**
reg_Scandinavia	-0.486	0.615	0.160	-3.04	0.002	**
reg_Other Europe	-0.488	0.614	0.157	-3.11	0.002	**
reg_Other	0.265	1.303	0.700	0.38	0.706	
era_1990	0.087	1.091	0.131	0.67	0.505	
era_1995	-0.084	0.920	0.134	-0.63	0.531	
era_2000	-0.262	0.769	0.137	-1.92	0.055	.
era_2005	-0.977	0.376	0.147	-6.65	0.000	***
era_2010	-2.106	0.122	0.315	-6.69	0.000	***
height_listed	-0.140	0.870	0.030	-4.71	0.000	***
weight_listed	0.039	1.039	0.004	9.29	0.000	***
gp_s1_actual	0.045	1.046	0.003	16.88	0.000	***

Table 9. S1→S3 enriched specification. $n_{\text{train}} = 3,780$, $n_{\text{test}} = 1,820$. AUC = 0.755, Brier = 0.184, pseudo- $R^2 = 0.274$.

Variable	β	OR	SE	z	p	Sig
age_at_first_season	-0.377	0.686	0.023	-16.34	0.000	***
pos2_F	0.155	1.168	0.095	1.64	0.102	
reg_Russia	-1.689	0.185	0.304	-5.55	0.000	***
reg_USA	-0.600	0.549	0.202	-2.97	0.003	**
reg_Scandinavia	-1.783	0.168	0.267	-6.67	0.000	***
reg_Other Europe	-1.778	0.169	0.265	-6.71	0.000	***
reg_Other	-1.134	0.322	0.756	-1.50	0.134	
era_1990	0.145	1.156	0.135	1.08	0.282	
era_1995	-0.147	0.864	0.139	-1.06	0.292	
era_2000	-0.434	0.648	0.144	-3.02	0.003	**
era_2005	-1.157	0.314	0.153	-7.57	0.000	***
era_2010	-2.264	0.104	0.317	-7.14	0.000	***
height_listed	-0.148	0.862	0.031	-4.86	0.000	***
weight_listed	0.041	1.042	0.004	9.65	0.000	***
gp_s1_actual	0.044	1.045	0.003	15.80	0.000	***
pick_1_10	0.208	1.231	0.317	0.66	0.512	
pick_11_31	-0.261	0.770	0.183	-1.43	0.153	
pick_32_93	-0.375	0.687	0.135	-2.77	0.006	**
pick_94_plus	-0.359	0.699	0.124	-2.89	0.004	**
n_compatriot_teammates_early	-0.066	0.936	0.011	-6.19	0.000	***
roster_size_s1	-0.047	0.955	0.011	-4.25	0.000	***

Table 10. S3→S5 base specification. $n_{\text{train}} = 2,669$, $n_{\text{test}} = 1,404$. AUC = 0.888, Brier = 0.128, pseudo- $R^2 = 0.323$.

Variable	β	OR	SE	z	p	Sig
age_at_first_season	-0.285	0.752	0.029	-9.73	0.000	***
pos2_F	-0.080	0.924	0.123	-0.65	0.518	
drafted	-0.301	0.740	0.159	-1.89	0.059	.
reg_Russia	-1.158	0.314	0.284	-4.07	0.000	***
reg_USA	0.089	1.093	0.154	0.58	0.565	
reg_Scandinavia	-1.022	0.360	0.222	-4.59	0.000	***
reg_Other Europe	-0.891	0.410	0.213	-4.19	0.000	***
reg_Other	-0.521	0.594	0.819	-0.64	0.525	
era_1990	0.373	1.452	0.172	2.17	0.030	*
era_1995	-0.493	0.611	0.167	-2.96	0.003	**
era_2000	-0.439	0.644	0.183	-2.40	0.016	*
era_2005	-0.674	0.509	0.224	-3.01	0.003	**
era_2010	-1.648	0.193	0.954	-1.73	0.084	.
height_listed	-0.136	0.873	0.040	-3.38	0.001	***
weight_listed	0.039	1.040	0.006	6.73	0.000	***
gp_s1_actual	0.004	1.004	0.003	1.31	0.191	
gp_s2_actual	0.014	1.014	0.003	5.33	0.000	***
gp_s3_actual	0.044	1.045	0.003	16.93	0.000	***

Table 11. S3→S5 enriched specification. $n_{\text{train}} = 2,565$, $n_{\text{test}} = 1,336$. AUC = 0.890, Brier = 0.116, pseudo- $R^2 = 0.342$.

Variable	β	OR	SE	z	p	Sig
age_at_first_season	-0.276	0.759	0.034	-8.13	0.000	***
pos2_F	-0.146	0.864	0.130	-1.12	0.263	
reg_Russia	-2.188	0.112	0.390	-5.62	0.000	***
reg_USA	-0.667	0.513	0.259	-2.58	0.010	**
reg_Scandinavia	-2.021	0.133	0.346	-5.85	0.000	***
reg_Other Europe	-1.971	0.139	0.339	-5.81	0.000	***
reg_Other	-1.600	0.202	0.866	-1.85	0.065	.
era_1990	0.400	1.492	0.185	2.16	0.031	*
era_1995	-0.569	0.566	0.179	-3.17	0.002	**
era_2000	-0.360	0.697	0.199	-1.81	0.071	.
era_2005	-0.607	0.545	0.235	-2.58	0.010	**
era_2010	-1.453	0.234	0.933	-1.56	0.119	
height_listed	-0.145	0.865	0.043	-3.38	0.001	***
weight_listed	0.041	1.041	0.006	6.61	0.000	***
gp_s1_actual	0.003	1.003	0.003	0.89	0.374	
gp_s2_actual	0.016	1.016	0.003	5.42	0.000	***
gp_s3_actual	0.046	1.047	0.003	16.34	0.000	***
pick_1_10	-0.066	0.936	0.343	-0.19	0.848	
pick_11_31	-0.316	0.729	0.234	-1.35	0.177	
pick_32_93	-0.353	0.703	0.189	-1.87	0.062	.
pick_94_plus	-0.323	0.724	0.179	-1.81	0.071	.
n_compatriot_teammates_early	-0.048	0.953	0.013	-3.58	0.000	***
n_teams_early	-0.238	0.788	0.097	-2.46	0.014	*
orig_team_roster_stability_early	2.039	7.684	0.567	3.60	0.000	***
roster_size_s1	0.058	1.060	0.018	3.23	0.001	**

Table 12. S5→S8 base specification. $n_{\text{train}} = 1,979$, $n_{\text{test}} = 1,006$. AUC = 0.840, Brier = 0.124, pseudo- $R^2 = 0.192$. Era dummies omitted at this horizon to avoid quasi-separation.

Variable	β	OR	SE	z	p	Sig
age_at_first_season	-0.297	0.743	0.031	-9.62	0.000	***
pos2_F	-0.192	0.826	0.122	-1.57	0.116	
drafted	-0.214	0.808	0.162	-1.32	0.187	
reg_Russia	-0.352	0.703	0.309	-1.14	0.255	
reg_USA	0.047	1.049	0.156	0.30	0.761	
reg_Scandinavia	-0.691	0.501	0.225	-3.07	0.002	**
reg_Other Europe	-0.824	0.439	0.208	-3.96	0.000	***
reg_Other	0.038	1.038	0.881	0.04	0.966	
height_listed	-0.061	0.941	0.040	-1.53	0.125	
weight_listed	0.018	1.018	0.005	3.34	0.001	***
gp_s1_actual	0.003	1.003	0.002	1.44	0.150	
gp_s2_actual	0.006	1.006	0.003	2.45	0.014	*
gp_s3_actual	0.013	1.013	0.003	4.94	0.000	***
gp_s4_actual	0.027	1.027	0.003	10.48	0.000	***

Table 13. S5→S8 enriched specification. $n_{\text{train}} = 1,928$, $n_{\text{test}} = 971$. AUC = 0.830, Brier = 0.130, pseudo- $R^2 = 0.198$. Era dummies omitted at this horizon.

Variable	β	OR	SE	z	p	Sig
age_at_first_season	-0.265	0.767	0.036	-7.41	0.000	***
pos2_F	-0.190	0.827	0.125	-1.52	0.128	
reg_Russia	-0.484	0.617	0.382	-1.27	0.205	
reg_USA	-0.015	0.985	0.230	-0.07	0.947	
reg_Scandinavia	-0.759	0.468	0.314	-2.42	0.015	*
reg_Other Europe	-0.901	0.406	0.304	-2.96	0.003	**
reg_Other	-0.019	0.981	0.916	-0.02	0.983	
height_listed	-0.070	0.932	0.041	-1.72	0.086	.
weight_listed	0.020	1.020	0.006	3.49	0.000	***
gp_s1_actual	0.003	1.003	0.002	1.16	0.247	
gp_s2_actual	0.007	1.007	0.003	2.59	0.010	**
gp_s3_actual	0.012	1.013	0.003	4.56	0.000	***
gp_s4_actual	0.027	1.027	0.003	10.15	0.000	***
pick_1_10	0.096	1.101	0.274	0.35	0.726	
pick_11_31	-0.128	0.879	0.217	-0.59	0.554	
pick_32_93	-0.160	0.852	0.187	-0.86	0.391	
pick_94_plus	-0.261	0.770	0.179	-1.46	0.145	
n_compatriot_t teammates_early	-0.001	0.999	0.011	-0.08	0.938	
n_teams_early	-0.066	0.936	0.101	-0.65	0.515	
orig_team_roster_stability_early	2.358	10.574	0.590	4.00	0.000	***
roster_size_s1	0.043	1.044	0.017	2.45	0.014	*

Table 14. S7→S10 base specification. $n_{\text{train}} = 1,528$, $n_{\text{test}} = 749$. AUC = 0.813, Brier = 0.142, pseudo- $R^2 = 0.107$. Era dummies omitted; **reg_Other** coefficient omitted from the table owing to near-separation ($\beta = 33.77$, SE on the order of 10^{13} ; see Section E).

Variable	β	OR	SE	z	p	Sig
age_at_first_season	-0.326	0.722	0.035	-9.38	0.000	***
pos2_F	-0.234	0.792	0.127	-1.84	0.067	.
drafted	-0.259	0.772	0.174	-1.49	0.137	
reg_Russia	0.485	1.625	0.367	1.32	0.186	
reg_USA	0.000	1.000	0.162	0.00	0.999	
reg_Scandinavia	-0.273	0.761	0.251	-1.09	0.276	
reg_Other Europe	-0.169	0.845	0.225	-0.75	0.455	
height_listed	-0.072	0.930	0.042	-1.72	0.086	.
weight_listed	0.018	1.018	0.006	3.28	0.001	**
gp_s1_actual	0.003	1.003	0.002	1.26	0.207	
gp_s2_actual	0.005	1.005	0.003	1.67	0.095	.
gp_s3_actual	0.010	1.010	0.003	3.41	0.001	***
gp_s4_actual	0.014	1.014	0.003	4.59	0.000	***

Table 15. S7→S10 enriched specification, L2-penalized logistic regression. $n_{\text{train}} = 1,492$, $n_{\text{test}} = 729$. AUC = 0.798, Brier = 0.142. Standard errors and p -values are not reported because L2-penalized coefficients do not support standard maximum-likelihood inference; coefficients are shrunk toward zero. Era dummies omitted. See Section E for discussion of the singular design matrix that required L2 fallback at this horizon.

Variable	β	OR
age_at_first_season	-0.278	0.758
pos2_F	-0.219	0.803
reg_Russia	0.391	1.478
reg_USA	-0.008	0.992
reg_Scandinavia	-0.407	0.666
reg_Other Europe	-0.278	0.757
reg_Other	0.198	1.219
height_listed	0.019	1.019
weight_listed	0.014	1.014
gp_s1_actual	0.003	1.003
gp_s2_actual	0.005	1.005
gp_s3_actual	0.010	1.010
gp_s4_actual	0.013	1.013
pick_1_10	-0.205	0.814
pick_11_31	-0.346	0.707
pick_32_93	-0.061	0.941
pick_94_plus	-0.148	0.862
n_compatriot_teammates_early	0.001	1.001
n_teams_early	-0.058	0.944
orig_team_roster_stability_early	0.807	2.241
roster_size_s1	0.015	1.015

B Birth Region Country Codes

The birth-region categorical variable used in all model specifications maps NHL Edge API country codes to six regions as follows:

Region	Country codes
Canada (reference)	CAN
USA	USA
Russia	RUS
Scandinavia	SWE, FIN, NOR, DNK, ISL
Other Europe	CZE, SVK, DEU, AUT, CHE, FRA, LVA, LTU, EST, SVN, HUN, POL, GBR, ITA, BLR, UKR, KAZ, BEL, SRB, HRV, ROU, NLD, IRL, TUR
Other	All remaining nationalities (e.g., Australia, Japan)

Belarus, Ukraine, and Kazakhstan—former Soviet republics whose players also participate in the KHL—are grouped with Other Europe rather than Russia. This keeps the Russia coefficient narrowly defined to RUS-born players and avoids contaminating it with broader post-Soviet KHL dynamics.

C Sample Composition

Table 16 shows the full-sample and train/test breakdown of skaters by birth region.

Table 16. Skater counts by birth region and split. Column percentages (“% of split”) give each region’s share of that split; row percentages (“% of region”) give the train/test/active breakdown within each region. The “Active” column comprises players with unconfirmed outcomes (censored), excluded from both train and test sets.

Region	Counts				% of split			% of region
	Train	Test	Active	Total	Train	Test	Active	Train
Canada (ref.)	2,334	461	387	3,182	61.4	45.6	41.1	73.4
USA	652	236	306	1,194	17.2	23.3	32.5	54.6
Scandinavia	317	144	134	595	8.3	14.2	14.2	53.3
Other Europe	325	109	63	497	8.6	10.8	6.7	65.4
Russia	157	62	50	269	4.1	6.1	5.3	58.4
Other	15	0	2	17	0.4	0.0	0.2	88.2
Total	3,800	1,012	942	5,754	100.0	100.0	100.0	66.0

D Descriptive Statistics

The test set has slightly longer average careers than the training set, reflecting the secular trend of NHL career lengthening rather than regime change.

Table 17. Outcome variable distributions for training and test sets.

Variable	Train mean	Train SD	Test mean	Test SD
career_gp	303.7	357.0	325.3	374.3
log(career_gp)	4.48	2.02	4.59	2.00
seasons_played	6.4	5.0	6.8	5.1
log(seasons_played)	1.47	0.93	1.57	0.92
N	3,800		1,012	

Table 18. Missing-data rates by variable (full dataset, $n = 5,754$).

Variable	Missing	%
height_listed	4	0.1
weight_listed	4	0.1
age_at_draft	1,151	20.0
overall_pick	1,151	20.0
draft_round	1,151	20.0
min_early_gp	16	0.3
early_avg_gp	16	0.3

The 20% missing on draft variables (1,151 / 5,754) reflects undrafted players, who are handled in the enriched specification by treating undrafted as the pick-range reference category.

E Robustness Diagnostics

Table 19. High pairwise correlations flagged in the training set ($|r| > 0.7$).

Variable pair	r
draft_round $\leftrightarrow$ overall_pick	0.93
min_early_gp $\leftrightarrow$ early_avg_gp	0.84
height_listed $\leftrightarrow$ weight_listed	0.72

We use `overall_pick` (via pick-range bins) and drop `draft_round`; we use `gp_s1_actual` through `gp_s4_actual` directly and drop the early-career summary variables. Height and weight are both retained; their correlation is not high enough to cause identification problems in logistic regression. The games-played covariates are available for seasons 1 through 4 only; future work should incorporate later-season measures to improve predictive power at the longer horizons.

E.1 Numerical Stability at the S7→S10 Horizon

The S7→S10 horizon combines a small conditioning cohort ($n_{\text{train}} \approx 1,500$) with the highest event-rate ceiling in the design and produces the closest-to-singular design matrices in the paper. Two numerical-stability issues arise at this horizon and warrant disclosure.

First, in the S7→S10 *base* specification (Table 14), unpenalized MLE converges overall but the `reg_Other` coefficient exhibits near-separation: the point estimate inflates to $\beta = 33.77$ with a Wald standard error on the order of 10^{13} , reflecting that the `Other` region cell contains effectively no events of the relevant type at this horizon. We omit this single coefficient from the reported table; all other coefficients in the same fit are well-identified and reported with standard maximum-likelihood inference.

Second, in the S7→S10 *enriched* specification (Table 15), the design matrix is rank-deficient under the full enriched feature set and standard MLE fails to converge. For this single horizon-specification combination we fall back to an L2-penalized (ridge) logistic regression via `scikit-learn`. The resulting coefficients are identified but shrunk toward zero by the penalty; standard maximum-likelihood inference (Wald standard errors, z -statistics, p -values) is not valid for penalized estimates and is therefore omitted. Predictive metrics (AUC, Brier) remain valid because they are functions of predicted probabilities rather than parameter-level inference.

The L2 fallback is confined to the S7→S10 enriched cell — the seven other primary specifications (base and enriched at S1→S3, S3→S5, S5→S8, plus base at S7→S10) all converge under unpenalized MLE and are reported with standard inference.

F Era-Control Robustness

The base-spec landmark logits include five-year era dummies indexed by draft year (`era_1985`, `era_1990`, ..., `era_2010`, `era_2015+`). Two features of these dummies merit note given the train/test design. First, the `era_2010` dummy is identified on the subset of train-cohort players whose final NHL season fell within five years of their draft, since the train cutoff (last season ≤ 2014 –15) implies this maximum tenure for that bucket. Second, the `era_2015+` bucket is unrepresented in the training cohort by construction and is therefore absorbed into the reference baseline (`era_1985`) when applied to test-cohort players from later draft years. To verify these features do not drive the headline findings, we refit the $S3 \rightarrow S5$ base logit with all era dummies removed; the continuous covariates retained in the model (age at first season, games played in seasons 1–3, height, weight, draft status, and pick-bin indicators) implicitly carry era-related signal. The Russia coefficient is essentially invariant ($\beta = -1.158 \rightarrow -1.103$; OR = $0.314 \rightarrow 0.332$; $p < 0.001$ in both); aggregate Brier and recall improve substantially under the no-era specification, while AUC declines by 0.005. We retain the era-dummy specification as primary for consistency with the broader landmark-survival literature.

References

1. van Houwelingen, H.C.: Dynamic prediction by landmarking in event history analysis. *Scand. J. Stat.* **34**(1), 70–85 (2007)
2. Dafni, U.: Landmark analysis at the 25-year landmark point. *Circ. Cardiovasc. Qual. Outcomes* **4**(3), 363–371 (2011)
3. Depken, C.A., Ducking, J., Groothuis, P.A.: Career duration in the NHL: pushing and pulling on Europeans? *Appl. Econ.* **49**(59), 5923–5934 (2017)
4. Christie, T., Lavoie, M.: Entry discrimination in the NHL: evolution and the KHL effect. *East. Econ. J.* **41**(2), 214–229 (2015)
5. Fenn, A.J., Gerdes, L., Rothstein, S.: Country of origin effects on the average annual values of NHL player contracts. *Int. J. Financ. Stud.* **7**, 24 (2019)
6. Schuckers, M.: What’s an NHL draft pick worth? A value pick chart for the National Hockey League. Technical report, St. Lawrence University and Statistical Sports Consulting (2011) https://myslu.stlawu.edu/~msch/sports/Schuckers_NHL_Draft.pdf
7. Shoniker, C.: Estimating player survivability in the NHL. Technical report, Stats Sports Consulting (2017), https://www.statsportsconsulting.com/wp-content/uploads/Shoniker_Survival-Analysis-in-the-NHL-Draft.pdf
8. Wang, Y., Csiernik, B., Johnston, K., Baker, J.: Age effects in NHL draftees: a data-driven review of a multi-dimensional concept. *Front Sports Act Living* **7**, 1593409 (2025)
9. Davis, J., Bransen, L., Devos, L., Jaspers, A., Meert, W., Robberechts, P., Van Haaren, J. and Van Roy, M.: Methodology and evaluation in sports analytics: challenges, approaches, and lessons learned. *Machine Learning* **113**, 6977–7010 (2024).
10. NHL Edge API, <https://api-web.nhle.com/v1>, last accessed 2026/04/14
11. Hockey Reference, <https://www.hockey-reference.com>, last accessed 2026/04/14

Who’s In by Game 15?

Multi-Model Per-Game Forecasting of NHL Playoff Qualification

Quinton J. Krueger^{1,2}[0009-0008-9340-2119]

¹ Probility AI, Minnesota, USA

² Northwestern University, Evanston, IL, USA
quinton@probility.ai

Abstract. Most published NHL playoff forecasting systems operate at pre-season or season-end resolution, yet the practical question is not who qualifies but when a forecast becomes trustworthy. We retrain nine heterogeneous classifiers and a uniform-mean ensemble at all 82 regular-season checkpoints, across 14 full seasons from a 40,600 team-game corpus (2008–09 through 2024–25) with 632 cumulative features. Tree-based learners reach ROC-AUC near 0.78–0.82 by game 10, exceed 0.90 by mid-season, and identify 13 of 16 playoff teams by game 15; the tuned random forest peaks at 0.970, while aggressive ℓ_1 regularization collapses the tuned logistic regression to a degenerate flat predictor. Against Win% and Pts% baselines the lift concentrates at 20 of 82 checkpoints. In a per-game betting test, no fixed model or public-metric picker profits over the season, but a Win%-anchored hybrid that resolves the contested bubble seats with the ensemble returns a 0.9% edge against a 7.1% best-model ceiling. The right model is a function of the game checkpoint.

Keywords: NHL · Playoff prediction · Machine learning · Ensemble · SHAP · Betting simulation

1 Introduction

Forecasting NHL playoff qualification well in advance of season end is crucial: a GM weighing a deadline acquisition cares whether the signal is strong enough to act on, and a bettor wants to know when the futures market has stopped tracking the true probability.

This paper builds that game-by-game estimate, retraining nine base learners plus a uniform-mean ensemble at every checkpoint $g \in \{1, \dots, 82\}$ on data observed through game g only, under a strict chronological protocol across 14 full-length seasons. The contributions: (i) we show that the best-performing family shifts across season phases, so a checkpoint-aware framework beats commitment to any single architecture; (ii) we expose the ℓ_1 -tuned logistic regression as a degenerate flat predictor; (iii) we characterize discrimination, ranking, attribution, and economic value against Pts% and Win%

baselines; (iv) we track feature importance shifting from possession proxies to goal share to standings.

Prior work frames the gap. Surveys of the sport-outcome literature report tree ensembles and feed-forward networks outperforming linear baselines, with no single family dominant [4]. Hockey-specific antecedents include Pettigrew's in-game win probabilities [12] and Schuckers's shot-quality-adjusted goaltending evaluation [13], now standard in practitioner workflows; the danger-tier goaltending features of Sect. 3 build on the latter. Khanmohammadi et al. [5] proposed a hybrid architecture for NBA playoff prediction; we compare rather than fuse architectures for the NHL, drawing on LSTMs [3], transformers [15], TreeSHAP [8], and Monte Carlo dropout [2]. No existing work evaluates a matched set of learners on one league, task, and dataset at every game of an 82-game season against public-metric baselines under the same protocol.

2 Dataset, Targets, and Task

The corpus comprises 40,600 team-game rows spanning 17 NHL seasons (2008–09 through 2024–25). Three shortened seasons (the 2012–13 lockout, the 2019–20 and 2020–21 COVID campaigns) are excluded from modeling, leaving 14 full-length seasons; all 17 inform the companion paper [7]. The 632 features are cumulative season-to-date variables covering possession, shot quality, goaltending by danger tier, special teams, and injury severity (man-games lost weighted by cap hit), from public data via Natural Stat Trick [11], plus coach- and GM-change flags. At checkpoint *g* the task is to predict qualification under the rule in force that year, the current format from 2013–14 forward and the six-division format for 2008–12, each yielding exactly 16 qualifiers; the 16 highest predicted probabilities form the model's field.

The proprietary component cannot be redistributed; an anonymized derived feature set reproducing the per-game results on public seasons is available to academic researchers on request, and the released artifacts (predictions CSV, SHAP table, saved checkpoints) reproduce all figures and the betting analysis.

3 Methodology

3.1 Features

All inputs are cumulative: at checkpoint *g*, a feature aggregates games 1..*g* only. The *c_* prefix marks cumulative-to-date; the second component identifies game state (*c_all_*, *c_ev_*, *c_5v5_*, and *score-* and *venue-*adjusted *c_sva_* supply the named SHAP features); the suffix names the metric, GF%/xGF%, SCF%/HDGF%, CF%/FF%, PDO, and *c_Points*, *c_W*, *c_L*. Thus *c_sva_GF%* is cumulative goal share in adjusted 5-on-5 play.

3.2 Models

Nine base classifiers plus an ensemble train at every checkpoint: default logistic regression (L2); ℓ_1 -tuned logistic regression ($C = 0.01$); default random forest [1] (200 trees); tuned random forest (100 trees, depth 10); simple RNN; LSTM [3]; transformer encoder [15]; a Bayesian dense network with Monte Carlo dropout [2]; and a deterministic dense network. The ensemble is a uniform mean of tuned logistic regression, tuned random forest, and LSTM, the combination with the lowest test-season variance.

3.3 Sensitivity

The deep architectures are fixed rather than re-searched per checkpoint. An offline experiment with deeper variants on a wider ($\approx 1,270$ -feature) input, including a stacked LSTM and Keras Tuner searches, indicates the lineup's discrimination is not capped by simple architectures; details in [6].

3.4 Protocol

Features are standardized with column-wise median imputation, scaler and imputer fit only on training data at that checkpoint (fixed `random_state = 42`). The split is chronological: the 10 oldest full seasons train, 2021–22 and 2022–23 validate, 2023–24 and 2024–25 are held out for all reported metrics. Every model retrains at each g on rows at game g ; no information from games $> g$ enters any training fold.

4 Results

4.1 Discrimination by Season Phase

Held-out ROC-AUC separates the models into season phases (Fig. 1). Tree-based learners and default logistic regression reach 0.78–0.82 by game 10; sequence models are volatile early and narrow late without closing the gap; trees cross 0.90 by mid-season, the tuned forest peaking at 0.970. The ℓ_1 -tuned regression collapses to a flat 0.500 everywhere: tuning a sparse model on ROC-AUC alone can leave a degenerate predictor in selection, a failure that a Brier-score or dispersion-floor criterion catches.

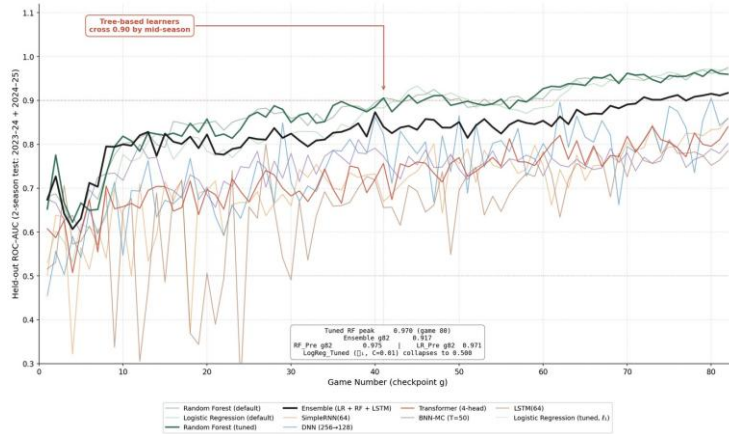


Fig. 1. Held-out ROC-AUC for the 10 model families, averaged across the two test seasons. The flat line near 0.500 is the degenerate ℓ_1 log regression.

4.2 Top-16 Field Identification

Two milestones anchor the title's framing on 2024–25 (Fig. 2): by game 15, the tuned random forest, BNN-MC, and the ensemble each identify 13 of the 16 eventual playoff teams, with three more families (default logistic regression, default forest, RNN) at 12; the first 14 of 16 occurs at game 28 (tuned forest), maintained from game 31 with one regression at game 58, and the ensemble reaches 14 at game 32. Sect. 4.4 reports the Win% convergence: 12 of 16 at game 15, 13 at game 28.

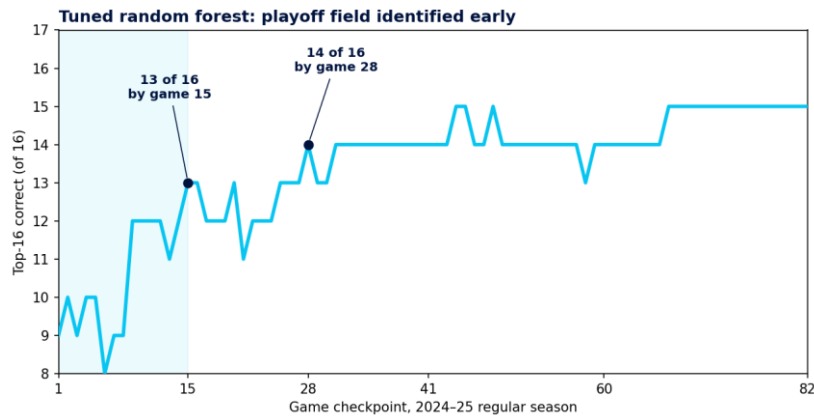


Fig. 2. Tuned random forest top-16 accuracy across 2024–25. The shaded band marks games 1–15; markers show the first 13-of-16 (game 15) and 14-of-16 (game 28).

4.3 Attribution and Ensemble Behavior

Attribution. TreeSHAP [8], extending the SHAP framework of Lundberg and Lee [9], shows the tuned forest's top features traversing three regimes (Fig. 3; naming per Sect. 3): possession and expected-goals proxies lead early ($c_all_GF\%$, $c_ev_xGF\%$, $c_sva_SCF\%$ in the top 8), $c_all_GF\%$ dominates games 21–41, and goal-for shares with the standings triple lead from game 42. Importance concentrates $2.75\times$ across the season, and PDO variants rise late as that signal stabilizes.

Ensemble mechanics. The ensemble is never the single best model at any checkpoint but is the most stable across phases. With the tuned logistic regression degenerate, $P_Ensemble = (0.5 + P_RF_Tuned + P_LSTM)/3$, an implicit shrinkage toward 0.5 on the RF-LSTM consensus, which accounts for part of its compressed range and stability; substituting the default logistic regression is a clean follow-up.

The family pattern has a structural reading: forests partition cumulative features at thresholds that carry signal even on small samples; sequence learners need roughly 20–30 ordered observations to extract trajectory shape; by game 60 cumulative features are nearly deterministic predictors, part of the late 0.97 ceiling (Sect. 6.1).

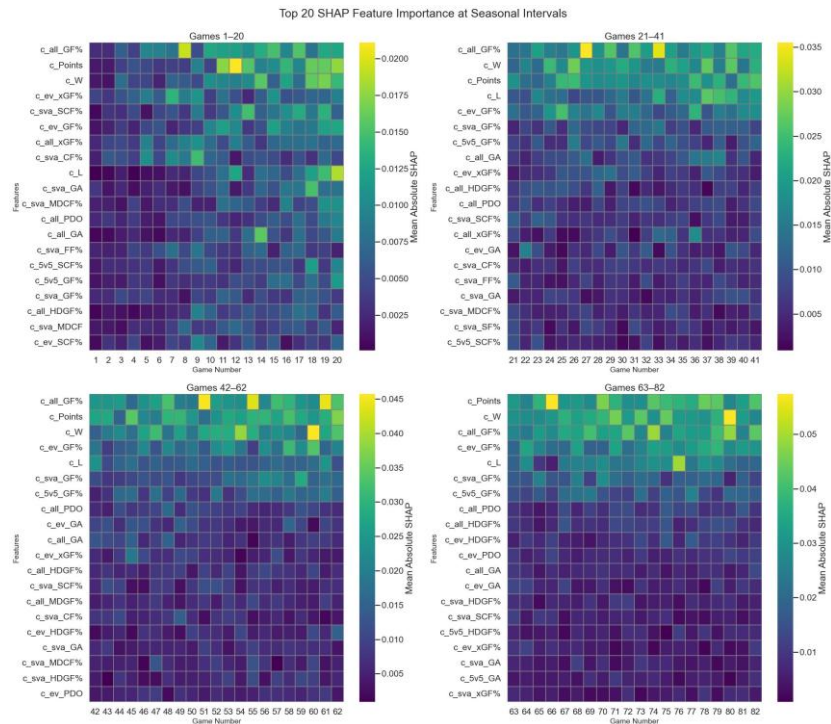


Fig. 3. Top-20 SHAP features (mean absolute value) for the tuned random forest at four season phases: games 1–20 (top-left), 21–41 (top-right), 42–62 (bottom-left), and 63–82 (bottom-right). Rows reordered per panel by within-panel importance; colorbar scales are panel-specific; naming per Sect. 3.

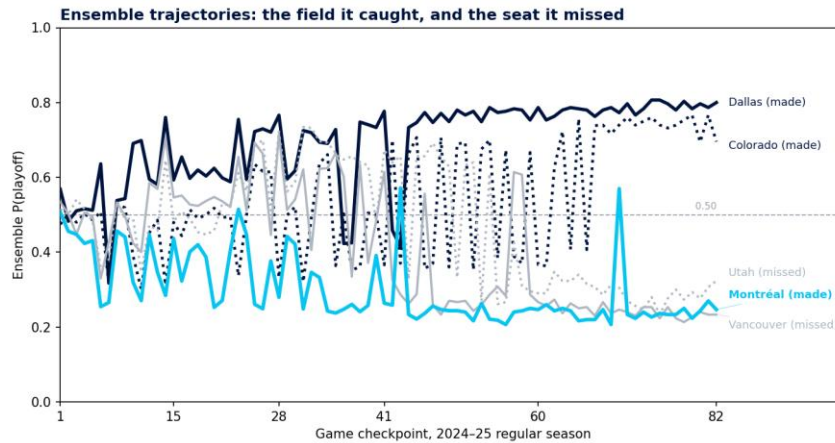


Fig. 4. Ensemble $P(\text{playoff})$ over games 1–82 for five teams; the dashed line marks 0.50. Montréal closes near 0.25 yet qualifies.

Selected 2024–25 trajectories illustrate the ensemble (Fig. 4). Dallas stayed comfortably above 0.50 throughout; Colorado dipped to 0.37 at game 60 before recovering to 0.70, tracking real slumps; Vancouver and Utah declined from game 50. Montréal is the one miss, closing at 0.247 despite qualifying, an honest illustration of why a probabilistic forecast is not a binary one.

4.4 Comparison to Public-Metric Baselines

Does the machinery add top-16 accuracy over a public standings predictor? We compute two baselines on 2024–25 under the identical protocol: Points% (standings points over $2 \times$ games) and Win% (wins over games). Win% is the more demanding reference, 13.62 of 16 mean accuracy against 13.15, because the loser point smooths away the playoff-propensity signal.

The lift over Win% is small and concentrated (Fig. 5). The best of the ten families beats Win% at 20 of 82 games, ties at 50, loses at 12; the ensemble beats at 7 and loses at 33; the tuned forest beats at 6. The lift sits at the tails: +2 lifts at games 2 and 4, six +1 wins from the tuned forest between games 15 and 60, and a perfect 16 of 16 from the RNN at six late checkpoints, first at game 72, where Win% stays at 15.

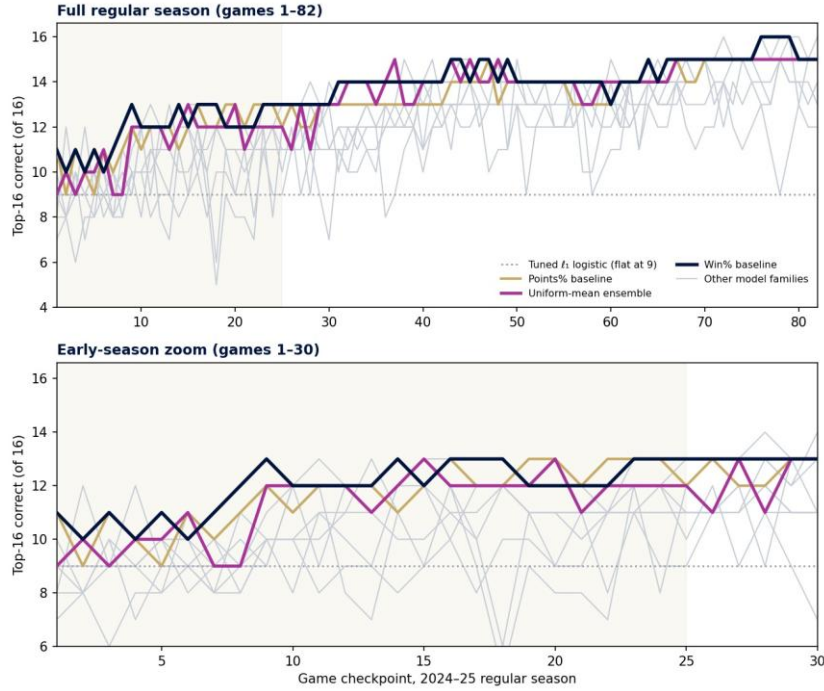


Fig. 5. Top-16 accuracy on 2024–25 for all ten families against the Points% and Win% baselines (bold). Top: full season. Bottom: early-season zoom, band marking games 1–25. The flat dotted line at 9 is the ℓ_1 logistic regression.

The comparison calibrates the framework: it supplies calibrated probabilities (Sect. 5), uncertainty bands, attribution, and protection against worst-case selection, the LSTM falling to 5–7 of 16 at games 18 and 30 and the ℓ_1 regression flat at 9, while the ensemble and tuned forest never fall below 11.

5 Application: Per-Game Betting Simulation

For an economic test, we price each team at every checkpoint against a synthetic qualification line and settle on actual qualification. Live mid-season futures for 2024–25 are not archived, so the game-1 price is the actual BetMGM two-way futures line from SportsOddsHistory [14], devigged, then blended with an Elo-seeded Monte Carlo estimate computed from the standings in place. The blend weight starts at 1 on the pre-season anchor and hands off to the Monte Carlo by about game 35, with a flat 10% two-way margin at every checkpoint. By game 82 the line collapses to the outcome, so a correct pick returns about a dollar on a \$100 stake.

A wager is a single \$100 stake placed at one checkpoint and settled at season end; the price is fixed when the stake is placed, each stake settles once, and payouts differ

across checkpoints because the line moves (Fig. 6). A stake on Montreal to qualify returns \$590 at game 1, \$791 at game 55, and \$1 at game 82; the line priced St. Louis at 4% and Montreal at 10% on 6 February, and each went on a documented late run to qualify [10], so the payout is collectable rather than an artifact. Two views are reported: one round of 32 stakes at a chosen game, MAKE on the strategy's top 16 and NOT-MAKE on the rest, and the cumulative sum of one round per checkpoint. Standings pickers converge by game 82 to a near-perfect field worth nothing in dollars: the Win% picker there finds 15 of 16 yet loses \$173.

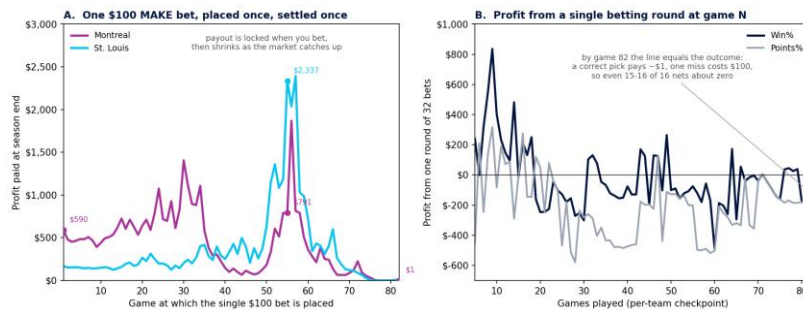


Fig. 6. Mechanics of a single wager. Left: season-end profit from one \$100 MAKE stake on a qualifier placed once at the indicated game, for Montreal and St. Louis. Right: settled profit from one 32-team round placed at each checkpoint for the Win% and Points% pickers.

The payoff is asymmetric by phase (Fig. 7). A bubble seat is one the Monte Carlo still prices between 10% and 90%; by game 40 about 11 of 16 seats are locked above 90%, where a correct stake returns about a dollar, and the four to five contested seats carry the entire payoff, \$634 on average for a contested stake at game 55. A method that resolves those seats half a team to two teams sooner than a Win% picker collects on prices the picker cannot reach, the economic content of the Sect. 4.4 lifts.

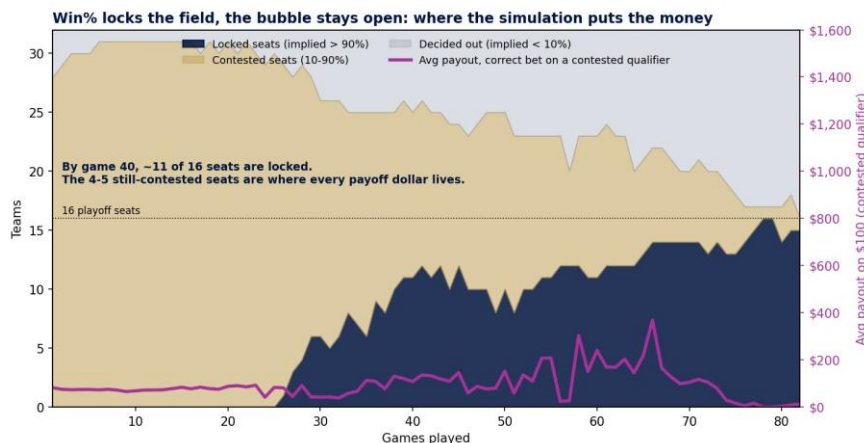


Fig. 7. Playoff-seat decomposition by checkpoint: navy marks seats priced above 90%, gold the contested 10–90% band, grey below 10%; the magenta line gives the average \$100 payout on a correct contested bet.

Fig. 8 reports cumulative net profit from one both-side round at every checkpoint, and Table 1 places the floor, the deployable hybrid, and the ceiling side by side. The deployable pickers split sharply: Points% loses \$15,200 despite tying Win% on end-of-season accuracy, because it keeps backing loser-point teams that miss the cut; the tuned forest loses \$3,200; a walk-forward rotation that switches wholesale to whichever model led most recently loses \$4,300, so rotating everything is the wrong move, and anchoring is the right one. A hybrid keeping Win% on the obvious seats and filling the four bubble seats with the ensemble returns \$2,400 at accuracy comparable to Win% (13.45 vs 13.62), the gain coming from more profitable resolution of the contested seats. The oracle picking the strongest family at each game finishes at \$18,600, and that headroom is structured rather than random, forests leading early and mid-season and the simple RNN late. As return on the \$262,400 staked, Points% loses 5.8%, Win% breaks even, the hybrid returns 0.9%, and the oracle 7.1%; the percentages hold at any turnover, \$9,000 on the hybrid and \$71,000 at the ceiling per \$1,000,000 staked. The deployable edge is small and expensive to earn: the hybrid clears 0.9% after every losing bet is netted out, and reaching even that took the full modeling stack rather than any single public number. An edge near 1% that survives out of sample is still meaningful in a market where the median participant is negative-sum once the vig is paid; the 0.9-to-7.1 gap is headroom rather than a finding.

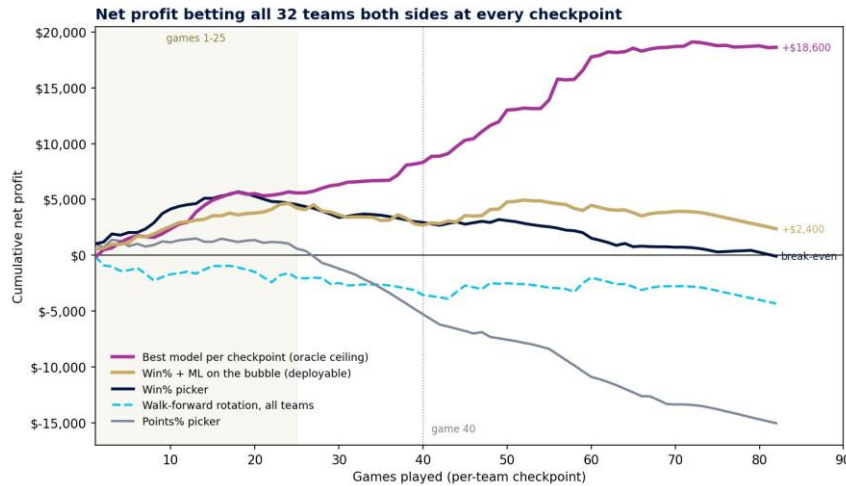


Fig. 8. Cumulative net profit betting all 32 teams both sides at every checkpoint: best model per checkpoint, Win%-anchored hybrid, Win% and Points% pickers, and walk-forward rotation.

Table 1. Full-season outcome by strategy: one both-side round of 32 \$100 stakes per checkpoint, settled on qualification; ROI = net profit / \$262,400 staked.

Strategy	Deployable	Net profit	ROI
Points% picker	Yes	−\$15,200	−5.8%
Tuned random forest	Yes	−\$3,200	−1.2%
Walk-forward rotation	Yes	−\$4,300	−1.6%
Win% picker	Yes	break-even	−0.1%
Win% anchor + ML on the bubble	Yes	+\$2,400	+0.9%
Best model per checkpoint	Oracle ceiling	+\$18,600	+7.1%

The operational recipe follows. Win% reliably finds about 12 of the 16 qualifiers, but those are the seats the market prices short; the money is in the contested seats, exactly the qualifiers Win% misses. Table 2 prices them at game 20: a \$100 stake on the long-est-priced of the four missed qualifiers returns \$642, and all four turn \$400 into \$1,168, a 292% blended return. The value is front-loaded onto the most mispriced team, so the first correct bubble resolution is worth more than the other three combined; repeated across the season, this is the gap between the Win% line and the rotation ceiling in Fig. 8. Anchor on Win%, add the model on the bubble.

Table 2. Value of resolving the contested bubble earlier than Win% at game 20, where Win% misses four qualifiers. Single \$100 stake per team held to season end; return is the single-bet payout, not netting losing stakes.

Beyond Win%	Bubble team	Payout	Return	Cumulative
+1	Montreal	\$642	642%	\$642
+2	St. Louis	\$266	266%	\$908
+3	Ottawa	\$215	215%	\$1,123
+4	Los Angeles	\$45	45%	\$1,168

6 Discussion

Meaningful playoff discrimination is available much earlier than commonly assumed: trees reach 0.78–0.82 ROC-AUC by game 10 and exceed 0.90 by game 41, so a front office or book can act on the top-16 field by roughly game 30. The right family changes across phases, the ensemble never best but most stable.

The baseline comparison bounds the claim: the tuned forest reaches 13 of 16 at game 15, thirteen games before Win%, with the aggregate lift concentrated at 20 of 82 checkpoints. The betting test gives the lift its economic reading: no fixed model or public picker profits alone, and a Win%-anchored hybrid clears \$2,400 where Win% breaks even. The framework's value is protection against worst-case model selection plus a deployable hybrid whose lift concentrates where the line is still long.

6.1 Limitations and Future Work

Cumulative goal-differential and xG-share baselines are the natural next benchmarks, and a leave-one-category-out ablation, refit at each checkpoint, is the natural attribution follow-up. The late-season ROC-AUC ceiling is partly structural, residual uncertainty shrinking as games run out. The betting test uses a synthetic line; a replay against archived quotes and a multi-season test of a learnable rotation are the open follow-ups.

The deep architectures are fixed per checkpoint, and the BNN uses inference-time dropout rather than tuned posteriors; expressive variational BNNs are the natural next step where probability quality matters more than discrimination. Calibration awaits a Platt or isotonic step, manpower-adjusted feature consolidation is the cleaner next feature design, and the ℓ_1 collapse motivates a Brier or dispersion floor when tuning sparse models.

On preprocessing, imputation and scaling are reported under a strict train-only regime; retraining the three deterministic learners at all 82 checkpoints under both regimes changed held-out ROC-AUC by at most 1×10^{-4} with identical top-16 identification, so the metrics are not preprocessing artifacts. Per-checkpoint scaler refitting is residual future work.

Acknowledgments. The author thanks thesis advisers Syamala Srinivasan and Joel K. Shapiro, and Probility AI colleagues Toby Velte, Doug Plocki, and Nick Carone [7].

Disclosure of Interests. The author is Founder and Chief Data Officer of Probility AI, whose proprietary dataset was used in this study and which serves agents, teams, players, and betting markets. No other competing interests are declared.

References

1. Breiman, L.: Random Forests. *Machine Learning* **45**(1), 5–32 (2001). <https://doi.org/10.1023/A:1010933404324>
2. Gal, Y., Ghahramani, Z.: Dropout as a Bayesian Approximation: Representing Model Uncertainty in Deep Learning. In: *Proc. 33rd International Conference on Machine Learning (ICML)*, pp. 1050–1059 (2016)
3. Hochreiter, S., Schmidhuber, J.: Long Short-Term Memory. *Neural Computation* **9**(8), 1735–1780 (1997). <https://doi.org/10.1162/neco.1997.9.8.1735>
4. Horvat, T., Job, J.: The Use of Machine Learning in Sport Outcome Prediction: A Review. *WIREs Data Mining and Knowledge Discovery* **10**(5), e1380 (2020). <https://doi.org/10.1002/widm.1380>
5. Khanmohammadi, R., Saba-Sadiya, S., Esfandiarpour, S., Alhanai, T., Ghassemi, M.M.: MambaNet: A Hybrid Neural Network for Predicting the NBA Playoffs. *SN Computer Science* **5**, 628 (2024). <https://doi.org/10.1007/s42979-024-02977-0>
6. Krueger, Q.J.: Multi-model NHL playoff prediction system. M.S. Thesis, Northwestern University (2025)
7. Krueger, Q.J., Velte, T., Plocki, D., Carone, N.: Net Man-Games Lost as a Coaching and GM Evaluation Metric in the NHL. LINHAC 2026 Submission #3

8. Lundberg, S.M., Erion, G., Chen, H., et al.: From local explanations to global understanding with explainable AI for trees. *Nature Machine Intelligence* **2**, 56–67 (2020). <https://doi.org/10.1038/s42256-019-0138-9>
9. Lundberg, S.M., Lee, S.-I.: A Unified Approach to Interpreting Model Predictions. In: *Advances in Neural Information Processing Systems 30 (NIPS 2017)*, pp. 4765–4774
10. National Hockey League: Stanley Cup Playoffs clinching scenarios and results, 2024–25 regular season. <https://www.nhl.com/>, accessed 2025
11. Natural Stat Trick: NHL statistics. <https://www.naturalstattrick.com/>, accessed 2025
12. Pettigrew, S.: Assessing the Offensive Productivity of NHL Players Using In-Game Win Probabilities. In: *9th Annual MIT Sloan Sports Analytics Conference* (2015)
13. Schuckers, M.E.: Statistical Evaluation of Ice Hockey Goaltending. In: Albert, J., Glickman, M.E., Swartz, T.B., Koning, R.H. (eds.) *Handbook of Statistical Methods and Analyses in Sports*. Chapman & Hall/CRC (2017)
14. Sports Odds History: NHL odds history. <https://www.covers.com/sportsoddshistory/nhl-odds/>, accessed 2026 (archive migrated from sportsoddshistory.com)
15. Vaswani, A., Shazeer, N., Parmar, N., et al.: Attention Is All You Need. In: *Advances in Neural Information Processing Systems 30 (NIPS 2017)*

Validating Expected Possession Value Against Expert Coach Assessments of Passing Intelligence in Elite Youth Ice Hockey

Miikka Palvalin

Finnish Sports Institute, Vierumäki, Finland

Abstract. The quantitative evaluation of passing-related game intelligence in ice hockey remains methodologically underdeveloped. This study examined the convergent validity of Expected Possession Value (EPV) based passing metrics against structured expert coach assessments of passing intelligence in elite youth ice hockey. Data were collected from 120 male players across 12 clubs competing in the Finnish U20 SM-sarja during the 2025–26 season. Coach assessments were obtained via a three-item structured rating instrument demonstrating excellent internal consistency ($\alpha = .950$) and good-to-excellent inter-rater reliability (ICC = .852). EPV-based passing metrics showed moderate convergent validity ($r = .40-.44$) with coach assessments across the full sample, consistently outperforming traditional passing metrics such as pass completion percentage. Positional analyses revealed substantially stronger associations for defensemen (EPV Passes+ / 60, $r = .70$) than for forwards ($r = .39$). The findings support EPV as a theoretically grounded operationalisation of passing intelligence and highlight player position as a meaningful moderator of metric validity.

Keywords: Game intelligence, Expected Possession Value, Passing metrics, Ice hockey, Player evaluation

1 Introduction

The quantitative evaluation of ice hockey players has undergone substantial transformation over the past two decades. Where traditional box-score statistics such as goals, assists, and plus-minus once dominated player assessment, the field has progressively moved toward metrics that capture the quality and context of actions rather than merely their occurrence [1,2]. The most recent advance in this progression is Expected Possession Value (EPV): a class of models that assigns a continuous value to every on-puck action; passes, carries, and turnovers alike based on its effect on the team's probability of scoring [3, 4, 5]. Unlike traditional metrics, EPV is outcome-independent and sensitive to decision quality, making it theoretically well suited to evaluating dimensions of play.

Game intelligence in team sports is broadly understood as the ability to perceive relevant cues, make appropriate decisions under time pressure, and execute those

decisions effectively a perception–decision–execution chain consistently identified as the cognitive basis of expert performance [6, 7, 8]. Passing is the team sport action that most directly engages all three stages simultaneously: the passer must scan available options, evaluate their risk and reward, and deliver the puck accurately to the chosen recipient before the defensive structure closes. Expert players differ systematically from less skilled performers in these perceptual and decision-making stages, using fewer but longer fixations, anticipating outcomes earlier, and generating higher-quality first options [9].

Despite this theoretical alignment, the empirical relationship between automated tracking data and expert human judgement of passing intelligence has not been established. Existing validation work in hockey analytics has benchmarked metrics against objective outcomes; goal rates, team success, player contracts, rather than against independent assessments of specific cognitive behaviours [1]. The closest precedent is Gréhaigne et al. [13], who demonstrated strong agreement between expert observer rankings and structured observational scores in ice hockey; however, that study employed controlled assessment conditions rather than season-long automated tracking data. Coach assessments, when structured and anchored, have been shown to constitute credible criterion measures in sport science contexts [10, 11, 14], yet their potential as a validation instrument for automated hockey metrics has not been exploited.

A further dimension of the validation question concerns playing position. Forwards and defensemen occupy structurally different roles in ice hockey, with different passing responsibilities, movement patterns, and tactical mandates. Radke et al. [12] demonstrated that defensemen and forwards exhibit systematically different passing behaviours in tracking data. Defensemen distributing the puck to forwards significantly more often than to their partners, forwards completing passes through narrower lanes under higher pressure. These structural differences suggest that the relationship between EPV metrics and perceived passing intelligence may not be uniform across positions, a hypothesis that position-stratified validation is necessary to test.

The present study addresses these gaps by examining the convergent validity of EPV-based ice hockey performance metrics against structured multi-rater coach assessments of passing-related game intelligence in a player level. Data were collected from 120 players across 12 clubs competing in the Finnish U20 SM-sarja (Under-20 National League) during the 2025-26 season. Two research questions structure the research paper:

RQ1: To what extent do EPV-based passing metrics correlate with coach expert assessments of passing-related game intelligence?

RQ2: Does the validity of performance metrics vary by player position (forward vs. defensemen)?

2 Related Work

2.1 Game Intelligence and Passing in Team Sports

Game intelligence is widely recognised as a decisive factor in team sport performance, yet it resists simple operationalisation. At its core, the construct refers to a player's capacity to read a dynamic environment, make appropriate decisions under time pressure, and execute those decisions effectively [15, 16]. Research across multiple team sports has converged on a three-stage cognitive chain perception, decision-making, and execution as the functional architecture underlying skilled play [17, 22]

A substantial body of research documents systematic differences between expert and novice performers in the perceptual and decision-making stages. A meta-analysis by Mann et al. [6] covering 42 studies and 388 effect sizes found that experts outperform non-experts in response accuracy and reaction time, and that they employ fewer, longer-duration visual fixations, a pattern indicative of more efficient extraction of task-relevant information. Ward and Williams [8] demonstrated that elite soccer players use situational probability cues to anticipate passing recipients more accurately than sub-elite players, with the advantage increasing with game complexity.

Passing is the team-sport action that most directly engages all three stages of the cognitive chain simultaneously. The passer must survey available options (perception), weigh their strategic value and risk (decision-making), and deliver the puck accurately to the chosen recipient (execution). Johnson and Raab [9] showed that professional handball playmakers generate action options sequentially and that the first generated option is typically of highest quality, supporting a fast heuristic model of expert decision-making in time-pressured situations. A recent multi-sport interview study including elite ice hockey players found that top performers use visual perception both to anticipate upcoming situations and to verify whether play is unfolding as expected before committing to an action [7]; a dual perceptual function consistent with the scanning-then-deciding sequence.

2.2 Measuring Passing Intelligence: From Coach Assessment to EPV

Coach Assessment as Expert Judgement. Expert coach assessments represent the historically earliest and still most commonly used method for evaluating player-level game intelligence in applied sport settings. Structured rating instruments that direct coach attention to specific behaviours have been shown to produce reliable and valid assessments [23]. Sieghartsleitner et al. [10] found that subjective coach ratings demonstrated acceptable concurrent validity against objective motor and tactical performance tests in soccer, and that coach assessments contributed unique predictive variance for future performance beyond what objective tests alone captured. McCormack et al. [14] similarly concluded that subjective and objective approaches provide complementary information when profiling athletes. More broadly, the question of how subjective expert

judgement relates to objective performance measurement is a longstanding methodological concern in sport science [19].

The validity of coach assessment is, however, moderated by several factors. Interrater reliability varies substantially depending on how clearly the rated construct is defined and whether raters share a common reference frame. A study of coach ratings of athlete mental toughness found good inter-rater agreement when a structured instrument was used [18], while studies using unstructured global ratings report considerably lower agreement. Coach assessments are also susceptible to the halo effect – whereby an overall impression of a player disproportionately influences ratings of specific attributes [11].

Traditional Volume and Accuracy Metrics. Traditional hockey statistics approach passing performance through two broad categories: volume measures, which count passes per unit of playing time, and accuracy measures, primarily completion percentage (Pass%). Their limitations for assessing passing intelligence are well documented. Volume metrics do not distinguish between passes that advance play into high-danger areas and passes that merely retain possession in low-risk locations. Completion percentage conflates decision quality with technical execution: a low-risk pass to a nearby team-mate is easily completed, while a high-risk pass that breaks defensive structure carries a lower expected completion rate even if it is the strategically superior choice. Neither metric is sensitive to the context in which a pass is made: the defensive pressure faced, the zone of the ice, or the scoring threat generated. The core problem is that traditional metrics are offensively biased and event-sparse, providing limited signal about the decision-making processes that precede shots and goals [1, 12].

Expected Possession Value. Expected Possession Value (EPV) addresses these limitations by assigning a continuous value to every action based on its effect on the team's probability of scoring. The framework was formalised by Cervone et al. [5] for basketball and subsequently extended to soccer [21] and ice hockey. Murto [3, 4] developed a real-time EPV model for ice hockey using synchronised event and positional tracking data from 872 Finnish Liiga games, implemented through the Wisehockey platform. The model frames possession value as a Markov decision process and defines the reward function as whether either team scores within ten seconds of an action. This model underpins the EPV-derived passing variables used in the present study. Radke et al. [12] confirmed, using NHL tracking data, that EPV-based passing metrics capture dimensions of passing performance not reflected in traditional statistics, and documented systematic positional differences between forwards and defensemen.

3 Methodology

3.1 Participants and Sampling

The study sample comprised 120 male ice hockey players drawn from 12 clubs competing in the Finnish U20 SM-sarja (U20 National League) during the 2025-26 regular season. The U20 SM-sarja is the highest tier of under-20 ice hockey competition in Finland, and the participating clubs represent the development programmes of Finland's top professional organisations. Of the 18 clubs competing in the league, 12 agreed to participate in the coach assessment component of the study, yielding a participation rate of 67%. From each participating club, the six forwards and four defensemen with the highest total full strength ice time in the 2025-26 regular season were included, producing a sample of 72 forwards and 48 defensemen.

3.2 Coach Assessment Instrument

Passing-related game intelligence was assessed using a structured coach rating instrument administered an online survey platform in March 2026, after the completion of the regular season. Each participating club received a survey containing three items assessing different dimensions of passing-related game intelligence, which coaches were asked to rate for each of their six forwards and four defensemen listed in the survey. Players were presented in randomised order within each survey to minimise sequential rating effects, whereby a coach's evaluation of one player might influence their rating of the subsequent player. To enhance the reliability of assessments, each club was asked to provide ratings from a minimum of two coaches.

These items were developed to correspond to the three stages of the perception–decision–execution chain identified in the game intelligence literature [6,7]. Each item (Table 1) was rated on a seven-point scale anchored at the extremes and midpoint: 1 = bottom 10% of the league, 4 = league average, 7 = top 10% of the league. This criterion-referenced anchoring was designed to encourage coaches to use the full range of the scale and to rate players relative to the league as a whole rather than relative to their own roster.

Table 1. Descriptive statistics and reliability of the coaches' assessment scale

Item / Scale	Item wording	<i>M</i>	<i>SD</i>	<i>α</i> if dropped
Perception	The player perceives and identifies multiple passing options.	4.24	1.18	0.906
Decision-making	The player produces decisive or game-breaking passes.	4.30	1.14	0.947
Execution	The player completes passes successfully.	4.26	1.10	0.926

Total (mean of 3 items)	4.27	1.09	$\alpha = .950$, $\omega = .951$
-------------------------	------	------	--------------------------------------

Note. N = 120. Items rated on a 1-7 scale. α if dropped = Cronbach's alpha with that item removed. Total scale reliability: Cronbach's $\alpha = .950$; McDonald's $\omega = .951$.

Prior to the main convergent validity analyses, the psychometric properties of the three-item coach instrument were examined to verify its suitability as a criterion measure. Internal consistency was assessed using Cronbach's α and McDonald's ω . All psychometric analyses were conducted in Jamovi (version 2.7) [24]. Internal consistency of the three-item composite was excellent (Cronbach's $\alpha = .950$, McDonald's $\omega = .951$). Inter-rater reliability, assessed via intraclass correlation (ICC2k, two-way random, average measures), was good-to-excellent (ICC = .852, 95% CI: .809-.887), supporting the use of the coach mean composite as a reliable criterion measure. Descriptive statistics (see also table 2) of the survey results and the density of Coaches assessment variable show that the coaches were able to use the scale well with high standard deviation.

3.3 Performance Data

Objective performance data were obtained directly from Wisehockey (Wisesport Ltd.), an automated real-time sports analytics platform that collects synchronised event and positional tracking data from all U20 SM-sarja games. Data covered the full 2025–26 regular season. All analyses were restricted to 5-on-5 (full-strength) situations and all per-player metrics were normalised to per-60-minutes of 5v5 ice time to ensure comparability across players who differed in total playing time. Across the analytic sample, 5v5 ice time ranged from 6 to 13 hours, with a mean of approximately 10 hours for defensemen and 8,5 hours for forwards.

EPV metrics are derived from the Wisehockey tracking based possession value model [3, 4], which frames puck possession as a Markov decision process and assigns a continuous value to each on-puck action based on its effect on the team's probability of scoring within the following ten seconds. Passes are weighted by the spatial context in which they occur, with passes progressing the puck into the offensive zone and high-danger areas receiving greater positive weight than lateral or retreating passes. EPV related variables in this study are EPV PassTotal / 60, which sums all positive and negative EPV value from each individual pass and is normalized for 60 minutes of 5v5 ice time. EPV Passes+ / 60 sums only EPV from all the passes that has had positive impact on EPV and EPV Passes- / 60 all the negative impacts including missed passes. EPV CarryTotal / 60 which sums all positive and negative EPV value from each individual puck possession that player has had. All variables are based on algorithms provided by Wisehockey.

Table 2. Descriptive statistics for coaches' assessment and game performance variables

Variable	<i>N</i>	<i>M</i>	<i>SD</i>	<i>Skw</i>	<i>Min</i>	<i>Max</i>
Coaches' assessment	120	4.27	1.09	-0.11	1.44	7.00
<i>Passing variables</i>						
EPV PassTotal / 60	120	0.215	0.127	0.972	-0.005	0.676
EPV Passes+ / 60	120	0.454	0.147	0.977	0.143	1.001
EPV Passes- / 60	120	-0.239	0.063	-0.502	-0.475	-0.088
Passes / 60	120	53.61	12.72	0.208	27.57	84.59
Pass%	120	0.807	0.038	-0.647	0.690	0.870
<i>Puck possession variables</i>						
EPV CarryTotal / 60	120	0.194	0.115	1.007	-0.008	0.626
PuckControl / 60	120	2.635	0.717	0.709	1.245	5.047
<i>Scoring threat variables</i>						
xG / 60	120	0.551	0.342	0.684	0.095	1.514
Shots / 60	120	9.60	3.241	1.048	3.30	20.50
xG%	120	0.506	0.066	-0.514	0.306	0.673
xGFor / 60	120	2.678	0.481	-0.334	1.332	3.736
xGAgainst / 60	120	2.604	0.441	1.263	1.754	4.074

The primary analyses examined Pearson *r* correlations between each of the 12 objective performance variables and the coach composite, conducted separately for three analytical groups: all players (*n* = 120), all defensemen (*n* = 48), all forwards (*n* = 72). Statistical significance was set at *p* < .05 throughout. The magnitude of correlations was interpreted using the benchmarks proposed by Hopkins et al. [20]: *r* = .10 small, *r* = .30 moderate, *r* = .50 large, and *r* = .70 very large. All correlation analyses were conducted in Jamovi [24].

4 Results

4.1 Overview

Table 3 presents Pearson correlations between each objective performance variable and the coaches' assessment mean (three-item composite) across all three analytical groups. Bold values indicate significance. Analyses were conducted separately for all players, all defensemen, all forwards.

Table 3. Correlations between coaches' assessment and game performance variables across player groups.

Variable	Full sample n = 120		Forwards n = 72		Defensemen n = 48	
	<i>r</i>	<i>p</i>	<i>r</i>	<i>p</i>	<i>r</i>	<i>p</i>
<i>Passing variables</i>						
EPV PassTotal / 60	.398***	< .001	.332**	0.004	.628***	< .001
EPV Passes+ / 60	.437***	< .001	.388***	< .001	.701***	< .001
EPV Passes- / 60	-.218*	0.017	-.265*	0.025	-.180	0.22
Passes / 60	.372***	< .001	.319**	0.006	.651***	< .001
Pass%	.234*	0.01	0.13	0.276	.430**	0.002
<i>Puck possession variables</i>						
EPV CarryTotal / 60	.249**	0.006	0.161	0.176	.434**	0.002
PuckControl / 60	.351***	< .001	.244*	0.039	.526***	< .001
<i>Scoring threat variables</i>						
xG / 60	0.136	0.139	.316**	0.007	.393**	0.006
Shots / 60	0.08	0.385	0.14	0.24	-.004	0.978
xG%	.391***	< .001	.427***	< .001	.352*	0.014
xGfor / 60	.340***	< .001	.376**	0.001	.318*	0.028
xGAgainst / 60	-.231*	0.011	-.242*	0.04	-.218	0.137

Pearson's *r*. *Italics indicate negative correlations.* * $p < .05$, ** $p < .01$, *** $p < .001$.

4.2 RQ1: Convergent Validity Across the Full Sample

Across all 120 players, EPV-based passing metrics showed the strongest associations with coach-assessed passing intelligence. EPV Passes+ / 60 emerged as the leading predictor ($r = .44$), followed by EPV PassTotal / 60 ($r = .40$), which is considered as moderate correlation [20]. Notably, both outperformed traditional measures: Passes / 60 ($r = .37$) and Pass% ($r = .23$), indicating that the quality-sensitive properties of EPV metrics capture dimensions of passing behaviour that volume and accuracy measures do not fully reflect. The negative association of EPV Passes- / 60 ($r = -.22$) further demonstrates that EPV metrics provide discriminative information in both directions; identifying not only value-generating but also value-reducing passing decisions; a distinction that traditional metrics such as Pass% cannot make.

Secondary associations were observed for puck possession and scoring-threat variables. PuckControl / 60 ($r = .35$), xG% ($r = .39$), and xGfor / 60 ($r = .34$) all reached significance, suggesting that passing intelligence as assessed by coaches also manifests

in broader offensive contributions. Shots / 60 showed no meaningful association ($r = .08$, ns), confirming that shot volume is orthogonal to the construct regardless of metric type.

The results support RQ1 affirmatively: EPV-based passing metrics correlate meaningfully with coach assessments of passing-related game intelligence and consistently outperform traditional passing metrics, with EPV Passes+ / 60 representing the strongest single predictor in the full sample.

4.3 RQ2: Position-Specific Validity

Defenseemen. Across all 48 defensemen, EPV-based passing metrics showed large associations with coach-assessed passing intelligence: EPV Passes+ / 60 ($r = .70$) and EPV PassTotal / 60 ($r = .63$) were the strongest predictors which is considered strong correlation [20]. EPV Passes- / 60 did not reach significance among defensemen ($r = -.18$), suggesting that value-reducing passing decisions are less diagnostically relevant for this position than the positive value generated.

Secondary associations were observed for puck possession and scoring-threat variables: PuckControl / 60 ($r = .53$), EPV CarryTotal / 60 ($r = .43$), xG / 60 ($r = .39$), and xGfor / 60 ($r = .32$) all reached significance, while Shots / 60 showed no association.

Forwards. For the 72 forwards, EPV-based passing metrics remained significant but showed weaker associations than among defensemen. EPV Passes+ / 60 ($r = .39$) and EPV PassTotal / 60 ($r = .33$) were the strongest passing predictors with moderate strength of correlation. EPV Passes- / 60 reached significance ($r = -.27$), suggesting that value-reducing passing decisions are more diagnostically relevant for forwards than for defensemen. Among secondary variables, scoring-threat metrics played a more prominent role for forwards: xG% ($r = .43$) and xGfor / 60 ($r = .38$) were among the strongest predictors overall, reflecting the position-specific emphasis on offensive production in how coaches evaluate forward play.

Across both positions, EPV-based passing metrics consistently outperformed their traditional counterparts. EPV Passes+ / 60 outperformed Passes / 60 among both defensemen ($r = .70$ vs. $r = .65$) and forwards ($r = .39$ vs. $r = .32$), and Pass% failed to reach significance among forwards ($r = .13$, ns) while EPV passing metrics remained significant. EPV Passes+ / 60 was the passing metric to reach the most significance across all three analytical groups, supporting its treatment as the most robust single-metric operationalisation of passing-related game intelligence in this dataset.

The results support RQ2 affirmatively and reveal a clear positional asymmetry. EPV-based passing metrics are substantially more strongly associated with coach-assessed passing intelligence for defensemen than for forwards, and consistently outperform traditional metrics across both positions.

5 Discussion

5.1 RQ1: EPV Metrics Converge with Expert Coach Assessments

The convergence between coach assessments and EPV metrics observed here is broadly consistent with prior work [10, 14] examining the relationship between expert judgement and objective performance data in sport. The present results extend this line of evidence to automated tracking-derived metrics in ice hockey, and further specify the conditions under which convergence is strongest: it is the quality-sensitive, context-weighted EPV metrics that align most closely with what coaches observe and evaluate. This suggests that EPV metrics are not merely a more sophisticated computational tool but a closer empirical approximation of the cognitive dimensions that expert observers attend to when assessing passing intelligence.

EPV Passes+ / 60 emerged as the strongest single predictor ($r = .44$), followed by EPV PassTotal / 60 ($r = .40$), while their traditional counterparts, Passes / 60 ($r = .37$) and Pass% ($r = .23$), showed systematically weaker associations. Correlation around 0.4 level is considered as moderate [20] in behavioral sciences. This ordering is theoretically coherent: EPV metrics weight passing decisions by the spatial context and scoring-threat consequences they produce, whereas volume and accuracy metrics treat all completed passes as equivalent regardless of where and under what pressure they occur. Coaches, as the results indicate, reward progressive decision quality rather than frequency; precisely the distinction the game intelligence literature identifies as central to expert passing behaviour [6, 9].

5.2 RQ2: Position-Specific Validity and Its Theoretical Interpretation

Defensemen. The strength of the defensemen results is theoretically coherent given the structural properties of the position. Radke et al. [12] documented that defensemen distribute the puck to forwards significantly more often than to their defensive partners, and that this directional consistency makes their passing behaviour more systematic and observable. When the primary passing role is to initiate offensive sequences from the defensive zone, the progressive value of those decisions; precisely what EPV Passes+ corresponds with is the central dimension of performance. EPV Passes- / 60 did not reach significance among defensemen ($r = -.18$), suggesting that value-reducing decisions are less diagnostically relevant for this position than the positive value generated a finding consistent with the directional asymmetry of the defensemen role.

Forwards. For forwards, convergent validity was present but consistently weaker, and the predictor profile shifted meaningfully toward scoring-threat metrics. xG% ($r = .43$) and xGfor / 60 ($r = .38$) were among the strongest predictors alongside EPV Passes+ / 60 ($r = .39$). However, xGfor / 60 and xG% are team-context metrics that partly reflect offensive zone time and teammate quality rather than individual passing decisions

specifically. One interpretation is that coaches evaluate forward passing intelligence partly through its immediate offensive outcomes: forwards operate closer to the opposition goal, where the connection between a good pass and a scoring chance is direct and visible, while defensemen's passing quality is more often expressed through play-building sequences whose value materialises several events later. Another reasoning could be that it is more difficult for the coaches to evaluate passing for the forwards for some reason. The forward associations with these variables should therefore be interpreted cautiously. High correlations with other than pass related metrics could potentially be also a cause of Halo-effect like Jekauc & Fritsch [11] have found. Disentangling these contributions would require teammate- and context-adjusted versions of the scoring-threat variables, which the present data do not provide. Notably, EPV Passes+ / 60 reached significance among forwards ($r = -.27$) but not defensemen, indicating that value-reducing decisions are more diagnostically relevant for forward evaluation consistent with the higher decision complexity and risk-reward trade-offs forwards face in offensive zone situations [7,9].

The consistent outperformance of EPV over traditional metrics holds across both positions. Pass% failed to reach significance among forwards ($r = .13$, ns) while EPV passing metrics remained significant, and EPV Passes+ / 60 outperformed Passes / 60 among both defensemen ($r = .70$ vs. $r = .65$) and forwards ($r = .39$ vs. $r = .32$). The results thus support RQ2 affirmatively: EPV-based metrics correspond more closely to coach assessments of passing intelligence across both positions, and the positional asymmetry in the magnitude of that correspondence reflects genuine structural differences in how passing intelligence is expressed and evaluated at each position.

6 Conclusion

The present findings carry several theoretical implications for both the game intelligence and hockey analytics literatures. First, the convergent validity between EPV-based passing metrics and structured coach assessments provides at least moderate empirical support for the theoretical proposition that EPV captures decision-quality dimensions of passing behaviour. Second, the positional asymmetry in convergent validity extends current theoretical understanding of position-specific performance structures in ice hockey. Moderate correlations with variables non-related to passing were surprising and require follow up studies to investigate their relationship with game intelligence.

The results offer concrete guidance for coaches and performance staff working in elite youth ice hockey. EPV Passes+ / 60 emerged as the strongest single-metric predictor of coach-assessed passing intelligence across both positions, suggesting that this variable provides a reliable and efficient complement to subjective observation in player evaluation. Coaches may use EPV-based passing metrics to identify players whose decision quality is systematically high or low relative to peers, supplementing what is visible in game observation with context-weighted quantitative evidence. The positional asymmetry in the findings has direct practical relevance: EPV metrics are

particularly informative for evaluating defensemen, where the play-building role produces a consistent and interpretable passing signal. For forwards, EPV passing metrics capture moderately the passing intelligence with the coaches judgement, but the question remains open if the bias is in the judgement or in the algorithm.

Several limitations warrant acknowledgement and point toward productive future directions. First, the correlational design of the study cannot establish whether EPV metrics causally reflect passing intelligence, especially for forwards. Future work should examine what causes the difference in correlation level between defenders and forwards and find answer to the question whether it is in the coaches bias or in the algorithm. Second, while the coach assessment instrument demonstrated excellent internal consistency and good-to-excellent inter-rater reliability at the composite level, inter-rater reliability at the individual item level was moderate, suggesting that the three dimensions of passing intelligence are assessed more consistently when aggregated than when evaluated separately. Test-retest reliability and sensitivity to halo effects remain uncharacterised, and discriminant validity analyses would further strengthen the evidence base for the instrument's use as a criterion measure in future work. Third, coach assessments in this study are intra-team ratings, meaning each player was evaluated by coaches from their own club. Future work should examine whether EPV metric validity holds when players are assessed on a common scale by external scouts or standardised scouting instruments. Finally, the data is collected from elite youth male players and it remains unclear if the results can be generalized to other player groups.

References

1. Nandakumar, N., Jensen, S.T.: Historical perspectives and current directions in hockey analytics. *Annual Review of Statistics and Its Application* 6, 19–36 (2019).
2. Macdonald, B.: An expected goals model for evaluating NHL teams and players. In: *Proceedings of the MIT Sloan Sports Analytics Conference* (2012).
3. Murto, F.: Towards a real-time possession value framework in ice hockey. In: *Proceedings of the Linköping Hockey Analytics Conference (LINHAC 2023)*, Linköping Electronic Conference Proceedings 201 (2023).
4. Murto, F.: Real-time expected possession value estimation in ice hockey. Master's thesis, Tampere University (2024).
5. Cervone, D., D'Amour, A., Bornn, L., Goldsberry, K.: A multiresolution stochastic process model for predicting basketball possession outcomes. *Journal of the American Statistical Association* 111(514), 585–599 (2016).
6. Mann, D.T.Y., Williams, A.M., Ward, P., Janelle, C.M.: Perceptual-cognitive expertise in sport: a meta-analysis. *Journal of Sport and Exercise Psychology* 29(4), 457–478 (2007).
7. Magnaguagno, L., & Beck, D. (2025). Decision-making process in game sports: what do top-level players think of current research?. *Frontiers in Sports and Active Living*, 7, 1653834.
8. Ward, P., Williams, A.M.: Perceptual and cognitive skill development in soccer: the multi-dimensional nature of expert performance. *Journal of Sport and Exercise Psychology* 25(1), 93–111 (2003).

9. Johnson, J.G., Raab, M.: Take the first: option-generation and resulting choices. *Organizational Behavior and Human Decision Processes* 91(2), 215–229 (2003).
10. Sieghartsleitner, R., Zuber, C., Zibung, M., & Conzelmann, A. (2019). Science or coaches' eye?—Both! Beneficial collaboration of multidimensional measurements and coach assessments for efficient talent selection in elite youth football. *Journal of sports science & medicine*, 18(1), 32.
11. Jekauc, D., Fritsch, J.: The assessment of the halo-effect in a questionnaire rating dimensions of coaches' competence. *International Journal of Sport and Exercise Psychology* 23(1), 163–179 (2025).
12. Radke, D., Lu, J., Woloschuk, J., Le, T., Radke, D., Liu, C., Brecht, T.: Analyzing passing metrics in ice hockey using puck and player tracking data. In: *Proceedings of LINHAC 2023, Linköping Electronic Conference Proceedings* 201 (2023).
13. Gréhaigne, J.F., Richard, J.F., Griffin, L.: The validity and reliability of a performance assessment procedure in ice hockey. *Physical Education and Sport Pedagogy* 14(1), 65–82 (2008).
14. McCormack, S., Jones, B., Elliott, D., et al.: Coaches' assessment of players' physical performance: subjective and objective measures are needed when profiling players. *European Journal of Sport Science* 22(4), 557–567 (2022).
15. Furley, P., Memmert, D.: Creativity and working memory capacity in sports: working memory as a limiting factor of successful tactical performance? *PLoS ONE* 8(7), e68903 (2013).
16. Roca, A., Ford, P.R., McRobert, A.P., Williams, A.M.: Identifying the processes underpinning anticipation and decision-making in a dynamic time-constrained task. *Cognitive Processing* 14(3), 259–266 (2013).
17. Voigt, L., Friedrich, J., Grove, P., Heinrich, N., Ittlinger, S., Iskra, M., ... & Raab, M. (2023). Advancing judgment and decision-making research in sport psychology by using the body as an informant in embodied choices. *Asian Journal of Sport and Exercise Psychology*, 3(1), 47-56.
18. Roberts, A. H., Greenwood, D., Humberstone, C., & Raynor, A. J. (2020). Pilot study on the reliability of the coach's eye: identifying talent throughout a 4-day cadet judo camp. *Frontiers in Sports and Active Living*, 2, 596369.
19. O'Donoghue, P.: *Research Methods for Sports Performance Analysis*. Routledge, London (2010)
20. Hopkins, W. G., Marshall, S. W., Batterham, A. M., & Hanin, J. (2009). Progressive statistics for studies in sports medicine and exercise science. *Medicine & Science in Sports & Exercise*, 41(1), 3–13.
21. Fernández, J., Bornn, L., Cervone, D.: A framework for the fine-grained evaluation of the instantaneous expected value of soccer possessions. *Machine Learning* 110(6), 1389–1427 (2021).
22. Ashford, M., Abraham, A., & Poolton, J. (2021). Understanding a player's decision-making process in team sports: a systematic review of empirical evidence. *Sports*, 9(5), 65.
23. Hendry, D. T., Williams, A. M., & Hodges, N. J. (2018). Coach ratings of skills and their relations to practice, play and successful transitions from youth-elite to adult-professional status in soccer. *Journal of Sports Sciences*, 36(17), 2009-2017.
24. The jamovi project: jamovi, Version 2.7 [Computer Software] (2025). <https://www.jamovi.org>.

Climate and Income Effects on U.S. Hockey Participation

Rodney J. Paul¹, Nick Riccardi¹, Gina Pauline¹, and Andrew Weinbach²

¹ Syracuse University, Syracuse NY 13244, USA

² Coastal Carolina University, Conway SC 29526, USA

Abstract. This paper examines youth hockey participation across U.S. states from 2009 to 2023 using USA Hockey data combined with state-level economic and climate measures. Fixed effects results show that colder climates strongly predict higher participation, while income plays an important enabling role. An interaction between income and winter temperature indicates that participation becomes more sensitive to income in warmer states, where access to indoor ice is required. Random forest models reinforce the strong predictive importance of winter temperature while highlighting the significance of income and rink infrastructure in explaining cross-state variation. Overall, the results show how economic resources can support participation beyond traditional hockey markets.

Keywords: Youth participation · ice hockey · climate · income.

1 Introduction

Participation in youth sports varies widely across geographic regions, reflecting differences in culture, climate, economic resources, and access to facilities. Ice hockey provides one of the most striking examples of these patterns in the United States. Historically, participation was concentrated in colder-weather regions of the country. Today, participation is still heavily concentrated in northern states such as Minnesota, Michigan, and Massachusetts, while remaining relatively limited across much of the southern United States. Understanding the factors associated with these patterns is important for explaining both the historical development of the sport and its potential for future growth in traditional and non-traditional markets.

The nature of the sport of ice hockey requires both expensive equipment and the availability of ice. Therefore, it is more heavily constrained among major youth sports by both environmental and infrastructure requirements. Climate conditions provide a natural starting point for understanding participation patterns. Historically, colder winter temperatures likely contributed to the early development of hockey culture, infrastructure, and participation patterns in northern regions. In warmer climates, however, participation depends almost entirely on artificial indoor ice facilities. Roller hockey and street hockey are close alternatives, but the growth of the game requires available ice. Because constructing

and maintaining ice rinks is costly, participation in warmer regions may require greater economic resources, facility access, and infrastructure investment. As a result, geographic variation in hockey participation reflects not only differences in climate but also the interaction between climate and economic conditions.

This study examines the determinants of hockey participation, focused mainly on youth, across U.S. states using panel data from 2009 through 2023. Participation data from USA Hockey, available on their website, are combined with state-level demographic, economic, and climate variables to analyze how participation varies across both time and geographic regions. The analysis focuses on the roles of winter temperature, income, and urbanization in shaping participation patterns.

The empirical results reveal three main findings. First, winter temperature is strongly associated with participation levels, with colder states exhibiting substantially higher participation rates. This result is broadly consistent with expectations. Second, income plays an important role in facilitating participation, reflecting the relatively high-cost structure of youth hockey, which would, again, be an anticipated result. However, third, and most importantly, the effect of income varies systematically with winter temperature. In warmer states where natural ice is unavailable, participation becomes more sensitive to income, suggesting that financial resources are more strongly associated with sustaining participation where infrastructure constraints are greater.

To examine this relationship more directly, the analysis estimates models that allow the marginal effect of income on participation to vary with winter temperature. The results show that the effect of income increases as winter temperatures rise. In colder regions, participation is less dependent on income because the sport is supported by established infrastructure and cultural traditions. In warmer regions, participation depends more heavily on economic resources that enable access to ice facilities.

Beyond the econometric analysis, this study employs a machine learning approach to evaluate the relative importance of explanatory variables in predicting participation patterns. Random forest models confirm the central role of climate conditions while also highlighting the importance of income and infrastructure in explaining cross-state variation.

Together, these findings help explain both the historical concentration of hockey participation in northern regions and the conditions associated with the expansion of the sport into non-traditional markets. While colder-weather regions remain strongly associated with higher participation, economic resources and infrastructure investment appear capable of supporting participation growth in warmer regions. While this study uses data from the United States, specifically, the findings are likely applicable to any country that has either a wide range of climate conditions across the country, or is predominantly a warm-weather environment.

The contribution of this study extends beyond documenting higher participation in colder-weather states. In particular, the analysis examines how economic resources and infrastructure constraints interact differently across climate condi-

tions and participation groups. The results suggest that participation in warmer regions becomes increasingly dependent on income and facility access, especially for girls' participation, where infrastructure and program availability may be more limited. These findings provide insight into the structural conditions associated with sustaining hockey participation outside traditional hockey markets.

The remainder of the paper proceeds as follows. Section 2 reviews related literature. Section 3 describes the data, Section 4 outlines the empirical model, Section 5 presents the results, Section 6 provides machine learning validation, Section 7 discusses the findings, and Section 8 concludes.

2 Related Literature

A large body of research in sports economics examines the determinants of sport and physical activity participation, with consistent emphasis on economic resources, access to facilities, and social context. Humphreys and Ruseski [1] and Downward and Riordan [2] show that participation in organized sports is influenced by financial costs, time allocation, and social interactions, with higher-income households generally more likely to participate in structured activities. In addition, broader economic conditions and public investment in recreation infrastructure can influence participation behavior [3].

A second strand of literature emphasizes the importance of facility access. Participation in many sports depends on the availability of appropriate infrastructure, which can act as a binding constraint. Prior work finds that proximity to facilities and the supply of recreational infrastructure are strongly associated with participation rates [4, 5]. This constraint is especially relevant for ice hockey, where participation requires consistent access to specialized ice facilities for both practice and competition.

A growing literature also highlights socioeconomic inequality in youth sport participation. Financial barriers, including equipment costs, league fees, and travel requirements, can limit access and create disparities across income groups. Tandon et al. [6] document significant socioeconomic inequities in youth sport participation, particularly in structured and resource-intensive activities. These constraints are especially important in hockey, which is widely recognized as one of the more expensive youth sports.

Related work further shows that participation disparities extend across demographic dimensions, including socioeconomic status and geographic location. Eime et al. [7] find that sport participation is strongly associated with these factors, which influence access to organized programs and facilities. These considerations are particularly relevant for understanding differences in hockey participation across regions and populations.

Despite these advances, relatively little research has examined how environmental constraints interact with economic resources in shaping participation decisions. Climate conditions may influence both the historical development of sports and the feasibility of maintaining necessary infrastructure. For ice hockey,

colder climates historically supported natural outdoor ice and early infrastructure development, while warmer regions require costly indoor facilities.

Finally, emerging hockey-specific research suggests that youth participation responds to broader structural and cultural factors. Kellenberg and Taylor [8] find that professional-level characteristics of the sport can influence youth hockey participation decisions. However, the interaction between climate constraints and economic resources remains largely unexplored.

While prior research has examined economic and infrastructure determinants of sport participation, limited attention has been given to how environmental constraints interact with economic resources in shaping participation in facility-dependent sports such as hockey. This study addresses this gap using state-level panel data.

3 Data

This study examines youth hockey participation across U.S. states using a panel dataset covering the period from 2009 through 2023. Participation data are obtained from annual registration reports published by USA Hockey, which provide counts of registered players by state and age category. These counts are converted into participation rates by dividing by state population estimates from the U.S. Census Bureau.

The primary dependent variable is the per capita hockey participation rate for each state-year observation. Because participation patterns differ by gender, the analysis examines boys' and girls' participation separately in several specifications.

Several explanatory variables are included to capture economic, demographic, and environmental factors that may influence participation. Economic conditions are measured using real per capita income at the state level. State-level income is used as a broad measure of economic resources and affordability conditions within each state rather than as a direct measure of the incomes of participating hockey families. The variable is intended to capture differences in the broader economic environments that may support access to organized hockey participation, including facility availability, equipment affordability, and local program resources. Hockey is a relatively expensive youth sport due to equipment costs, travel requirements, and facility access, suggesting that participation may be highly dependent on income. Demographic characteristics are captured using the share of the population residing in urban areas, which may affect access to organized sports programs and recreational facilities. Climate conditions are measured using average winter temperature, reflecting the importance of ice availability for hockey participation.

In addition, data on ice rink infrastructure were compiled from publicly available listings of ice arena locations across the United States and aggregated to the state level. These data are used to construct a measure of rinks per capita, which serves as a proxy for access to hockey infrastructure and broad cross-sectional differences in facility availability across states. Access to ice facilities

represents a key constraint on participation because practices and games require regular access to rink time. Because rink availability primarily reflects long-run infrastructure differences rather than substantial annual within-state changes, it exhibits limited within-state variation over time.

As a result, rink availability is not included directly in the fixed effects regression models, where identification relies primarily on within-state changes over time. Instead, the rink variable is incorporated later in the machine learning analysis to assess its importance in explaining cross-state differences in participation. Rink availability likely reflects both historical hockey development and current facility access across states.

Table 1 presents summary statistics for the variables used in the analysis, highlighting substantial variation across states in both participation rates and the explanatory variables.

Table 1. Summary statistics.

Sample	N	ln(participation)	ln(income)	Urban (%)	Winter temp (°F)
Boys HS	696	-5.202 (1.370)	11.220 (0.176)	72.36 (14.68)	32.03 (11.96)
Girls HS	682	-7.372 (1.714)	11.221 (0.173)	72.40 (14.61)	31.51 (11.42)
Male All	700	-6.077 (1.323)	11.220 (0.176)	72.44 (14.67)	32.23 (12.22)
Female All	693	-8.245 (1.620)	11.219 (0.176)	72.30 (14.68)	31.87 (11.76)

Notes: HS refers to high-school-age participants, while All refers to all registered participants within each gender category. Participation and income variables are expressed in natural logarithms. Values in parentheses are standard deviations. N represents the number of state-year observations.

To illustrate the relationship between winter temperature and participation, Figure 1 plots state-level hockey participation rates against average winter temperature for all state-year observations in the sample. Separate clusters emerge for male and female participation rates, reflecting substantially higher participation levels among males. Point colors reflect median income levels across observations. The figure illustrates a clear descriptive relationship, with colder-weather states generally exhibiting higher participation levels.

Infrastructure availability also varies widely across states. Figure 2 plots ice rinks per capita against hockey participation rates across the state-year observations in the sample. The fitted line represents the estimated linear relationship between rink availability and participation, while the shaded region reflects the associated confidence interval. States with greater access to ice facilities generally exhibit higher participation rates, highlighting the importance of facility availability in supporting organized hockey participation. While this relationship is descriptive and reflects cross-state variation rather than causal effects, it provides visual evidence that infrastructure availability is associated with participation levels, particularly outside traditional hockey markets.

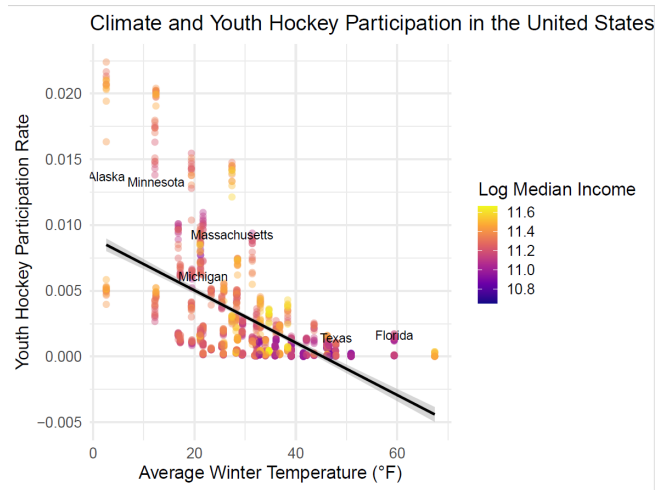


Fig. 1. State-level hockey participation rates versus average winter temperature.

Figure 2. Ice Rinks and Youth Hockey Participation Across U.S. States

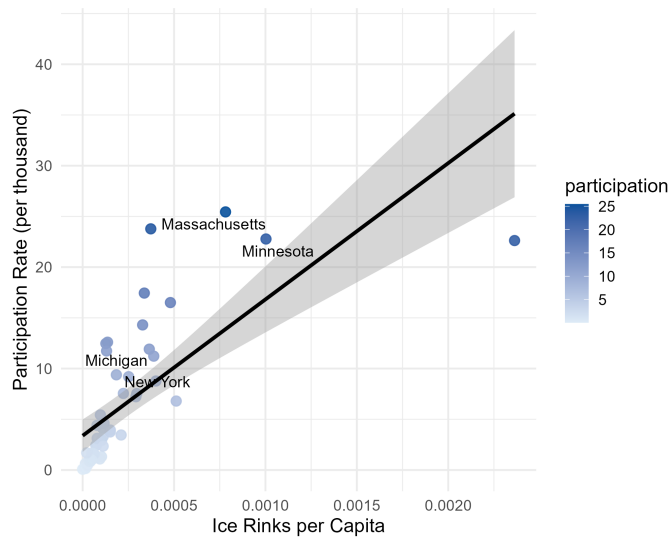


Fig. 2. Ice rinks and youth hockey participation across U.S. states.

Together, these descriptive patterns highlight the importance of climate, economic resources, and infrastructure in shaping youth hockey participation across the United States.

4 Empirical Model

To examine the determinants of youth hockey participation, the analysis employs a state-level panel regression framework. The baseline specification relates participation rates to economic conditions, demographic characteristics, and climate conditions while controlling for unobserved state and year effects.

The baseline empirical model is specified as:

$$\ln(\text{Participation}_{st}) = \beta_0 + \beta_1 \ln(\text{Income}_{st}) + \beta_2 \text{Urban}_{st} + \beta_3 \text{Temp}_{st} + \alpha_s + \gamma_t + \epsilon_{st} \quad (1)$$

where $\text{Participation}_{st}$ represents the hockey participation rate in state s and year t . The variable Income_{st} represents real per capita income, Urban_{st} measures the share of the population living in urban areas, and Temp_{st} captures average winter temperature. We do have an estimation of ice hockey rinks, by state (Rinks_{st}), but they are not included in the econometric specification, as the focus is on variables that exhibit meaningful within-state variation over time and can cleanly be identified in a fixed effect framework. The rinks variable will appear in the machine learning model later in the paper.

The model includes state fixed effects (α_s) to account for time-invariant characteristics such as regional hockey culture, historical participation patterns, and long-standing infrastructure differences across states. Year fixed effects (γ_t) capture national trends affecting youth hockey participation, including broader changes in youth sports participation and economic conditions.

To examine whether economic resources play a larger role in warmer climates, the analysis introduces an interaction between income and winter temperature. This specification allows the effect of income on participation to vary with climate conditions. The interaction model can be written as:

$$\begin{aligned} \ln(\text{Participation}_{st}) = & \beta_0 + \beta_1 \ln(\text{Income}_{st}) + \beta_2 \text{Temp}_s \\ & + \beta_3 [\ln(\text{Income}_{st}) \times \text{Temp}_{st}] + \alpha_s + \gamma_t + \epsilon_{st} \end{aligned} \quad (2)$$

In this specification, the marginal effect of income on participation depends on winter temperature. If participation in warmer states depends more heavily on access to indoor ice facilities, the effect of income should increase as winter temperatures rise. All continuous variables are expressed in natural logarithms where appropriate.

All models are estimated separately for male and female participation in order to examine potential differences in how economic and environmental factors influence participation patterns by gender. Goals of expanding female participation may need different policy intervention than male participation and we aim to explore this idea through the models and their results.

5 Results

Table 2 presents the baseline fixed effects regression results for youth hockey participation. Across specifications, income is positively associated with participa-

tion, consistent with what would be expected of a relatively high-cost structure of youth hockey. States with higher income levels tend to exhibit higher participation rates, reflecting the importance of financial resources for equipment, travel, and facility access.

Urbanization does not appear to be a significant predictor of participation across specifications. In contrast, winter temperature is consistently negative and highly statistically significant, indicating that colder states exhibit higher participation rates. This result aligns with the historical development of hockey infrastructure, participation traditions, and regional hockey culture in colder-weather states.

Table 2. Reduced fixed effects models (year fixed effects).

	Boys HS	Girls HS	Male All	Female All
ln(income)	1.671 ⁺ (0.868)	3.079*** (0.773)	1.230 (0.757)	2.679*** (0.716)
Urban (%)	0.014 (0.010)	0.003 (0.011)	0.012 (0.009)	0.002 (0.011)
Winter temperature (°F)	-0.090*** (0.013)	-0.107*** (0.012)	-0.090*** (0.012)	-0.102*** (0.010)
Num. Obs.	696	682	700	693
R ²	0.748	0.751	0.770	0.781

⁺ $p < 0.1$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$. Notes: Values in parentheses are state-clustered standard errors. Year fixed effects included. HS refers to high-school-age participants, while All refers to all registered participants within each gender category. Num. Obs. represents the number of state-year observations.

To further examine how economic resources interact with winter temperature, Table 3 extends the baseline specification by introducing an interaction between income and winter temperature. For this model, we focus solely on boys and girls of high school age in order to examine gender-specific participation differences at a key youth-development stage where organized participation and retention decisions are especially relevant. The broader male and female participation models were presented first to establish overall participation patterns, while the high-school specifications provide a more targeted comparison of gender differences.

A positive interaction coefficient indicates that the estimated relationship between income and participation becomes stronger as winter temperatures increase, consistent with the idea that economic resources become more important in warmer states where participation depends more heavily on indoor ice infrastructure.

The interaction models reveal important differences between boys' and girls' participation. For boys, neither the income coefficient nor the interaction term is statistically significant, suggesting that participation at the high school level is less sensitive to income variation once climate and fixed effects are accounted

Table 3. Climate interaction models (HS only).

	HS Boys	HS Girls
ln(income)	-0.389 (1.046)	-1.347** (0.484)
ln(income) $\times$ Winter temperature	0.028 (0.035)	0.057*** (0.015)
Num. Obs.	696	682
R^2	0.985	0.974

[†] $p < 0.1$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$. Notes: Values in parentheses are state-clustered standard errors. State and year fixed effects included. HS refers to high-school-age participants. Num. Obs. represents the number of state-year observations.

for. This pattern is consistent with the idea that boys' participation may be more strongly driven by established infrastructure and historical participation patterns.

In contrast, the results for girls indicate a statistically significant interaction between income and winter temperature. The negative coefficient on income, combined with a positive and significant interaction term, implies that the effect of income becomes more positive as winter temperatures increase. In colder states, income plays a limited role in explaining girls' participation, but in warmer states, higher income levels are increasingly important for sustaining participation. This finding suggests that economic resources appear particularly important for supporting girls' hockey in non-traditional markets, where participation depends more heavily on access to indoor facilities and organized programs.

Figure 3 illustrates national participation trends over time. A substantial gap remains between boys' and girls' participation, although girls' participation has generally increased over the sample period, with a temporary decline during the COVID-19 period.

The geographic distribution of hockey participation is shown in Figure 4. Participation is heavily concentrated in northern states, particularly in the Upper Midwest and Northeast, while rates remain comparatively low across much of the southern United States.

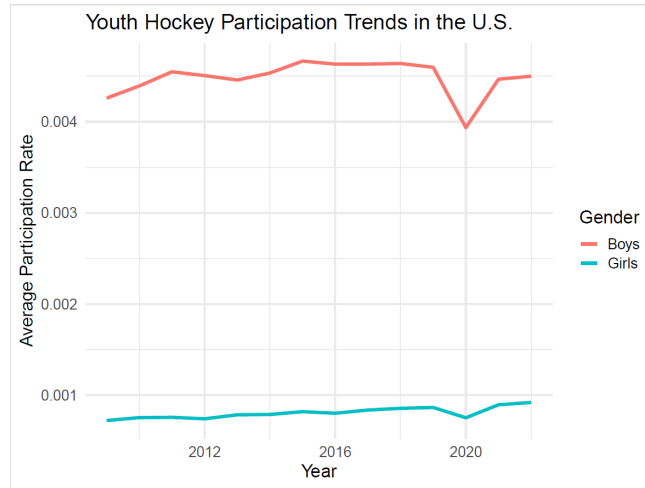


Fig. 3. Youth hockey participation trends in the U.S.

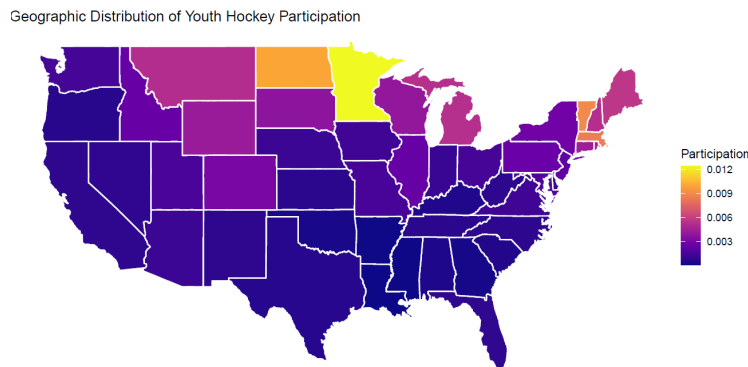


Fig. 4. Geographic distribution of youth hockey participation.

While these descriptive patterns highlight strong geographic clustering, they do not fully capture how economic resources interact with climate conditions. To examine this relationship more directly, Figure 5 illustrates the estimated marginal effect of income on hockey participation across different winter temperature levels using the interaction models reported in Table 3. The figure is based on the high-school-age participation samples and shows how the estimated

relationship between income and participation changes as winter temperatures become warmer. The solid line represents the estimated marginal effect of income, while the shaded region reflects the associated confidence interval. The results indicate that the estimated relationship between income and participation becomes stronger as winter temperatures increase, particularly for girls' participation.

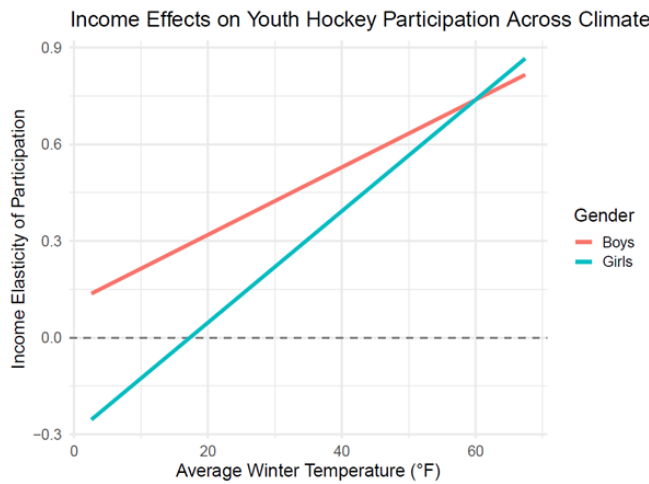


Fig. 5. Estimated marginal effect of income across winter temperature levels.

Figure 5 illustrates a clear and consistent pattern. In colder states, where hockey infrastructure and participation are already well established, the marginal effect of income is relatively small. In these regions, access to the sport is less dependent on financial resources.

As winter temperatures increase, however, the marginal effect of income becomes larger. In warmer states, participation relies more heavily on indoor facilities and organized programs, both of which require substantial financial investment. As a result, higher-income areas are more likely to sustain hockey participation in these regions. This pattern is observed for both boys and girls, although it is more precisely estimated and statistically stronger for girls. Overall, the results indicate that economic resources are strongly associated with sustaining participation in warmer regions where hockey infrastructure and facility access may be more limited.

6 Predictive Validation Using Machine Learning

In addition to the econometric analysis, a machine learning approach is used to evaluate the relative predictive importance of the key variables discussed in pre-

dicting hockey participation. Unlike the fixed effects models, the machine learning framework allows for the inclusion of variables that vary primarily across states, such as rink infrastructure, while also accommodating nonlinear relationships and interactions. While the regression models provide interpretable estimates of the relationships between participation, climate, and income, the machine learning analysis serves as a complementary predictive validation framework that helps assess how strongly variables contribute to participation patterns when fewer functional form restrictions are imposed. Random forests are well suited for this application because they can capture nonlinear relationships and complex interactions without requiring specific functional form assumptions.

Figure 6 presents the variable importance measures from the random forest model. %IncMSE measures the impact of each variable on predictive accuracy, while IncNodePurity captures its contribution to reducing prediction variance across the trees. Winter temperature emerges as the most important predictor of hockey participation across states in both the %IncMSE and IncNodePurity metrics. This indicates that winter temperature is strongly associated with predictive variation in participation across states.

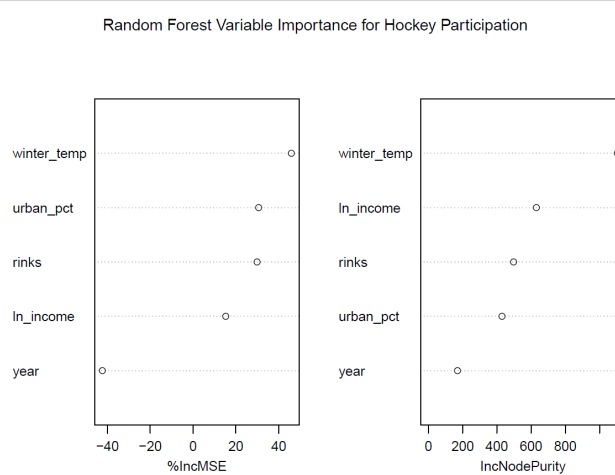


Fig. 6. Random forest variable importance for hockey participation.

Income is the second most important predictor in IncNodePurity, reinforcing the results from the econometric analysis that highlight the role of financial resources in supporting participation, particularly in higher-cost environments. Notably, rink infrastructure also ranks among the top predictors, indicating that access to facilities is a key structural component associated with participation differences across states.

The importance of rink infrastructure in the machine learning model contrasts with its absence in the fixed effects regression results, where limited within-

state variation makes it difficult to identify econometrically. However, its strong predictive contribution in the random forest model suggests that infrastructure is strongly associated with cross-state participation differences.

Overall, these findings complement the regression results by demonstrating that while climate conditions establish the baseline feasibility of hockey participation, economic resources and infrastructure are central to sustaining and expanding participation, particularly in non-traditional markets.

Figure 7 plots actual versus predicted participation from the random forest model using log participation rates across the pooled state-year observations in the sample. Observations are closely aligned with the 45-degree line, indicating strong overall predictive performance. Two distinct bands emerge, reflecting systematic differences in participation across genders. The upper predicted band corresponds to girls' participation, while the lower predicted band reflects boys' participation. Relative to the 45-degree line, the model tends to overpredict girls' participation and underpredict boys' participation. This pattern indicates that while the model captures the primary factors associated with participation, it compresses differences across genders, suggesting that additional gender-specific factors influencing participation are not fully captured by climate, income, and infrastructure variables alone. This compression toward the mean is consistent with earlier regression results showing differential sensitivity to economic and environmental conditions across genders.

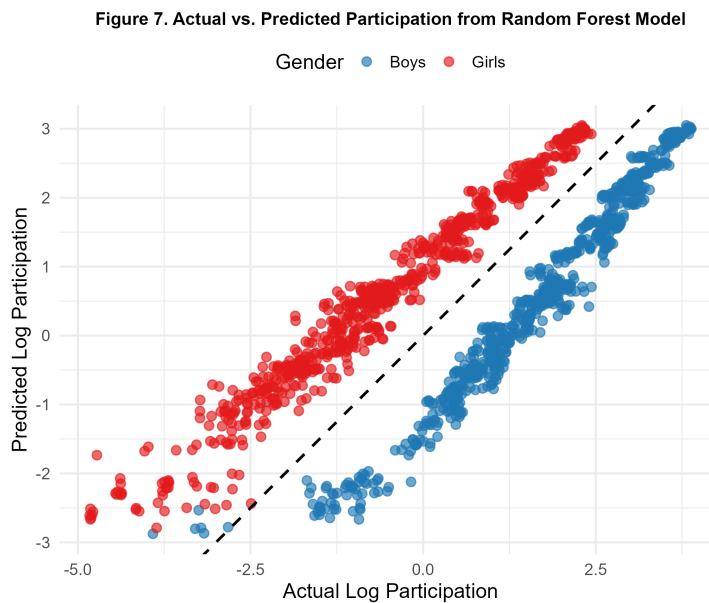


Fig. 7. Actual vs. predicted participation from random forest model.

Taken together, these results reinforce the conclusions from the econometric analysis while also providing additional insight into the role of rink availability and sport infrastructure. Winter temperature remains strongly associated with hockey participation patterns across states, while economic resources and infrastructure availability appear important in supporting participation in non-traditional markets.

7 Discussion

The results of this study highlight the joint roles of winter temperature, infrastructure, and economic resources in shaping hockey participation patterns across the United States. Participation remains strongly concentrated in colder-weather states that historically developed extensive hockey infrastructure and participation cultures. Historically, colder regions likely benefited from conditions that supported the long-term development of hockey infrastructure, participation traditions, and regional hockey culture, contributing to the persistent concentration of participation in northern states such as Minnesota, Michigan, and Massachusetts.

At the same time, the findings show that economic resources appear to play an important enabling role outside traditional cold-weather markets. In warmer regions, participation appears more dependent on access to indoor ice facilities, which require substantial capital investment and ongoing operating costs. The positive interaction between income and winter temperature indicates that participation becomes more sensitive to income as climates become warmer, consistent with greater reliance on costly infrastructure.

Notably, this relationship differs across boys' and girls' participation. The interaction effect is stronger and more precisely estimated for girls, suggesting that female participation is more sensitive to economic conditions, particularly in warmer climates. This suggests that financial and infrastructure constraints may be more binding for girls' hockey, where participation levels are lower and programs may be less established. In contrast, boys' participation appears less responsive to income variation, likely reflecting more established participation pathways and infrastructure.

These results help explain recent patterns of hockey expansion into nontraditional markets. Growth in youth participation in states such as Florida, Texas, and California has occurred alongside investments in rink infrastructure and organized development programs. However, the findings suggest that expansion in these regions is likely to remain more resource-intensive than in colder states where infrastructure and hockey culture are already established.

The results also have implications for youth hockey development strategies. Efforts to expand participation in warmer regions may depend on targeted investments in rink capacity, improved access to ice time, and programs designed to reduce financial barriers for families. Policies that lower participation costs, such as equipment subsidies or community-based development programs, may be particularly effective in regions where climate conditions raise the cost of

participation. These policies may be much more important in encouraging girls' participation in the sport, as girls' participation continuously lags behind boys' participation in ice hockey in the United States.

More broadly, the findings suggest that participation in facility-dependent sports reflects the interaction of environmental conditions, infrastructure availability, and economic resources. While colder-weather regions historically developed stronger hockey ecosystems, infrastructure and income appear associated with the extent to which participation can be sustained and expanded in nontraditional markets. Several limitations should be noted. The analysis is conducted at the state level and therefore captures broad structural relationships rather than individual-level participation decisions. In addition, rink infrastructure exhibits limited within-state variation over time, making it difficult to separately identify within the fixed effects framework. The findings should therefore be interpreted as evidence of strong associations between winter temperature, income, infrastructure, and participation patterns rather than causal effects.

8 Conclusion

This study examines the determinants of youth hockey participation across U.S. states using panel data from 2009 through 2023. By combining participation data from USA Hockey with state-level economic, demographic, and climate variables, the analysis provides evidence on how environmental and economic factors jointly shape participation patterns.

The results show that hockey participation remains strongly concentrated in colder-weather states that historically developed extensive hockey infrastructure and participation cultures. At the same time, economic resources play an important enabling role, particularly in warmer regions where participation appears more dependent on access to indoor ice facilities. The interaction between income and winter temperature indicates that participation becomes more sensitive to income as climates become warmer. This relationship appears particularly important for girls' participation, so to increase female participation in the sport, policy investments in warmer climates to mitigate economic costs may be necessary.

Machine learning results reinforce these findings by highlighting the strong predictive importance of winter temperature while also emphasizing the roles of income and infrastructure in explaining cross-state variation in participation. Overall, the findings help explain the historical concentration of hockey participation in colder-weather regions while also highlighting the importance of infrastructure and economic resources in supporting expansion into nontraditional markets. More broadly, the results illustrate how environmental conditions, infrastructure availability, and economic resources are jointly associated with participation patterns in facility-dependent activities.

Acknowledgments. We thank the conference organizers for the opportunity to submit research.

Disclosure of Interests. The authors have no competing interests.

References

1. Humphreys, B.R., Ruseski, J.E.: Participation in physical activity and government spending on parks and recreation. *Contemporary Economic Policy* **25**(4), 538–552 (2007)
2. Downward, P., Riordan, J.: Social interactions and the demand for sport: An economic analysis. *Contemporary Economic Policy* **25**(4), 518–537 (2007)
3. Humphreys, B.R., Ruseski, J.E.: An economic analysis of participation and time spent in physical activity. *The B.E. Journal of Economic Analysis & Policy* **11**(1), Article 47 (2011)
4. Reimers, A.K., Wagner, M., Alvanides, S., Steinmayr, A., Reiner, M.: Proximity to sports facilities and sports participation. *BMC Public Health* **14**, 1–10 (2014)
5. Wicker, P., Hallmann, K., Breuer, C.: Analyzing the impact of sport infrastructure on sport participation. *Sport Management Review* **16**(2), 211–220 (2013)
6. Tandon, P.S., Kroshus, E., Olsen, K., Garrett, K., Qu, P., McCleery, J.: Socioeconomic inequities in youth participation in physical activity and sports. *Preventive Medicine* **152**, 106716 (2021)
7. Eime, R.M., Charity, M.J., Harvey, J.T., Payne, W.R.: Participation in sport and physical activity: associations with socioeconomic status and geographical remoteness. *BMC Public Health* **15**, 434 (2015)
8. Kellenberg, D., Taylor, M.P.: Externalities of violence: The effects of National Hockey League fighting on youth hockey participation. *Journal of Sports Economics* **26**(5), 562–588 (2025)

Non-Transitivity in the National Hockey League: Ranking Teams, Lines, and Players using Game Theory

David Radke and Troy Mulholland

Chicago Blackhawks, Chicago, IL, USA
`{dradke,tmulholland}@blackhawks.com`

Abstract. The game theoretical complexity of ice hockey creates a large strategy space for teams to exploit. This suggests teams can counteract any opponent strategy; however, the clear distributions in team strength place emphasis on roster construction and collective performance. We analyze the game theoretical landscape of the National Hockey League (NHL) by modeling the transitive and non-transitive interactions between teams based on possession value of events. We utilize recent methods from artificial intelligence (AI) to rank teams, forward lines, and individual players in the NHL over three seasons. Our rankings satisfy a core social choice axiom that Elo and Bradley-Terry rankings violate in 6.8% and 3.7% of rankings respectively due to extensive non-transitivity.

1 Introduction

Sports represent real-world instantiations of many game theoretical concepts. Game theory can provide insight into a game’s underlying structure, or *geometry*, contextualizing the dominance hierarchy of different strategies [7]. While game theory is commonly studied in testbed environments, sports incorporate strategy, skill, and overall ability to shape general game play and competition. Insights about game geometry have been shown to be valuable for evaluating, developing, and understanding artificial intelligence (AI) training algorithms; however, these methods are yet to be widely explored in complex real-world sports.

Ice hockey is classified as an *invasion game* sport, where teams intermingle throughout the environment, invading the opponent’s territory to score points [8, 19]. Invasion games represent direct zero-sum competition between two teams, while the league-wide environment, namely the National Hockey League (NHL) or other professional leagues, involves many more competing interests. Invasion games are rich with multiagent complexity, increasing the overall strategy space for teams to exploit and promoting the emergence of strategy cycles, or *non-transitivity* [32]. Strategies are considered non-transitive if they fail to completely dominate groups of other strategies and become locked in a cycle (i.e., $i \succ j$, $j \succ k$, but $k \succ i$). Popular ranking algorithms typically assume transitive relationships between entities [9, 5]. In pursuit of general evaluation frameworks,

the AI community has proposed ranking algorithms that overcome the documented limitations of assuming transitivity [2, 20]; however, these methods are still typically used across the sporting world [34, 37].

In this paper, we utilize core game theoretical principles to analyze the transitive and non-transitive interactions between NHL teams in the 2021-22, 2022-23, and 2023-24 regular seasons. We observe similar geometrical structures to games studied throughout the broader AI literature and implement an evaluation technique that is guaranteed to satisfy a key social choice axiom that Elo and Bradley-Terry violate in 6.8% and 3.7% of comparisons respectively. Similar to recent work evaluating with AI agents [21], we expand our analysis to evaluate forward lines from their sparse matchup information and we modify that model to distill line value to individual players. We compare our line and player rankings with a player performance metric from [hockeystats.com](https://www.hockeystats.com) and observe moderate correlations [16]. We make the following contributions:

- We implement a possession value model using concepts from reinforcement learning to ground our analysis in sequential decision making.
- We analyze the game geometry of the NHL at the team level and observe similarities to other games studied across AI. We implement a ranking algorithm that is guaranteed to satisfy key social choice axioms.
- We evaluate forward line performance from head-to-head interactions and expand the model to distill line value to individual players.

2 Background

Our work is inspired by several areas across game theory and AI. We provide a brief background of relevant materials and refer the reader to the cited resources for more nuanced background. In line with game theory and social choice vocabulary, we often label teams and lines as “alternatives” or “agents.”

2.1 Strategy Space and Game Geometry

In game theory, a strategy is defined as the course of action that an agent may choose to follow in a specific game [36]. A strategy *space* represents the complete set of all possible strategies a player can play. The dominance relationships between different strategies forms a strategy landscape for the game, or geometry. Understanding this geometry has been a growing topic of interest within AI to support training algorithms in games [7, 28].

Let i , j , and k represent three strategies, or teams. *Transitive* strategy spaces suggest if $i \succ j$ and $j \succ k$, then $i \succ k$. In *non-transitive* strategy spaces, it is plausible that $k \succ i$ due to some exploitable properties or specializations. Commonly used ranking algorithms, such as Elo [9] or Bradley-Terry [5], assume transitive relationships between agents [3]; however, there has been extensive work in AI analyzing the impact of non-transitivity in evaluation and the limitations of existing ranking algorithms [2, 28, 7, 3].

2.2 Voting as Evaluation

Voting has emerged as an alternative technique for evaluating agents in scenarios where other methods have been shown to fail [2, 20]. Voting schemes are designed to aggregate a set of *votes* over a set of *alternatives*. A social choice function is one that selects a “winning” alternative from a set of votes (i.e., rankings or preferences). Given a set of votes, a *Condorcet* winner i is an alternative that is ranked more highly than another alternative j more than vice versa, for all other alternatives j [20]. Voting rules that are guaranteed to return a Condorcet winner when one exists are *Condorcet-consistent*, satisfying a key axiom in social choice theory. Elo and Bradley-Terry are not Condorcet-consistent and have been shown to rank $j \succ i$ in scenarios where i is a Condorcet winner. Condorcet consistency is significant given that the failure to satisfy this axiom can result in pairwise rankings that do not accurately capture the transitive dynamics between pairs of alternatives. Put simply, methods that fail to satisfy this axiom can top-rank an alternative that is actually transitively dominated by another alternative.

We use a probabilistic social choice function called maximal lotteries when analyzing the game geometry of the NHL [12]. Maximal lotteries is calculated over a set of pairwise rankings over alternatives (i.e., $i \succ j$). The collection of all votes forms a voter preference matrix N , encoding the number of votes agent i is preferred over agent j . We transform N into a margin matrix M by calculating $M(i, j) = N(i, j) - N(j, i)$. This margin matrix is interpreted as a zero-sum game where the Nash equilibrium is either a pure strategy (all weight placed on one alternative, the Condorcet winner) or a mixed strategy (non-zero probability assigned to multiple alternatives). Maximal lotteries is computable in polynomial time in the size of the population and can easily identify non-transitive cycles among alternatives at the equilibrium.

Iterative Maximal Lotteries (IML) ranks entire populations of agents by iteratively running maximal lotteries on partitions of the population [20]. IML first runs maximal lotteries on the entire population and removes the alternatives with non-zero probability mass in the equilibrium. These “winning” alternatives create a *level* (i.e., they dominate the remaining alternatives). These alternatives are removed and the process repeats on the remaining population until all alternatives are ranked. Non-transitivity exists between alternatives in the same level but transitive relationships are preserved between levels.

3 Related Work

Ranking and evaluating agents, teams, or players is a core component across both AI and sports. Elo is a simple and common algorithm to rank competitors in zero-sum competition, traditionally used in the game of chess [9] and more recently to rank Large Language Models [6]. Bradley-Terry models are also designed to rank from pairwise comparisons, assume transitivity, and are commonly used in sports [5]. Recent work in AI has shown several limitations of ranking algorithms that assume transitivity depending on the population, including completely flipped and incorrect rankings [2, 20].

To better understand the relationship between transitive and non-transitive strategies, Czarnecki et al., [7] propose the “game of skill” and “spinning top” hypotheses. They argue that real world games have two critical features: the notion of progress as a function of practice (i.e., transitivity) and the availability of diverse game styles (i.e., non-transitivity). These features impose a geometry that resembles a “spinning top,” with a vertical transitive axis and horizontal non-transitive axis. The number of non-transitive strategies is largest in the middle-layers of the transitive axis and smallest in areas with the highest and lowest-skill strategies. They show that purely transitive or non-transitive games fail to exhibit a spinning top geometry. While the goal of that work is to formally understand how AI algorithms learn, the underlying phenomena serves as a powerful tool to contextualize the geometry of ice hockey as we do in this paper.

IML has been shown to effectively rank LLMs, reinforcement learning agents, and humans board game players [20]. We use IML to study the game geometry of the NHL and ensure Condorcet-consistency. In settings with incomplete matchup data, Soft Condorcet Optimization (SCO) learns rankings for all alternatives from data using gradient descent [21]. Cooperative game theory has been shown to measure marginal contributions of agents to teams, both in simulation and basketball scenarios with complete datasets [43]. We implement SCO to evaluate forward lines from their head-to-head interactions, a sparse and incomplete dataset, and expand the SCO model to evaluate individual players.

4 Dataset

We perform our study using Stathletes NHL event data from three NHL regular seasons: 2021-22, 2022-23, and 2023-24. We only consider 5-on-5 even strength play and the following event types: pass attempt, pass receipt, possession gain, area play, carry, deke, and shot attempt. This results in approximately 2,048 events per-game. Shots come with an expected goals (xG) value from Stathletes which we use to train our possession value model, discussed in Section 5. Our event data is augmented with information about player shift to calculate the time-on-ice (TOI) of each forward matchup. We represent three forwards on the ice during an event as a “line,” even if they may be a mixture of multiple lines in the middle of a change.

5 Possession Value Model

Possession value models are common in invasion game sports to learn the expected value of any game state [11, 40]. These models often take the form of estimating the probability of next-goal [1, 11, 18, 35], probability of an action being successful [13], or disruption of opponent structure [14]. We train our possession value model to estimate the sum of future discounted goal value by utilizing core concepts from reinforcement learning (RL) and sequential decision making [39]. Similar to existing work in ice hockey, we treat the problem as a

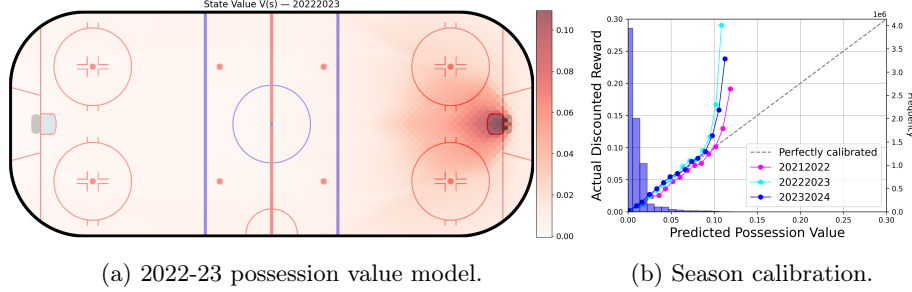


Fig. 1: Our possession value model is trained as a state-value function, predicting the sum of future discounted rewards (xG), consistent with value-based reinforcement learning methods [38]. Darker red represents higher value (attacking right).

Markov Decision Process [42, 27, 15, 40]. However, instead of focusing on a specific future time window, we learn the sum of discounted future rewards directly from trajectories of events. In value-based RL, agents learn a value function to estimate the value of a specific state conditioned on their behavior, or *policy*, π . A value function is formally defined as $v_\pi(s) = \mathbb{E}_\pi[\sum_{k=0}^{\infty} \gamma^k R_{t+k+1} | S_t = s]$, where R is the environment’s reward function and $\gamma \leq 1$ is a hyperparameter to discount future rewards.

We learn a value function using the Stathletes event data. States are represented by their 2-dimensional coordinates and xG values are rewards. The model is trained to estimate the sum of future discounted rewards with a discount factor of $\gamma = 0.99$. We utilize the zero-sum dynamics to inform the direction of value estimates for events that belong to either team. Figure 1a shows our learned possession value model for the 2022-23 season on a rink (attacking right) and Figure 1b shows the calibration plots for each season. The model has higher value closer to the offensive net and is well calibrated over most of the data distribution, only becoming uncalibrated at the edge of the distribution.

6 Rankings From Matchups

We calculate the utility of teams and lines as the sum of state-values for events using the possession value model in Section 5. For any team or line i , we define $V_i = \sum v_\pi(s) \forall s \in S$ when they are on the ice (including events by teammates).

6.1 Team-Level Analysis

NHL teams played 82 regular season games per-year throughout the three seasons in our dataset. Games are distributed across divisional opponents (three to four games), non-divisional conference opponents (three games), and non-conference opponents (two games). This schedule results in a dense matchup matrix since every team plays every other team; however, division and conference opponents have greater impact on a team’s overall performance due to more

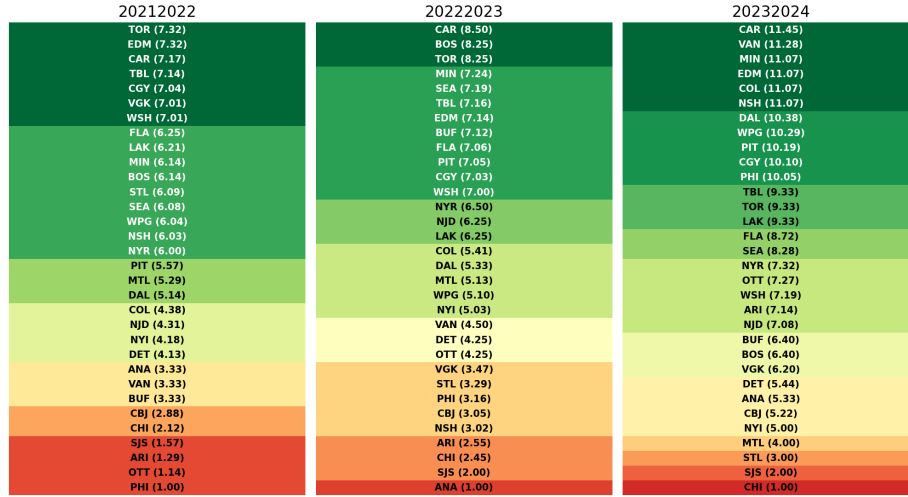


Fig. 2: Results of IML over three seasons. Each color within a season represents a level along the transitive axis, with higher levels dominating lower levels. Teams within the same level have a non-transitive relationship. Scores indicate the level and probability assigned to teams in the Nash equilibrium of the game.

matchups. We define a vote to be a pairwise ranking between teams based on their possession value at the game/period level, where $i \succ j$ when $V_i^\alpha > V_j^\alpha$ for unique game/period pair α .

Team Rankings Figure 2 shows the results of IML rankings in each of the three seasons. We implement IML using the OpenSpiel open source library [22]. The numbers next to each team can be interpreted as their score, or relative strength. The value left of the decimal represents the *level* the team is assigned, where higher scores indicate the team is higher along the transitive axis (i.e., stronger teams). The value right of the decimal represents the probability mass assigned to each team in the equilibrium of the game at that level. Since IML is Condorcet-consistent, integer scores (such as PIT = 6.00 in 2021-22) represent a single Condorcet winner when considering the subset of the population at lower levels (NYR being 6.00 is a function of rounding). Non-integer scores, such as DAL (3.42), CHI (3.33), and BUF (3.25) in 2021-22, represent the mixed strategy Nash equilibrium given the population of teams DAL and below. In this scenario, DAL receives 42% probability mass in the equilibrium, CHI receives 33%, and BUF receives 25%. While teams at the same level may have different probability masses, they do not have a purely transitive relationship between them.

Seasons may have different numbers of total levels based on the amount of observed transitivity, though we observe that middle levels tend to contain most of the population. IML and Elo using this voting system have a mean Kendall-tau distance of 0.64 suggesting they have strong agreement; however, Elo fails to

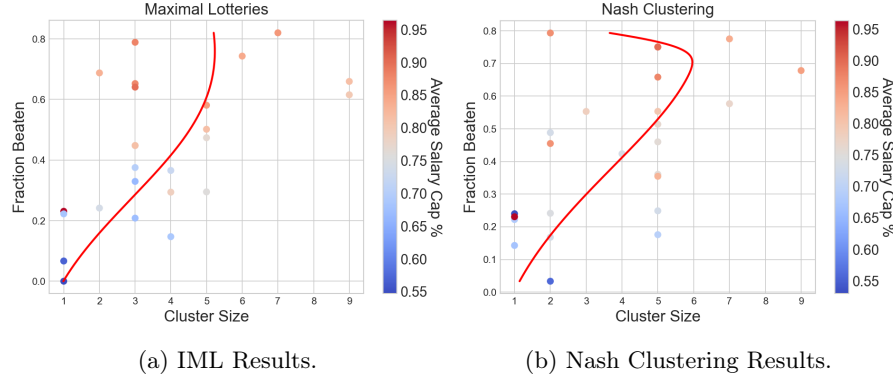


Fig. 3: Evaluating the spinning top game geometry of the NHL. We analyze the geometry with both IML and Nash Clustering. In both cases, we observe similar shapes in the Skewed Gaussian (red line) to the games studied in [7]. The red curve represents a segment of the spinning top geometry.

top-rank a Condorcet winner in 6.8% of pairwise comparisons and Bradley-Terry fails in 3.7%. One example being that Elo ranks $DAL \succ CAR$ in 2023-24 despite CAR beating DAL in the voter preference matrix in every game/period pair. Analyzing levels of non-transitivity may be informative for teams to prioritize relevant opponents or strategies that prevent them from becoming Condorcet winners.

Spinning Top Geometry We recreate the “spinning top” geometry analysis from Table 1 in [7] using two different approaches: IML shown in Figure 3a and Nash Clustering shown in Figure 3b. Nash Clustering, the original method used in [7], is a similar game theory-based approach but with a normalized voter preference matrix. This reduces the higher impact of additional games against division and conference rivals in the Nash strategy. Each dot represents a cluster, or level, of non-transitivity. The cluster sizes are represented on the x-axis and the fraction of opponents beaten in head-to-head matchups is on the y-axis. The red curve shows a fitted Skewed Gaussian highlighting the geometrical shape of the points. Dots are colored by the average percent of the salary cap that teams in the cluster allocate on their roster, discussed further in Section 8.

We observe characteristics of the spinning top geometry in both figures: smaller clusters lower on the y-axis, larger clusters higher on the y-axis, and a return to smaller clusters highest on the y-axis. Whereas most of the analysis in [7] simulates the full range of strategies, our study is limited to observed data, similar to the Alphastar analysis in [7]. This explains why we observe only a subset of the spinning top geometry which is significant to confirm the existence of transitive and non-transitive dynamics.

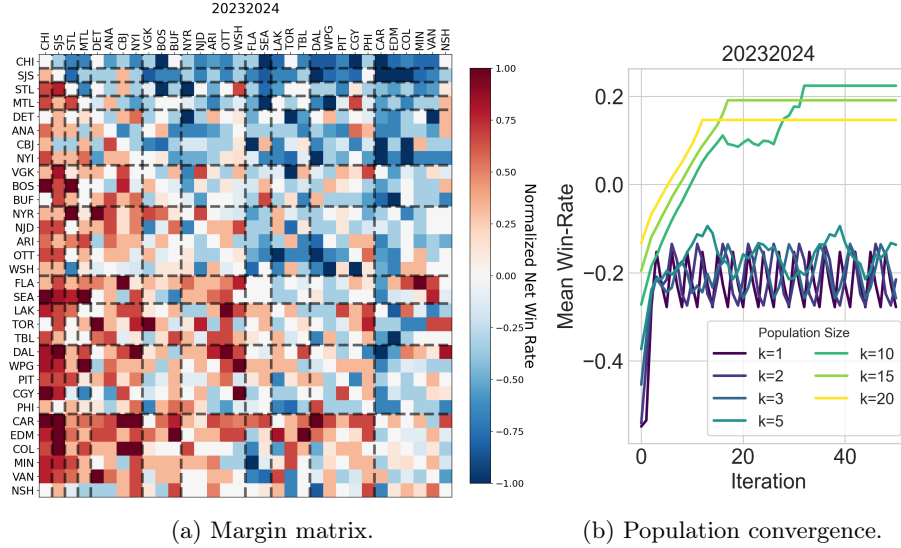


Fig. 4: (a) The 2023-24 margin matrix (normalized). Dark red represents row teams beating column teams and blue represents the inverse. (b) A test that further analyzes the transitivity of the population [7]. We replace the oldest strategy from the collection of k teams with one that beats the whole population on average. Once the sampled population is big enough, the system converges to the strongest strategy. Purely transitive games convergence at all sizes and purely non-transitive games never converge. We observe convergence only when $k \leq 10$ which suggests the existence of both transitivity and non-transitivity.

Margin Matrix Figure 4a shows the empirical margin matrix for all 32 teams against each of their opponents in 2023-24. We normalize the matrix for the number of games for consistent visualization, where red represents a winning margin and blue represents a losing margin. While we observe a general trend of transitivity, where the red area generally fades to blue, there remain matchup clusters that induce non-transitivity. These clusters of non-transitivity (levels in Figure 2) are separated by black dotted lines. The dark red square for row-player CAR and column-player DAL shows the voter preference dominance of CAR $\succ$ DAL that Elo fails to capture, discussed above.

Population Convergence We further evaluate transitivity by recreating another analysis from [7]. We sample k teams from the population and record their win rate against the remaining teams. The oldest team is iteratively replaced using a pessimistic oracle, by the weakest team that beats the entire population on average. Games that are purely transitive converge with a population of 1, while purely non-transitive games are shown to never converge. Figure 4b shows our results in 2023-24 with different values of k . The y-axis shows the mean

win-rate of the k teams against the rest of the population and the x-axis shows iterations of replacement. We find the population fails to converge with small values of $k \leq 5$ but reliably converges with a size $k \geq 10$. This further confirms that ice hockey has significant transitive and non-transitive components.

6.2 Forward Line-Level Analysis

One unique feature of ice hockey compared to many other invasion sports is that players are deployed in targeted coalitions (i.e., lines or defensive pairings) [23]. It is common for coaches to be strategic about using specific lines to against opponent lines; thus, we extend our analysis to consider line matchups. This results in approximately 12,801 unique forward lines per-season. The main challenge in this analysis is that only 0.04% of the margin matrix contains non-zero entries while IML and Nash Clustering assume complete data. We instead aim to learn rankings and scores for all lines by solving an optimization problem on a dataset.

Soft Condorcet Optimization (SCO) has been shown to effectively rank AI agents in settings with incomplete matchup data [21]. Let i and j represent two lines that play against each other for t seconds in game g . We define a vote to be an pairwise ranking between lines at the game level based on possession value, $i \succ j$ when $V_i^g > V_j^g$. Let $\mathcal{V}$ be the set of all votes and $\mathcal{T}$ be the set of all matchup time-on-ice (TOI) values, $t \in \mathcal{T}$. SCO learns a set of numerical values θ over the set of lines by minimizing the sigmoid loss function shown in Equation 1 where $\sigma(z) = (1 + e^{-z})^{-1}$ and T is a temperature hyperparameter. We weight each gradient update by $w_{ij} = \frac{t}{\max(\mathcal{T})}$ so that matchups with higher TOI are weighted more heavily. This aims to minimize the Kendall-tau distance between values θ and the set of votes in the training dataset. Consistent with past work, the model is trained on 80% of matchups and evaluated on a held-out test set. We ensure that every line in the test set appears in the training set but there is no overlap in actual matchups between the two datasets.

$$\mathcal{L}(\theta) = \sum_{(i,j) \in \mathcal{V}} w_{ij} \sigma\left(\frac{\theta_j - \theta_i}{T}\right) \quad (1)$$

Figure 5a shows a CDF plot for the distribution of learned SCO θ values in each season (minimum 100 matchups). Each season learns similar distributions of θ values and the shorter tails on the lower end of the distribution highlights an interesting trend of survivorship bias: poorly performing lines tend to not achieve the matchup threshold we impose when plotting.

Figure 5b shows accuracy results when predicting the winner of matchups in the test set based on the line with higher SCO θ value. We perform 10-fold cross validation over sampled minibatches of votes from the training set. The x-axis shows training iterations and the y-axis shows mean accuracy on the test set with 95% confidence intervals. We compare predictions with a public wins-above-replacement (WAR) metric from [hockeystats.com](https://www.hockeystats.com) [16] which ranks $i \succ j$ according to the mean WAR of the line. Their WAR metric is based on a regularized adjusted plus-minus (RAPM) model to measure individual player

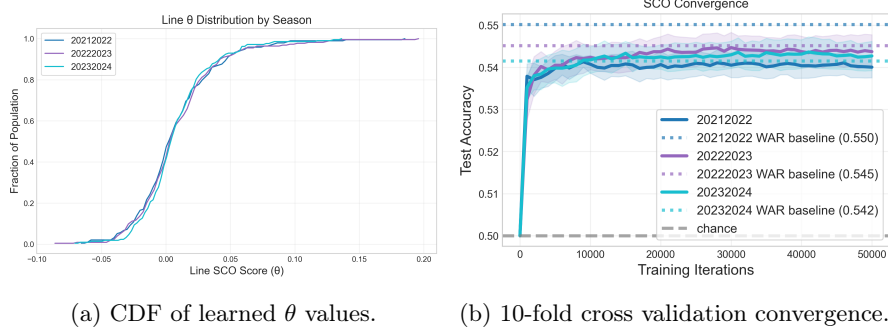


Fig. 5: (a) A CDF plot showing the distribution of learned SCO θ values for lines in each season (minimum 100 matchups). (b) We perform 10-fold cross validation to show the ranking accuracy of the SCO θ values on a held-out test set. Lines show the mean accuracy with 95% confidence intervals.

impact as a function of xG and game context. Notably, they consider special teams information whereas we only consider 5-on-5 play. SCO reliably learns θ values significantly above chance, showing that the model learns signal with tight confidence intervals. The SCO model achieves similar performance in two of the three seasons and is only slightly below WAR in 2021-22.

6.3 Player-Level Analysis

SCO learns θ values by treating lines as unique entities. In reality, lines are composed of individual players that can belong to multiple lines. Line-level matchups are more natural units of analysis than individual matchups in a sport that emphasizes coalition structures; thus, we utilize line-level matchup data but expand the SCO formalization to distribute gradient updates to evaluate individual players. This player SCO model learns values over the set of unique players in the dataset ϕ instead of line values θ . Let $\Phi(L_i) = \frac{1}{|L_i|} \sum_{p \in L_i} \phi_p$ represent the estimated value of line L_i , instead of θ_i in Section 6.2. We minimize the same objective as Equation 1 but replace θ_i and θ_j with $\Phi(L_i)$ and $\Phi(L_j)$ respectively. Gradients flow evenly to player SCO ϕ values for individual players in each line.

Baseball Proof-of-Concept In a sport like baseball, value is more easily attributed to individual players compared to invasion games. To evaluate our extension to SCO, we create an analogous baseball scenario and compare learned SCO ϕ values to observed individual player performance. We use MLB plate appearance data collected from MLB’s Stats API for regular season games from 2020 to 2025 [25]. We define a line to be the set of pitcher and batters for each team in unique game/inning matchups. A vote ranks two opposing lines based on cumulative weighted on-base average (wOBA) or runs above average (RAA) in that inning. Linear weights of plate appearance outcomes are calculated using a

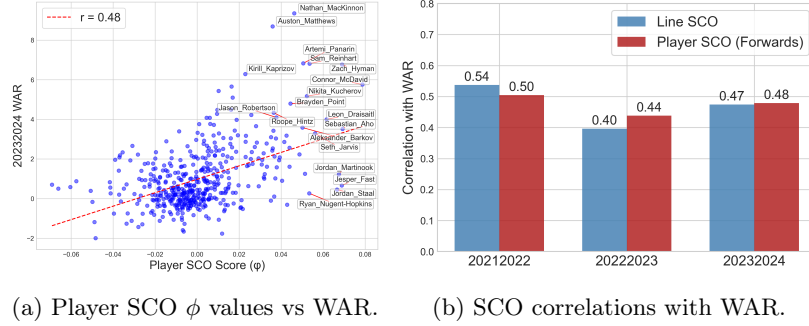


Fig. 6: (a) The 2023-24 season scatterplot comparing player SCO ϕ values and WAR from [16]. We label the players high along both axes. (b) Correlations with WAR between line-level SCO θ values and player SCO ϕ values.

run expectancy matrix over those seasons [10]. We train a simple model to predict batted ball outcome probability given batted ball exit velocity, spin, spray angle, and launch angle to estimate plate appearance expected wOBA (xwOBA) and expected RAA (xRAA). Our learned SCO ϕ values achieve moderately strong correlations over the population of considered players (greater than 75 pitches or 300 at-bats): 0.48 with xwOBA values (0.6 for pitchers and 0.39 for batters) and 0.55 with xRAA values (0.7 for pitchers and 0.44 for batters). These results indicate that our extension can effectively disentangle player performance from group-level matchups in baseball.

Ice Hockey Results Figure 6a compares the learned player SCO ϕ values for individual players along the x-axis with their observed WAR from [16] along the y-axis in the 2023/2024 season. We label several players that achieve high scores along either axis. We observe a moderate correlation of 0.472 which is consistent across other seasons (red bars in Figure 6b, line SCO θ values are the blue bars).

SCO ϕ values rank players like Connor McDavid, Sebastian Aho, and Zach Hyman among the highest in the league. An interesting takeaway is the cluster of Carolina linemates in the bottom right corner: Jesper Fast, Jordan Martinook, and Jordan Staal. The line of Fast/Martinook/Staal ranks highest in line-level SCO θ value in 2023/2024, winning 64% of their matchups against other lines and achieving the highest net-difference in possession value generated vs conceded. They also posted a +74 rating in just under 390 minutes together, ranking fifth-best in the NHL that season despite the WAR model ranking all three players below two wins individually. These players also played a considerable amount in this line until Seth Jarvis replaced Jesper Fast on the line half-way through the season, another player who is ranked highly by SCO.

One possible source of inconsistency between WAR and SCO could be the distribution of opponents. While SCO does adjust for opponent strength in the sigmoid loss, one limitation is that adjustments are ultimately a function of

margin matrix shape and density. If there is minimal overlap between specific clusters of lines, the model may never observe overlapping clusters that could adjust gradient updates properly. Even though the Fast/Martinook/Staal line faces opponents with slightly positive SCO values, an average value of $\theta = 0.004$, more observations against a range of different opponents or future simulation techniques could diagnose the root of this inconsistency.

7 Discussion

The Nash equilibrium strategy is positioned at the top of the transitive axis in the spinning top geometry. AI algorithms that continue to improve from experience are expected to converge towards this Nash equilibrium. This idea is intriguing in the context of sports, considering the sentiment around analytics making sports less entertaining by optimizing for single objectives [44, 26]. What does a Nash equilibrium look like in sports? Is there a unique Nash strategy or multiple?

Absent of any rulebook-based optimization strategy that reduces the search space (i.e., goals are worth the same), converging on a Nash strategy in ice hockey would rely on discovering effective ways to dominate opponents in any area of play. Player skill and ability has a large impact on the ability to execute strategies. In a world without roster constraints, it is theoretically possible to construct a team capable of climbing higher along the transitive axis towards the game's Nash equilibrium by simply acquiring the best talent.

As discussed in Section 6.1, we observe only partial segments of the spinning top geometry in the NHL due to a skewed sample of strategies. The shape of the Skewed Gaussian resembles the lower and middle portions of the geometry despite no strategy actively trying to lose (what should be the bottom of the transitive axis). In the context of sports, we argue that the spinning top geometry instead highlights the long-term lifecycle of teams imposed by roster constraints. We color the clusters in Figure 3 according to the average percentage of the salary cap allocated where darker red represents teams closer to the cap ceiling. We observe a strong positive correlation of 0.78 between the transitive axis and average salary cap allocation.

Making the connection between spending and performance is not a novel insight. Instead, we emphasize the relationship between key market pressures such as the hard salary cap and free agency and the observed underlying game geometry of the league. In a sport where constraints limit their ability to climb the transitive axis, the best teams will likely remain in non-transitive cycles and be forced to emphasize strategy to win.

8 Conclusions and Future Work

We show how fundamental concepts from game theory can be useful in analyzing real-world sports. One of our core contributions is highlighting the extent to which non-transitivity exists in the NHL and the geometrical similarities with games studied across AI. Given the extensive research in AI towards better

evaluation techniques that consider non-transitive spaces, our work makes a key contribution towards motivating future work at the intersection of AI and sports.

Several areas of future work exist to extend our analysis. Team and line rankings could expand to consider other manpowers or values specific to different areas of play, such as in-zone or off the rush, similar to how recent ranking algorithms in AI measure the impact of different tasks [24]. Recent research with ice hockey tracking data could better contextualize possession values and strategies of teams or lines [4, 17, 29, 31]. While we expand SCO to allocate the line-level gradient to players evenly, recent work in AI has shown that groups of agents may benefit from having different amounts of shared incentives [30, 41]. While SCO is designed for incomplete data, further investigation around densifying the margin matrix through simulation or normalization could be beneficial [33, 37].

Our work makes several connections to the broader AI and game theory literature. We utilize three seasons of Stathletes NHL event data to train a possession value model grounded in reinforcement learning theory and sequential decision making. We contextualize our work within the broader discussion of game geometry and hypothesize about how external constraints on roster construction shape the underlying landscape of the NHL to emphasize strategy.

Acknowledgments. We greatly appreciate the support of Meghan Chayka and Neil Lane from Stathletes for providing the data that serves as the basis of this study. We also thank Dr. Kate Larson from the University of Waterloo and Google DeepMind for initial conversations and feedback during the ideation phase and Dr. Alexi Orchard from the University of Notre Dame for thoughtful reviews and feedback on earlier versions of this manuscript. Finally, we'd like to thank the reviewers and program committee members for their feedback.

References

1. Awosoga, D., Cui, J., Li, A., Majumdar, J., Wang, J.: Comparing defensive pressures using possession retention probability and expected goals. In: Linköping Hockey Analytics Conference Student Competition (2023)
2. Balduzzi, D., Tuyls, K., Perolat, J., Graepel, T.: Re-evaluating evaluation. *Advances in Neural Information Processing Systems* **31** (2018)
3. Bertrand, Q., Czarnecki, W.M., Gidel, G.: On the limitations of the elo, real-world games are transitive, not additive. In: *International Conference on Artificial Intelligence and Statistics*. pp. 2905–2921. PMLR (2023)
4. Bischofberger, J., Baca, A.: Dangerous accessible space: a unified model of space and value in team sports. University of Vienna (2025)
5. Bradley, R.A., Terry, M.E.: Rank analysis of incomplete block designs: I. the method of paired comparisons. *Biometrika* **39**(3/4), 324–345 (1952)
6. Chiang, W.L., Zheng, L., Sheng, Y., Angelopoulos, A.N., Li, T., Li, D., Zhang, H., Zhu, B., Jordan, M., Gonzalez, J.E., et al.: Chatbot arena: An open platform for evaluating llms by human preference, 2024. URL <https://arxiv.org/abs/2403.04132> **2**(10) (2024)

7. Czarnecki, W.M., Gidel, G., Tracey, B., Tuyls, K., Omidshafiei, S., Balduzzi, D., Jaderberg, M.: Real world games look like spinning tops. *Advances in Neural Information Processing Systems* **33**, 17443–17454 (2020)
8. Ellis, M.: Similarities and differences in games: A system for classification. In: *International association for physical education in higher education (AIESEP) Conference* (1983)
9. Elo, A.: *The rating of chess players, past & present*. second printing (2008) (1978)
10. fangraphs.com: fangraphs.com. <https://www.fangraphs.com/>, accessed: 2026-04-20
11. Fernández, J., Bornn, L., Cervone, D.: Decomposing the immeasurable sport: A deep learning expected possession value framework for soccer. In: *13th MIT Sloan Sports Analytics Conference* (2019)
12. Fishburn, P.C.: Probabilistic social choice based on simple voting comparisons. *The Review of Economic Studies* **51**(4), 683–692 (1984)
13. Goes, F., Schwarz, E., Elferink-Gemser, M., Lemmink, K., Brink, M.: A risk-reward assessment of passing decisions: comparison between positional roles using tracking data from professional men’s soccer. *Science and Medicine in Football* **6**(3), 372–380 (2022)
14. Goes, F.R., Kempe, M., Meerhoff, L.A., Lemmink, K.A.: Not every pass can be an assist: a data-driven model to measure pass effectiveness in professional soccer matches. *Big data* **7**(1), 57–70 (2019)
15. HockeyGraphs, Kumagai, B., Nahabedian, M., Châtel, T., Stokes, T.: Bayesian space-time models for expected possession added value. <https://hockey-graphs.com/2021/07/06/bayesian-space-time-models-for-expected-possession-added-value-part-1-of-2/> (2021), accessed: 2026-05-18
16. hockeystats.com: hockeystats.com. <https://hockeystats.com/>, accessed: 2026-04-20
17. Inayatoli, H., Chan, T.: Evaluating space creation in the national hockey league using puck and player tracking data. In: *Linköping Hockey Analytics Conference*. pp. 13–25 (2024)
18. Kervin, E.: Optimus reim: A sequence-aware expected threat framework for player evaluation in professional hockey. <https://019d40b7-44ff-0a16-e037-1e0bebdcd87.share.connect.posit.cloud/> (2026), accessed: 2026-05-18
19. Lamas, L., Barrera, J., Otranto, G., Ugrinowitsch, C.: Invasion team sports: strategy and match modeling. *International Journal of Performance Analysis in Sport* **14**(1), 307–329 (2014)
20. Lanctot, M., Larson, K., Bachrach, Y., Marris, L., Li, Z., Bhoopchand, A., Anthony, T., Tanner, B., Koop, A.: Evaluating agents using social choice theory. *arXiv preprint arXiv:2312.03121* (2023)
21. Lanctot, M., Larson, K., Kaisers, M., Berthet, Q., Gemp, I., Diaz, M., Maura-Rivero, R.R., Bachrach, Y., Koop, A., Precup, D.: Soft condorcet optimization for ranking of general agents. *Proceedings of the 24th International Conference on Autonomous Agents and Multiagent Systems* (2025)
22. Lanctot, M., Lockhart, E., Lespiau, J.B., Zambaldi, V., Upadhyay, S., Pérolat, J., Srinivasan, S., Timbers, F., Tuyls, K., Omidshafiei, S., Hennes, D., Morrill, D., Muller, P., Ewalds, T., Faulkner, R., Kramár, J., Vylder, B.D., Saeta, B., Bradbury, J., Ding, D., Borgeaud, S., Lai, M., Schrittwieser, J., Anthony, T., Hughes, E., Danihelka, I., Ryan-Davis, J.: OpenSpiel: A framework for reinforcement learning in games. *CoRR* **abs/1908.09453** (2019), <http://arxiv.org/abs/1908.09453>

23. Ljung, D., Carlsson, N., Lambrix, P.: Player pairs valuation in ice hockey. In: MLSA@PKDD/ECML (2018)
24. Marris, L., Liu, S., Gemp, I., Piliouras, G., Lanctot, M.: Deviation ratings: A general, clone-invariant rating method. International Conference on Learning Representations (2025)
25. MLB Advanced Media, L.P.: Mlb stats api documentation (2026), <https://statsapi.mlb.com/>, accessed: 2026-04-20
26. Moeller, J.: Behind the plate: Are analytics ruining mlb? <https://www.tnonline.com/20260327/behind-the-plate-are-analytics-ruining-mlb/>, accessed: 2026-04-20
27. Murto, F.: Real-time expected possession value estimation in ice hockey. In: Master's Thesis, Tampere University (2024)
28. Omidshafiei, S., Tuyls, K., Czarnecki, W.M., Santos, F.C., Rowland, M., Connor, J., Hennes, D., Muller, P., Pérolat, J., Vylder, B.D., et al.: Navigating the landscape of multiplayer games. *Nature communications* **11**(1), 5603 (2020)
29. Radke, D., Brecht, T., Radke, D.: Identifying completed pass types and improving passing lane models. In: Linköping Hockey Analytics Conference. pp. 71–86 (2022)
30. Radke, D., Larson, K., Brecht, T.: The importance of credo in multiagent learning. The 22nd International Conference on Autonomous Agents and Multiagent Systems (2023)
31. Radke, D., Lu, J., Woloschuk, J., Le, T., Radke, D., Liu, C., Brecht, T.: Analyzing passing metrics in ice hockey using puck and player tracking data. In: Linköping Hockey Analytics Conference. pp. 25–39 (2023)
32. Radke, D., Orchard, A.: Presenting multiagent challenges in team sports analytics. In: Proceedings of the 22nd International Conference on Autonomous Agents and Multiagent Systems (2023)
33. Radke, D., Tilbury, K.: Simulating tracking data to advance sports analytics research. In Proceedings of the 24th Conference of Autonomous Agents and Multiagent Systems (2025)
34. Ratings, W.E.: World football elo ratings. <https://www.eloratings.net/>, accessed: 2026-04-08
35. Schuckers, M., Curro, J.: Total hockey rating (thor): A comprehensive statistical rating of national hockey league forwards and defensemen based upon all on-ice events. In: 7th Annual MIT Sloan Sports Analytics Conference, Boston. pp. 1–10 (2013)
36. Straffin, P.D.: Game theory and strategy, vol. 36. MAA (1993)
37. Štrumbelj, E., Vračar, P.: Simulating a basketball match with a homogeneous markov model and forecasting the outcome. *International Journal of Forecasting* **28**(2), 532–542 (2012)
38. Sutton, R.S., Barto, A.G.: Reinforcement learning: An introduction. MIT press (2018)
39. Sutton, R.S., Barto, A.G., et al.: Introduction to reinforcement learning (1998)
40. Thomas, A.C.: The impact of puck possession and location on ice hockey strategy. *Journal of Quantitative Analysis in Sports* **2**(1) (2006)
41. Tilbury, K., Radke, D.: Adaptive reward sharing to enhance learning in the context of multiagent teams. In: Reinforcement Learning Conference (2025)
42. urto, M.: Towards a real-time possession value framework in ice hockey. In: Linköping Hockey Analytics Conference. pp. 2–10 (2023)
43. Yan, T., Kroer, C., Peysakhovich, A.: Evaluating and rewarding teamwork using cooperative game abstractions. *Advances in Neural Information Processing Systems* **33**, 6925–6935 (2020)

44. Zukiewicz, P.: Analytics ruined the nba. here's how they can save it. <https://medium.com/@zukiewicz.piotr/analytics-ruined-the-nba-heres-how-they-can-save-it-7f24798d56d9>, accessed: 2026-04-24

NHL Home-Ice Offensive Advantage and the Power Play: Exploring Differences Among Game Contexts

Nick Riccardi, Tim Campbell Jr., and Rodney J. Paul

Syracuse University, Syracuse NY 13244, USA

Abstract. This study investigates the home-ice offensive advantage in the National Hockey League (NHL) across game contexts. Using four seasons (2021-2025) of NHL play-by-play data, we estimate home and away team offensive production during even strength and one-man power plays. Additionally, we model power play success as a function of home status and offensive production. Wilcoxon rank-sum and Mann-Whitney U tests show the home team has a significantly higher Corsi rate, Fenwick rate, xG rate, and average xG per shot compared to the away team during even strength stints and in comparison of one-man power plays. Further tests reveal that the home-ice advantage is disproportionately concentrated during the power play. After controlling for shot quantity and quality, home teams maintain roughly 10% higher odds of power play success. Home teams produce 4.27% more Corsi during the power play compared to 2.02% more during even strength and 3.88-3.91% higher average xG per shot compared to 1.55-2.76%.

Keywords: Expected goals · shot quality · crowd effects · penalty discipline.

1 Introduction

The home advantage (HA) phenomenon is well documented across sports [1–3]. This phenomenon stipulates that teams playing in their home venues, *ceteris paribus*, hold a slight advantage over their opponents. The advantage held by home teams is theorized to be most pronounced in indoor sports such as ice hockey and manifests as a performance increase on the offensive end of the game rather than on defense [1]. These considerations motivate the present study.

A difficulty in evaluating offensive performance in a sport like hockey is that scoring events are infrequent. In the past four completed National Hockey League (NHL) seasons, just one team eclipsed the four-goals per-game mark over an entire season. Therefore, the evaluation of offensive performance is best suited through the use of shot volume and shot quality metrics. An important aspect of hockey gameplay that drastically changes offensive production is that of the power play. Teams are awarded power plays—short-term advantages in the number of players on the ice—when their opponent commits a penalty. While the

offensive advantage during a power play is apparent, how it differs between when the home team is on the power play compared to when the away team is on the power play is unknown.

This paper aims to empirically test whether the HA is present in offensive performance in modern NHL play. Using play-by-play data from four recent NHL seasons (2021-2022 to 2024-2025), we derive stints—continuous periods of time where the relative strength between teams remains stable—to test how the HA varies across different game contexts. That is, we compare offensive production at even strength and during one-man power plays. To the best of our knowledge, this is the first study to assess the HA specifically in offensive production metrics (shot quantity and quality) across game contexts, offering a more granular view of where within the structure of a game the HA is most concentrated. Our findings suggest the home-ice offensive advantage is present across all game contexts and metrics. The home team systematically creates more shot attempts and greater scoring chances at even strength and during one-man power plays compared to the away team. Additionally, home teams convert on power plays at a higher rate than away teams, even after controlling for their greater offensive production during the power play. Lastly, the HA ranges from 1.55-6.09% at even strength and is 1.15-2.33% greater during the power play for metrics where the difference is statistically significant.

The remainder of this paper is organized as follows: Section 2 explores the literature on the HA. Section 3 describes the data and methodology utilized. Section 4 provides the results of our analyses and Section 5 concludes the paper.

2 Related Literature

The HA phenomenon has prevalent support across many sports [1–3]. Courneya and Carron [2] provide a framework for the sources of the HA, identifying four game location factors that may benefit the home team: crowd, learning (familiarity with the venue), travel, and rules. We use this framework to organize the relevant literature before turning to how these factors bear on offensive production in the NHL.

Home crowds are a major theorized element of a home team’s advantage, where attendees may provide social support for the team and present a hostile environment for the visiting team. Beyond direct motivational effects, crowds are suggested to influence refereeing decisions in favor of the home team in German football [4] and the National Basketball Association [5]. In the NHL context, home teams receive more power play opportunities than the away team [6, 7], and controlling for home team status, each one-unit increase in power play opportunities has been found to increase win probability by 1.6% [7], a finding that directly links referee behavior to game outcomes.

The COVID-19 pandemic provided an ideal natural experiment to isolate the crowd contribution to the HA. In both the NHL and the Canadian Hockey League (CHL), Guérette et al. [8] find no significant differences in penalties awarded between home and away teams when fans are absent, compared to

a significant home bias in their presence. Similarly, Thrane [9] finds that away teams scored more goals and were awarded more power play opportunities during no-crowd pandemic games in the top Norwegian ice-hockey league. Marek and Vávra [10] find that the HA was dampened in the top professional ice-hockey league in the Czech Republic during the pandemic as well. Comparatively, Losak and Sabel [11] do not find any difference in the HA during the pandemic in Major League Baseball games without fans, which may suggest the crowd effect on the HA operates primarily through influencing refereeing decisions rather than player performance directly.

Beyond referee behavior, crowd effects may also shape player performance through psychological and physiological channels. Carré et al. [12] tracked pre-game hormonal and psychological assessments for players of a CHL team before home and away games, finding that players report greater self-confidence and lower anxiety before home games and measure higher testosterone and cortisol levels before home games compared to away games. These differences in pre-game state plausibly translate into the offensive performance advantages documented in the literature [1, 13].

Familiarity with the home venue may have a particular effect in ice hockey as it relates to geographic regions. As Conetta et al. [14] find, regional weather patterns can impact the playing surface of ice hockey rinks. Areas with higher temperatures and greater humidity can increase the temperature inside the venue, causing wetter, slower ice compared to colder Northern areas.

The necessary travel required from visiting teams can negatively impact performance by introducing fatigue and circadian rhythm disruptions [15]. Circadian disruptions have been shown to impact the number of penalties committed [16]. Therefore, part of the disparity in penalties committed between home and away teams might also be attributed to travel requirements rather than simple crowd influence on referees.

The NHL provides several rules that favor the home team, such as last line changes and last stick position on faceoffs. Multiple studies find that part of the HA effect is absorbed when including faceoff win differentials, which may be influenced through these rules [6, 7]. Rule changes can also affect the HA in less direct ways. Hoffmann et al. [17] replicated Hoffmann et al. [18] to examine the HA across regulation, overtime, and shootouts, extending the earlier work to account for the rule change that moved overtime from 4-on-4 to 3-on-3. While this rule change was not designed to affect the HA, Hoffmann et al. [17] find that superior and evenly-matched home teams are more likely to win in regulation than overtime, while inferior home teams are more likely to win in overtime than regulation. This suggests that reducing player counts serves as an equalizer that partially offsets the HA.

While the factors above help explain why the HA exists, Schwartz and Barsky [1] posit that from a team performance perspective, the HA largely presents as better offensive performance from the home team with an insignificant difference in defensive performance. Indeed, Riccardi [13] finds that top CHL players record a greater number of points when playing at home. At the NHL level, Doyle and

Leard [7] show that Corsi percentage—a shot attempt differential metric—is higher for the home team across their sample, providing aggregate-level evidence that offensive shot generation reflects the HA. However, this aggregate finding does not speak to whether the advantage is uniform across game contexts or whether it is amplified during specific situations such as the power play.

The studies reviewed above establish that home teams win more often, receive more power play opportunities, and show higher aggregate Corsi percentages. However, none directly test whether the offensive production advantage differs in magnitude across game contexts, nor whether the home team’s power play conversion rate exceeds that of the away team and if so, what drives that difference. These gaps motivate the present study.

3 Data and Methodology

We collected play-by-play data from the NHL’s API (api-web.nhle.com) as well as shot-centric data from Evolving-Hockey (evolving-hockey.com) and MoneyPuck (moneypuck.com) for all regular season games over the past four completed NHL seasons (2021-22, 2022-23, 2023-24, 2024-25). The play-by-play data represents our main data source comprised of time-stamped major events in each NHL game such as shots, hits, goals, and penalties. The shot-centric data from Evolving-Hockey (EH) and MoneyPuck (MP) are joined to the play-by-play data to provide expected goal (xG) estimates from each site’s xG model as well as contextual data such as shot distance and angle calculated by EH and rebound and rush dummy variables from MP. The rebound variable takes a value of one if the most recent event before the shot was a previous shot by the same team within the past three seconds and the rush variable takes a value of one if the shot was preceded by an event that took place in a different zone within the past four seconds.

3.1 Deriving Stints for each Game

To accurately assess offensive production by the home and away teams across different strengths, we first derive "stints" within each game. A stint is considered a period of a game in which the relative strength between the two teams does not change. For example, the first stint of each game starts at the initial puck drop and ends when the first power play begins. We do not consider overtime or shootout periods for our analysis. A visual representation of the time elapsed in a game and the derived stints for such game is provided in Figure 1. There are seven distinct stints throughout this game which are labeled in chronological order and colored by the strength state during the stint. Summary statistics for stint durations across strength differences are provided in Table 1. The mean, median, min, and max columns represent the duration of stints in seconds. Just over half of all stints are played at even strength. The home team maintains a small advantage in power play opportunities as 24.1% and 22.2% of stints are home and away one-man power plays, respectively. Two-man advantage stints are rare events and are not included in further analysis.

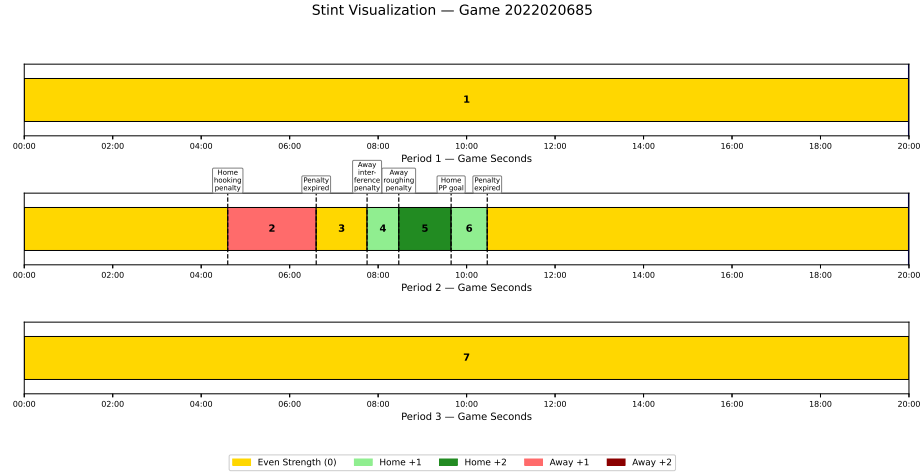


Fig. 1. Visualization of each stint within a game.

Table 1. Summary statistics of stints by HA state.

Strength State	N	% of Stints	Mean (s)	Median (s)	Min (s)	Max (s)
Away +2	452	0.70	48.63	42.0	1.0	120.0
Away +1	14,299	22.20	99.91	120.0	1.0	386.0
Even	33,604	52.18	472.76	337.0	1.0	3600.0
Home +1	15,522	24.10	98.55	120.0	1.0	346.0
Home +2	526	0.82	48.71	45.0	1.0	120.0
All	64,403	100.00	293.35	120.0	1.0	3600.0

3.2 Comparing Offensive Production Between Teams and Game Contexts

Deriving stints for each game allows us to compare offensive production by the home and away teams based on different game contexts and stint durations. Since goals are infrequent events in hockey, better measurements of offensive production are available. In particular, rate statistics may better capture how well a team’s offense is performing. Analysis using rate statistics requires a time duration to be known, motivating the creation of stints within each game.

We adopt the analysis conducted by Paul et al. [19] which compares Corsi rates between teams before and after fights occur. Corsi rates allow us to test for differences in shot volume between the home and away team.

$$\text{Corsi Rate} = \frac{\sum_{i=1}^n 1[\text{event}_i \in \{\text{Goal, Shot on Goal, Missed Shot, Blocked Shot}\}]}{t} \quad (1)$$

Additionally, we compare Fenwick rates, which represent all unblocked shots, which are the sample of shots that xG is calculated for.

$$\text{Fenwick Rate} = \frac{\sum_{i=1}^n 1[\text{event}_i \in \{\text{Goal, Shot on Goal, Missed Shot}\}]}{t} \quad (2)$$

Shot quantity is not fully indicative of team performance, however, as not all shots have equal probability of scoring. Therefore, we leverage the xG model from EH and create a "naive" xG model to compare shot quantity and quality. Similar to Corsi/Fenwick rate, we calculate xG rate across stints by summing xG values for all Fenwick shots by a team during a stint divided by the stint duration. We originally set out to compare team performance using MP's xG model as well, but found their xG values for away team shots during the power play to be problematic. This is apparent in the summary statistics provided in Table 4, where the median xG from the MP model during one-man away power plays is significantly lower compared to the other two models and the standard deviation is significantly higher.

$$\text{xG Rate} = \frac{\sum_{i=1}^n \text{xG}_i \cdot 1[\text{event}_i \in \{\text{Goal, Shot on Goal, Missed Shot}\}]}{t} \quad (3)$$

Creating a Naive xG Model While we leverage the xG model created by EH for offensive production comparisons, the essential research question for this study is whether home teams generate better offense than away teams across different game contexts. EH's model incorporates contextual data such as a home dummy variable and the current strength state. The home dummy should account for any variation in goal probability by team status, *ceteris paribus*, although EH's XGBoost model's gain from this variable is rather small and the effect itself is unknown [20]. While these variables increase the accuracy of predicting goals, they naturally bias the estimates of shot quality. Rather, our naive model only utilizes location and gameplay data and therefore provides a team-independent measurement of shot quality.

We train a logistic regression model to compute xG for the shots in our dataset using location and prior event data. Since our goal with this model is to provide a team-independent measure, we are simply aiming to maximize predictive accuracy. Therefore, we relax concern for multicollinearity in a traditional econometric interpretation of the model. The independent variables used for this model are described in Table 2. We focus on leveraging shot distance and angle, while including contextual information such as whether the shot comes as a rebound, on a rush opportunity, and the type of shot. Many of our chosen variables are interacted with one another to capture how these factors jointly influence scoring probability.

As can be seen in Table 3, which displays the diagnostic measurements of each model, the MP model performs with the highest Area Under the Curve (AUC) while the naive model performs the worst, albeit in close range to both EH and MP. All three models have similar mean xG values, with the Naive model having the lowest average prediction and the EH model having the highest. The

Table 2. Features included in the Naive xG logistic regression model.

Feature Group	Description
Shot distance	Distance from net (ft); quadratic term included
Shot angle	Absolute angle from center of net (degrees)
Distance $\times$ angle	Interaction between distance and angle
Rebound	Binary indicator (shot within 3 seconds of prior event)
Rush	Binary indicator for rush shots
Time since last event	Continuous (capped at 10s); interacted with distance
Rebound $\times$ distance	Interaction capturing close-range rebound danger
Rebound $\times$ angle	Interaction capturing angular rebound danger
Shot type	Categorical dummies (wrist as reference): snap, slap, backhand, tip-in, deflected, wrap-around, poke, bat, between-legs, cradle
<i>Note: Model estimated via L2-regularized logistic regression ($C=1.0$). All continuous features standardized prior to estimation. Trained on 2021–22 through 2023–24 seasons; evaluated on 2024–25.</i>	

strongest correlation of xG values across all shots is between the EH and MP models.

Table 3. Comparison of expected goals (xG) models across 447,708 Fenwick shot attempts.

Model	Mean xG	AUC	Log Loss	Corr. w/ EH	Corr. w/ MP
EH (Evolving Hockey)	0.0725	0.7552	0.2262	—	0.7892
MP (MoneyPuck)	0.0710	0.7864	0.2141	0.7892	—
Naive	0.0702	0.7183	0.2349	0.6759	0.5806

Furthermore, the summary statistics for each xG model’s predictions across game contexts are provided in Table 4. The mean and median values from all three models naturally increase in the power play contexts. Importantly, as previously mentioned, the MP away power play mean, median, and standard deviation deviate significantly from the values for the home power play, whereas these corresponding values for both the Naive and EH models only show a 0.001 difference.

3.3 Empirical Strategy

Our first analysis evaluates whether the HA exists in offensive production during even strength and power plays. We run a series of Wilcoxon rank-sum and Mann-Whitney U tests for Corsi rate, Fenwick rate, xG rate, and average xG per shot comparing offensive production between the home and away teams. We conduct Wilcoxon rank-sum tests for even strength stints since these are dependent observations between the two teams in the same stint. For comparing the offensive production of power plays across home versus away status, we leverage

Table 4. Summary statistics of xG values by game context and shooting team.

Model	Context	<i>n</i>	Mean	Med.	SD
Naive xG	Even Strength (Home)	189,698	0.069	0.052	0.055
	Even Strength (Away)	182,346	0.067	0.050	0.054
	Home PP	33,098	0.082	0.063	0.055
	Away PP	29,582	0.081	0.062	0.056
EH xG	Even Strength (Home)	189,698	0.067	0.037	0.088
	Even Strength (Away)	182,346	0.066	0.035	0.088
	Home PP	33,098	0.105	0.082	0.082
	Away PP	29,582	0.104	0.081	0.081
MoneyPuck xG	Even Strength (Home)	189,698	0.066	0.036	0.087
	Even Strength (Away)	182,346	0.067	0.035	0.090
	Home PP	33,098	0.100	0.077	0.082
	Away PP [†]	29,582	0.087	0.011	0.198

[†] MoneyPuck xG exhibits a notably problematic distribution for away team shots during away one-man power plays (median = 0.011, SD = 0.198), diverging substantially from both Naive xG and EH xG for the same context.

the Mann-Whitney U test since we are not comparing the same stints. For tests comparing average xG per shot, we restrict the even strength stints sample to those which feature at least one Fenwick shot per team (76% of even strength stints) and we restrict the power play stints sample to those that feature at least one Fenwick shot for the power play team (78-79% of power play stints).

We follow this initial analysis with Mann-Whitney U tests at the game-level using games which feature 1) at least one home one-man power play, 2) at least one away one-man power play, and 3) at least one Fenwick shot attempt recorded for each team during their power plays, accounting for roughly 96% of all games in our dataset. These tests at the game-level compare the percentage difference in rates between the home and away teams at even strength compared to percentage difference in rates during power play periods. Therefore, these tests answer whether HA differs during different games contexts.

Lastly, we test whether the HA exists in power play success. A successful power play ends with the power play team scoring a goal while an unsuccessful power play either terminates through the expiration of the penalty or the power play team committing a penalty, ending their advantage. Using "clean" one-man power play stints, we specify logistic regression models using a home team dummy, the power play team's score margin, time remaining in the game, score margin interacted with time remaining, and the total xG produced by the power play team to control for offensive production. Clean power play stints fit the following criteria:

- The penalty is a two-minute minor.
- The stint begins with no other active penalties.
- Another penalty is not committed during the power play by the power killing team.

Clean power plays represent about 83% of all -1 or +1 stints. While these criteria reduce heterogeneity in power play situations, we acknowledge that the types of infractions leading to penalties and the game contexts in which they occur may not be uniform across home and away teams, representing a potential source of selection bias that future work could address more explicitly.

4 Results

The results of our initial comparisons are displayed in Table 5. We present the mean and median values for the home and away teams for each metric within the specific game context. Furthermore, the n column represents the number of stints included in the analysis (where the leftmost value for the power play rows represents the home sample size and the rightmost value the away sample size), the p column represents the p-value of the statistical test, and the $|r|$ column represents the effect size of the test, where larger values indicate a stronger practical difference between groups. All tests are statistically significant for the home team, although effect sizes are quite minimal. These results suggest home teams produce more shots and greater scoring chances during even strength and in comparison of one-man power plays.

Table 5. Home vs. away performance rates (per 60 min) and average xG per shot at even strength and during one-man power plays.

Metric	Comparison	Home	Home	Away	Away	n	p	$ r $
		Mean	Med.	Mean	Med.			
Corsi Rate	Even strength [†]	58.72	55.30	57.43	54.00	33,604	< 0.001	0.027
	Power play [‡]	101.38	75.79	96.06	62.07	15,522/14,299	< 0.001	0.037
Fenwick Rate	Even strength [†]	43.49	38.98	41.84	37.50	33,604	< 0.001	0.033
	Power play [‡]	99.53	69.90	94.11	60.00	15,522/14,299	< 0.001	0.037
Naive xG Rate	Even strength [†]	3.08	2.39	2.90	2.24	33,604	< 0.001	0.045
	Power play [‡]	8.62	5.18	8.03	4.73	15,522/14,299	< 0.001	0.042
Naive xG Avg. [§]	Even strength [†]	0.068	0.065	0.067	0.063	24,918	< 0.001	0.038
	Power play [‡]	0.080	0.073	0.079	0.071	13,333/12,037	< 0.001	0.025
EH xG Rate	Even strength [†]	3.10	2.02	2.91	1.86	33,604	< 0.001	0.047
	Power play [‡]	11.36	6.27	10.47	5.73	15,522/14,299	< 0.001	0.041
EH xG Avg. [§]	Even strength [†]	0.069	0.057	0.067	0.055	24,918	< 0.001	0.039
	Power play [‡]	0.102	0.089	0.100	0.087	13,333/12,037	0.003	0.022

[†] Wilcoxon signed-rank test (paired within stint); [‡] Mann-Whitney U test (un-paired); [§] Restricted to stints with at least one Fenwick shot from each team.

The results of the game-level Mann Whitney U tests are provided in Table 6. In this table, we present the mean and median percentage differences between the home and away teams, from the perspective of the home team, for the given metric and game context. The p and $|r|$ columns report the results of one-sided Mann-Whitney U tests for whether the power play HA is statistically greater

than the even strength HA. All means and medians are positive, suggesting that the home team produces higher Corsi rates, Fenwick rates, xG rates, and average xG per shot across even strength and power play contexts at the game-level. Further, the comparative tests for Corsi rate, average Naive xG, and average EH xG are significant, suggesting the HA effect is 1.15-2.33% greater (median difference) during the power play compared to even strength in these metrics. While the Fenwick rate, Naive xG rate, and EH xG rate are not statistically significant, the percentage difference between the home and away teams is higher in the power play context compared to even strength.

Table 6. Game-level power play vs. even strength HA differentials.

Metric	Comparison	ES Diff	ES Diff	PP Diff	PP Diff	<i>n</i>	<i>p</i>	<i>r</i>
		Mean	Med.	Mean	Med.			
Corsi	Rate per 60	2.00	2.02	5.88	4.27	5,038	0.005	0.030
Fenwick	Rate per 60	3.96	3.28	6.25	4.02	5,038	0.141	0.012
Naive xG	Rate per 60	5.97	5.43	8.05	5.80	5,037	0.169	0.011
	Avg. per shot [§]	2.10	1.55	5.25	3.88	5,037	0.017	0.024
EH xG	Rate per 60	5.92	6.09	8.19	6.58	5,037	0.104	0.015
	Avg. per shot [§]	2.15	2.76	5.45	3.91	5,037	0.043	0.020

All tests are Mann-Whitney U (one-sided: PP diff > ES diff). Diff columns report the percentage difference in home vs. away rates per game. [§] Restricted to games where at least one team generated a Fenwick shot on the power play.

The results of the comparative analyses in Table 6 are visualized in Figure 2. The median percentage difference for the home team is compared between game contexts for each metric, where the left (blue) bar represents even strength and the right (orange) bar represents power plays. The significance level of the Mann-Whitney U tests are denoted above the power play bars. Each bar features a 95% confidence interval for the median percentage difference. These confidence intervals are constructed by drawing 5,000 bootstrapped samples with replacement and displaying the 2.5th and 97.5th percentile medians as the lower and upper bounds, respectively.

The power play success logistic regressions are displayed in Table 7. Standard errors are clustered by game and team-season fixed effects are supplied. The team-season fixed effects are included to absorb baseline variation in team power play performance across seasons, isolating the home/away effect from differences in overall team power play quality. For each model and independent variable, the top value represents the log-odds estimate and the bottom value the clustered standard error. *N* represents the number of clean power play observations included in each model. The log-likelihood measures how well the model fits the observed data, with less negative values indicating better fit. The pseudo *R*² represents the proportion of variance explained relative to a null model, and the AIC provides a measure of relative model fit penalized for complexity, where lower values indicate a better fit.

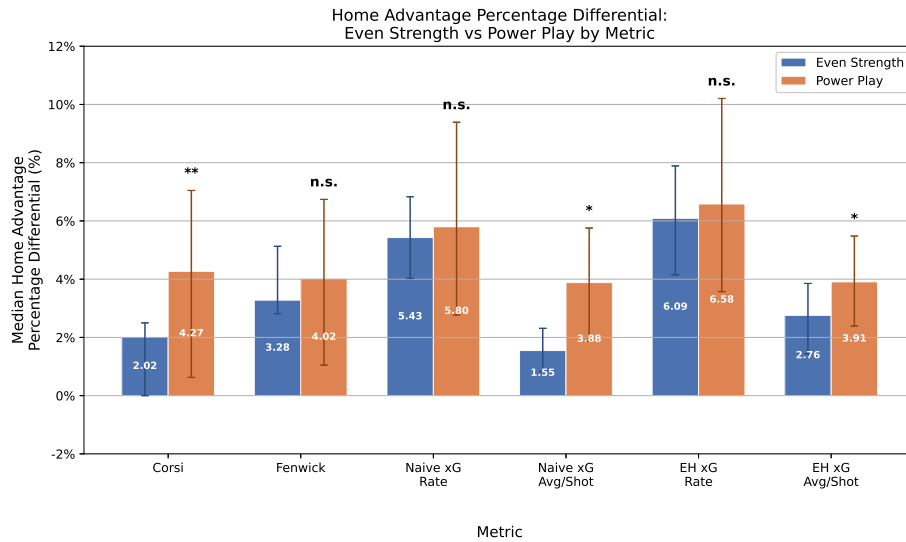


Fig. 2. HA differences across contexts and metrics.

The first model only considers the home dummy, which suggests the home team has about 11% greater odds in converting. The score margin and the amount of time remaining in the game do not appear to have a substantial impact on power play success. Both xG estimates are statistically significant in Models 2 and 3. However, the home dummy's statistical significance also persists. Furthermore, the controlling of offensive production through the xG variables hardly reduces the home dummy's estimate to roughly 10% greater odds in converting.

Table 7. Power play success logistic regression results.

	(1) No xG	(2) EH xG	(3) Naive xG
<i>Intercept</i>	-1.2761*** (0.1667)	-1.5013*** (0.1691)	-1.4917*** (0.1686)
<i>Home PP</i>	0.1059*** (0.0316)	0.0923*** (0.0319)	0.0921*** (0.0318)
<i>Score Margin</i>	-0.0184 (0.0170)	-0.0161 (0.0172)	-0.0127 (0.0172)
<i>Time Remaining</i>	0.0000 (0.0000)	0.0000 (0.0000)	-0.0000 (0.0000)
<i>Score Margin × Time Remaining</i>	0.0000* (0.0000)	0.0000* (0.0000)	0.0000 (0.0000)
<i>EH xG (PP Team)</i>	—	1.0700*** (0.0628)	—
<i>Naive xG (PP Team)</i>	—	—	1.3465*** (0.0843)
<i>Team × Season FE</i>	Yes	Yes	Yes
<i>N</i>	24,837	24,837	24,837
<i>Log-Likelihood</i>	-12759.14	-12629.12	-12649.43
<i>Pseudo R²</i>	0.0105	0.0206	0.0190
<i>AIC</i>	25782.28	25524.24	25564.87

*** $p < 0.01$, ** $p < 0.05$, * $p < 0.10$. Dependent variable: PP success (1 = PP team scored). Clustered standard errors by `game_id` in parentheses. Score margin is measured from the PP team's perspective (positive = leading).

5 Conclusion

In this paper, we aim to evaluate the home advantage in modern National Hockey League games. In particular, we investigate how HA is present in offensive production and how this may differ across game contexts. By deriving stints of gameplay where relative team strength remains stable, we compare shot quantity and quality between home and away teams at even strength and during one-man power plays. Furthermore, we test whether the magnitude of the HA differs between even strength and power play instances. Lastly, we modeled power play success as a function of home status and offensive production.

The results of our analysis support the notion from Schwartz and Barsky [1], which suggests the HA is particularly pronounced in offensive performance. Across the stint-level and game-level comparative analyses, the home team produces significantly more and higher quality shots than the away team. This superior production occurs both during even strength and power play contexts.

Furthermore, the results from Table 6 provide evidence that the HA is about 1-2% greater during the power play than during even strength for Corsi Rate and xG average per shot. Additionally, the HA impact on power play success goes

beyond the increase in shot quality, as accounting for xG on the power play does not absorb the significance of the home dummy in our models. This suggests that the HA in power play conversion operates through channels beyond shot generation. This plausibly includes crowd-driven pressure on opposing penalty killers, heightened player confidence in home settings [12], or officiating tendencies that influence how power play situations unfold [8, 9]. While this study does not consider the impacts of rule advantages for the home team, future work could address the impact of faceoffs and the location of each team's bench relative to the offensive zone as it changes each period.

These results are particularly noteworthy, as power plays provide heightened scoring opportunities, which seemingly would provide away teams an equalizer. However, coupling these results with the finding that home teams are generally awarded more power play opportunities [6, 7], the home team gains a major advantage on the power play. Therefore, while the away team may be inhibited by the home crowd's influence on referee decisions [8, 9] and circadian rhythm disruptions [16], a less-aggressive playing style during road games may help to diminish the HA, particularly given that the power play represents the context in which that disadvantage is most acute.

From a coaching perspective, these results underscore the importance of road penalty discipline. Away teams that limit penalties not only avoid surrendering power play opportunities but deny the home team the game context in which their advantage is most concentrated. Road teams may therefore benefit from prioritizing defensive structure over offensive aggression, accepting a trade-off in shot generation in exchange for reducing the home team's most favorable conditions.

Acknowledgments. We thank the anonymous referees for their helpful comments.

Disclosure of Interests. The authors have no competing interests.

References

1. Schwartz, B., Barsky, S.F.: The home advantage. *Social Forces* **55**(3), 641–661 (1977)
2. Courneya, K.S., Carron, A.V.: The home advantage in sport competitions: A literature review. *Journal of Sport and Exercise Psychology* **14**(1), 13–27 (1992)
3. Nevill, A.M., Holder, R.L.: Home advantage in sport: An overview of studies on the advantage of playing at home. *Sports Medicine* **28**(4), 221–236 (1999)
4. Dohmen, T.J.: The influence of social forces: Evidence from the behavior of football referees. *Economic Inquiry* **46**(3), 411–424 (2008)
5. Price, J., Remer, M., Stone, D.F.: Subperfect game: Profitable biases of NBA referees. *Journal of Economics and Management Strategy* **21**(1), 271–300 (2012)
6. Leard, B., Doyle, J.M.: The effect of home advantage, momentum, and fighting on winning in the National Hockey League. *Journal of Sports Economics* **12**(5), 538–560 (2011)
7. Doyle, J.M., Leard, B.: Variations in home advantage: Evidence from the National Hockey League. *Journal of Quantitative Analysis in Sports* **8**(2) (2012)

8. Guérette, J., Blais, C., Fiset, D.: The absence of fans removes the home advantage associated with penalties called by National Hockey League referees. *PLOS ONE* **16**(8), e0256568 (2021)
9. Thrane, C.: No crowd, no roar, no home advantage: evidence from ice hockey. *European Sport Management Quarterly* **24**(4), 921–933 (2024)
10. Marek, P., Vávra, F.: Home advantage in ice hockey matches without spectators. *Austrian Journal of Statistics* **53**(3), 39–52 (2024)
11. Losak, J.M., Sabel, J.: Baseball home field advantage without fans in the stands. *International Journal of Sport Finance* **16**(3), 148–162 (2021)
12. Carré, J., Muir, C., Belanger, J., Putnam, S.K.: Pre-competition hormonal and psychological levels of elite hockey players: Relationship to the ‘home advantage’. *Physiology and Behavior* **89**(3), 392–398 (2006)
13. Riccardi, N.: Canadian Hockey Leagues game-to-game performance. *Academy of Economics and Finance Journal* **10**, 71–78 (2019)
14. Conetta, C., Weinberg, J., Paul, R.: Impact of ice quality on team and player performance in the NHL. Presented at MIT Sloan Sports Analytics Conference, Cambridge, MA (2016)
15. Roy, J., Forest, G.: Greater circadian disadvantage during evening games for the National Basketball Association (NBA), National Hockey League (NHL) and National Football League (NFL) teams travelling westward. *Journal of Sleep Research* **27**(1), 86–89 (2018)
16. Bourgon, V., Fournier, J., Duval, F.G., Forest, G.: Travel-induced circadian shift between teams influences penalties in the National Hockey League. *Chronobiology International*, 1–5 (2026)
17. Hoffmann, M.D., McEwan, D., Baumeister, R.F., Barnes, J.D., Guerrero, M.D.: Home team (dis)advantage patterns in the National Hockey League: Changes through increased emphasis on individual performance with the 3-on-3 overtime rule. *Perceptual and Motor Skills* **128**(1), 424–438 (2021)
18. Hoffmann, M.D., Loughhead, T.M., Dixon, J.C., Crozier, A.J.: Examining the home advantage in the National Hockey League: Comparisons among regulation, overtime, and the shootout. *Psychology of Sport and Exercise* **28**, 24–30 (2017)
19. Paul, R.J., Riccardi, N., Kuchenbaur, H.: Dropping the gloves, driving the play? Reassessing the role of fighting in modern NHL games. In: Linköping Hockey Analytics Conference, pp. 68–80 (2025)
20. Younggren, J., Younggren, L.: A new expected goals model for predicting goals in the NHL. *Evolving-Hockey* (2021). <https://evolving-hockey.com/blog/a-new-expected-goals-model-for-predicting-goals-in-the-nhl/>. Accessed 16 Apr 2026

A Bayesian Approach to Estimating a National Hockey League Draft Pick Value Chart with Bounds

Yufei Zou, Elliott Kervin, and Michael Schuckers^[0000-0002-4817-7089]

University of North Carolina Charlotte, NC, USA
schuckers AT charlotte DOT edu

Abstract. For leagues like the National Hockey League (NHL), that use a draft to allocate players to team, the valuation of draft picks is an important topic. There is a robust literature of approaches to the creation of draft pick value charts (DPVCs). Our focus will be on a DPVC that values picks based upon the performance of past selected players. In this paper we propose two additions to this literature. First, we introduce two new Bayesian methodologies for estimation of the values in a DPVC. Both of these approaches create a flexible monotonic DPVC. These methodologies, along with a traditional LOESS approach, are compared and evaluated via cross-validation. Second, we generate credible intervals to move beyond point estimation for the evaluation of draft picks. Our methods are applied to three different responses variables: games played, time on ice, point shares and games above replacement. We evaluate these techniques using cross-validation on players selected as part of the NHL drafts from 2009 to 2018 and we find that a monotonic Bayesian B-spline model does best based upon cross-validation performance. For the resulting model, we provide tables (with interval bounds) for each of our three responses.

Keywords: Bayesian monotonic regression · credible intervals · draft pick value chart · sports analytics · ice hockey.

1 Introduction

Leading up to the 2024 NHL draft, the San Jose Sharks traded draft picks 14 and 42 to the Buffalo Sabres for pick 11. With the 11th selection the Sharks chose Sam Dickinson, while the Sabres chose Konsta Helenius and Adam Kleber with the 14th and 42nd picks, respectively. Teams, fans, media and analysts often want to evaluate trades such as this. While hindsight will likely tell us who ‘won’ this trade, in this paper we aim to develop a robust and flexible methodology for the evaluation of such trades when they occur.

Many North American professional sports leagues use a draft to allocate newly eligible players to the teams in those leagues. Among the leagues that use a draft in this manner are Major League Baseball (MLB), the National Basketball Association (NBA), the National Football League (NFL), the National

Hockey League (NHL) and the Professional Women’s Hockey League (PWHL). Generally, a draft consists of several rounds, each team starts with a single selection or pick in each round and teams that finished the previous season with the worst records select earlier in each round. Draft selections are considered assets and can be traded either for other draft picks or for players. Consequently, how to value those assets has been a source of research across sports. This is especially true since many of the collective bargaining agreements (CBAs) between teams set a fixed, often below market rate, contract structure for some of the draft selections. See, for example, the CBA for the NFL [7] or for the NHL [1].

The questions of how to value draft picks has an extensive history. Most authors attribute the first draft pick value chart (DPVC) to the Dallas Cowboys of the NFL in the 1990’s. Since that time, there has been a proliferation of DPVC’s in many sports, e.g. basketball, football, and hockey. Generally, there are three basic approaches to the valuation of draft picks. First picks can be valued based upon historical trade data. Recently, Brill and Wyner [2] added to the literature in this area by comparing the trade market to the probability of gaining an elite player with each draft pick. In the NHL, seminal work by Tulskey [17] summarized some previous work on trades and generated a trade based DPVC. Second, player performance during some portion of their career can be used to define a DPVC. Among the first to consider this approach were Stuart [16] and Schuckers [15] in the NFL and Schuckers [14] in the NHL. Third, the first contract that a player signs after their rookie contract expires can be used to value draft picks. In a classic paper, Massey and Thaler [11] argued that early first round picks were overvalued based upon the subsequent contracts that they signed. A nice summary of the methodological approaches was provided by Lopez [9].

In this paper we will focus on the NHL draft. In addition to the aforementioned articles, there has been no shortage of attempts to create a DPVC for the NHL. Notable research has been done by Schuckers [13], Moreau [12] and Luszczyszyn [10]. One limitation of the current DPVC’s is that they provide a point estimate for each draft pick. From a statistical perspective it is important to acknowledge the variability in the estimated values. In this paper, we propose two new methodological approaches to estimation of a draft pick value chart. In particular, we propose flexible Bayesian monotonic regression methods for evaluation of draft picks. Additionally, these methods create credible intervals for the value of each selection. We approach valuing draft picks using several player performance metrics.

All of the methods for this paper can be found in a github repository: <https://github.com/Yufeiiii/draft-research/tree/main>.

2 Data

Our methodology is driven by an application to the National Hockey League (NHL) draft. To that end, we utilize data on players from the 2009 to 2018 NHL drafts. For each selected player, we aggregate performance over the first seven

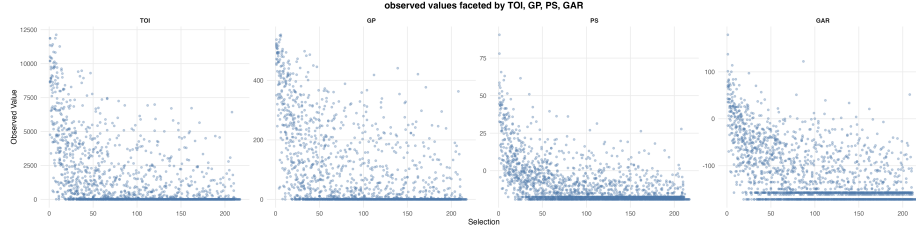


Fig. 1. Observed Raw Values for Response Variables by Draft Selection

NHL seasons following [14] and [13]. Seven seasons was chosen as the number of years that teams have control over drafted players during this period [1]. The latest collective bargaining agreement (CBA) alters this window so that team's have contractual control over players until age 25, so that players drafted after age 18, so-called 'overagers' have smaller windows. We consider four response variables of player performance: Games Played (GP), Time On Ice (TOI), Points Shares (PS) and Games Above Replacement (GAR). Data on GP, TOI, and PS were collected from www.hockey-reference.com [8], GAR data was obtained from evolving-hockey.com [19]. Point Shares is measured in NHL standing points, Kubatko [6], and GAR is measured in goals.

Since our data include the COVID-affected seasons of 2019–2020 and 2020–2021, we applied a scaling adjustment to account for the shortened number of games in those seasons. Specifically, player performance metrics were rescaled to an 82-game equivalent season. For the 2019–2020 season, we used a multiplier of $82/69.8$, and for the 2020–2021 we used a multiplier of $82/56$. We used 69.8 in the first adjustment since this was the average number of games that teams played that season. This adjustment allows performance metrics to be more comparable across seasons with different schedule lengths.

For players who were not in the NHL in a given season, we assigned a value of 0 to their TOI and GP. For PS, we assigned a value of -2.5 for each season a player was not in the NHL and for GAR, we imputed a value of -22.5 . Those values were chosen because they were one unit below the minimum value for players that appeared in NHL games. The logic is that a player playing in the NHL should be more valuable to a team than a player who does not appear in the NHL. This mirrors the difference between 1 GP and 0 GP and 1 minute of TOI and 0 minutes of TOI. Because we want each draft pick to have a positive value in the draft pick chart, we then subtracted the minimum of the cumulative values for each response variable from each cumulative value. That is, $y_i = y_i^{raw} - \min_{i \in \mathcal{I}} y_i^{raw}$ where y_i^{raw} is the sum total of the first seven years of production for player i , $\mathcal{I}$ is the set of all players, and y_i is the adjusted value we will use for modeling below. Clearly the cumulative minimums for GP and TOI are zero. The cumulative minimum for PS and GAR are then -17.5 and -157.5 , respectively.

Figure 1 has scatterplots for each of our response variables plotted against the selection or draft pick.

3 Methods

In this section we discuss methodologies for estimation of the DPVC. Since one of our goals is to have inferential intervals, we focus on statistical methods which produce those bounds. Below we consider three approaches to estimation of the DPVC: a baseline LOESS regression, a monotonic Bayesian B-spline model and a Bayesian nonparametric monotone regression. We will evaluate these methods using a K-fold cross validation approach, for example, James et al. [5].

3.1 Modelling Approaches

LOESS Regression Model For a baseline, we use a locally estimated scatter-plot smoothing regression (LOESS) Regression Model following previous work by Schuckers [13] among others. Let x_i denote draft selection and y_i denote the corresponding player outcome. LOESS estimates the regression function, $f(x_i)$, non-parametrically by fitting a sequence of local weighted polynomial regressions across the range of the predictor. At each target value, x_i , observations with draft selections closer to x_i receive larger weights, while more distant observations receive smaller weights, producing a smooth fitted curve, $\hat{f}(x_i)$. LOESS methodology smooths the relationship between the target/response and predictor(s) in a window that covers a range of values for x_i , the draft pick value for player i , to get a prediction and then the window is moved to estimate the relationship between a response y_i and the draft pick for player i denoted by x_i . LOESS methods produce a smooth estimated DPVC. Implementation of this methodology was done in *R* using the `stats` package with degree 1. We considered a range of spans for the LOESS models and chose final values based upon cross-validation as described below. Smaller spans allow more local flexibility, whereas larger spans produce smoother curves. We selected the span for each outcome using cross-validation described below.

Although LOESS is useful as a flexible benchmark, it does not enforce monotonicity. As a result, the fitted curves may exhibit local non-monotonicity, especially in later portions of the draft where the signal is weak. For this reason, we use LOESS primarily as an unconstrained comparison method for the monotone Bayesian approaches described below.

Monotonic Bayesian B-spline model (B-spline) For this approach, we modeled each response, y_i as a smooth monotone function of draft selection. The response was then transformed to a normalized scale from 1 to 100 using min-max scaling,

$$y_i^* = 1 + 99 \cdot \frac{y_i - y_{\min}}{y_{\max} - y_{\min}},$$

where $y_{\min}$ and $y_{\max}$ are the minimum and maximum observed values of y_i in the full dataset for each response. This normalization was used primarily for computational stability and prior calibration rather than because min-max scaling is intrinsically required for B-spline regression. In our Bayesian monotone spline model, the fitted curve is represented through a sequence of nonnegative coefficient decrements, so the overall scale of the response directly affects the scale of the spline coefficients, the step parameters, and the residual variance. Rescaling each response to the common interval $[1, 100]$ places these quantities on a comparable and numerically stable scale, which improves posterior sampling and makes weakly informative priors easier to specify across outcomes with very different raw units and ranges.

We restricted the modeling range to draft picks 1 through 224, and represented the mean draft-value curve using a cubic B-spline basis. Specifically, if $B_1(x), \dots, B_K(x)$ denote the spline basis functions, then the conditional mean was written as $\mu_i = \sum_{j=1}^K B_j(x_i)\beta_j$, with observation model

$$y_i^* \mid \mu_i, \sigma \sim \mathcal{N}(\mu_i, \sigma^2).$$

The spline basis was constructed with degree 3 and internal knots at draft selections 5, 15, 30, 50, 100, 150, 200. We chose these knot locations to ensure that we had flexibility with the functional form of $f(x_i)$ both during the early portions of the draft (picks less than 30) and later portions (picks after 100).

To enforce a monotonic non-increasing draft-value curve, we imposed a constraint through the spline coefficients. Rather than estimating the coefficients $\beta_1, \dots, \beta_K$ directly, we defined them through a positive starting value and a sequence of nonnegative decrements. Let $\beta_0 \sim \text{HalfNormal}(s_0)$, and let

$$\delta_j \sim \text{HalfNormal}(\tau_j), \quad j = 1, \dots, K-1.$$

We then defined the cumulative decrement process by $\beta_j^{\text{raw}} = \beta_0 - \sum_{\ell=1}^j \delta_\ell$, $j = 1, \dots, K-1$, and set $\beta_j = \max(\beta_j^{\text{raw}}, 0)$. Equivalently, the coefficient vector is constructed as a nonincreasing sequence of nonnegative values. Because each decrement δ_j is constrained to be nonnegative, this parameterization induces a monotone decreasing spline coefficient structure, which in turn yields a decreasing fitted draft curve over the observed draft range.

The prior scale for the observation variance was based on the empirical variability of the normalized training response. Specifically, $\sigma \sim \text{HalfNormal}(s_y)$, where s_y is the sample standard deviation of the normalized outcome in the training data. The prior scale for the initial coefficient was set using the observed upper range of the normalized response, and the decrement parameters were regularized with different scales across the spline basis. The first several decrements were given relatively larger prior scales to allow more flexibility near the top of the draft, while later decrements were assigned smaller scales to encourage smoother behavior in later draft positions. In the implementation, the first four decrement terms used scale $s_y/2$, and the remaining decrement terms used scale $s_y/20$.

Posterior inference was performed in a Bayesian framework using the No-U-Turn Sampler (NUTS) of Hoffman and Gelman [4]. For the final model fit, we used four Markov chains with 1,000 warmup iterations and 2,000 posterior draws, together with a target acceptance probability of 0.90. Convergence was monitored using the potential scale reduction statistic $\hat{R}$, and the maximum $\hat{R}$ across sampled parameters was recorded as a summary diagnostic, cf Gelman and Rubin [3]. The value of $\hat{R}$ was below 1.1 for all posterior runs.

After fitting the model, posterior samples of the spline coefficients were propagated to a grid of draft selections to obtain posterior samples of the mean draft-value curve,

$$\mu(x) = \sum_{j=1}^K B_j(x) \beta_j.$$

For each draft position, we computed the posterior mean and pointwise posterior interval bounds using the 5th and 95th posterior percentiles, producing a 90% credible interval. We chose 90% to ensure that we captured much of the variability in our estimated curves without producing intervals that were overly wide.

Overall, this modeling strategy provides a smooth Bayesian estimate of draft value as a function of selection number, while incorporating the substantive constraint that expected player value should decline as draft position increases. The Bayesian framework further allows straightforward quantification of uncertainty through posterior credible intervals and supports out-of-sample assessment through year-based cross-validation.

Bayesian nonparametric monotone regression (BNMR) As a second approach, we modeled draft value using Bayesian nonparametric monotone regression (BNMR), adapted from Wilson et al. [18]. The original BNMR formulation is built on a Bernstein polynomial basis with a Dirichlet process prior on transformed regression coefficients, and is designed to enforce a monotone increasing regression function. In our application, expected player value should decrease as draft selection number increases, so we modified the coefficient-constraint construction to enforce a *monotone decreasing* draft curve instead.

Let x_i denote the scaled draft selection for player i , with $x_i \in [0, 1]$, and let y_i denote the observed outcome of interest, such as cumulative games played over a fixed post-draft horizon. We assume

$$y_i = f(x_i) + \varepsilon_i, \quad \varepsilon_i \sim \mathcal{N}(0, \sigma^2),$$

where $f(\cdot)$ is an unknown smooth monotone function. Following [18], we represent f using a Bernstein polynomial expansion of order M :

$$f(x) = \sum_{k=0}^M \psi_k(x, M) \beta_k = \Psi(x, M) \beta,$$

where $\psi_k(x, M) = \binom{M}{k} x^k (1-x)^{M-k}$, $\Psi(x, M) = [\psi_0(x, M), \psi_1(x, M), \dots, \psi_M(x, M)]$, and $\boldsymbol{\beta} = (\beta_0, \beta_1, \dots, \beta_M)^\top$. In the original formulation, monotone increase is obtained by requiring $\beta_{k-1} \leq \beta_k$, $k = 1, \dots, M$, which is implemented through the reparameterization $\boldsymbol{\theta} = A\boldsymbol{\beta}$, where A is the lower first-difference matrix

$$A_{\text{inc}} = \begin{bmatrix} 1 & 0 & 0 & \cdots & 0 & 0 \\ -1 & 1 & 0 & \cdots & 0 & 0 \\ 0 & -1 & 1 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & -1 & 1 \end{bmatrix},$$

so that $\theta_0 = \beta_0$, $\theta_k = \beta_k - \beta_{k-1}$, $k = 1, \dots, M$. Under this construction, $\theta_k \geq 0$ for all $k \geq 1$ which implies that f is monotone increasing.

For the draft-curve setting, we reverse the monotonicity direction by negating the non-intercept rows of the constraint matrix while leaving the intercept row unchanged. Thus, we define

$$A_{\text{dec}} = \begin{bmatrix} 1 & 0 & 0 & \cdots & 0 & 0 \\ 1 & -1 & 0 & \cdots & 0 & 0 \\ 0 & 1 & -1 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & 1 & -1 \end{bmatrix},$$

which yields $\theta_0 = \beta_0$, $\theta_k = \beta_{k-1} - \beta_k$, $k = 1, \dots, M$. Under this modified parameterization, the constraints $\theta_k \geq 0$, $k = 1, \dots, M$, imply $\beta_0 \geq \beta_1 \geq \cdots \geq \beta_M$, and, therefore, enforce a monotone decreasing regression function over draft position. The fitted regression function can then be written as

$$f(x) = \Psi(x, M) A_{\text{dec}}^{-1} \boldsymbol{\theta}.$$

As in the original BNMR framework, we assigned a spike-and-cluster prior to the transformed coefficients $\theta_1, \dots, \theta_M$. Specifically, for $j = 1, \dots, M$,

$$Y_i \mid \boldsymbol{\theta}, \sigma^2 \sim \mathcal{N}(\Psi(x_i, M) A_{\text{dec}}^{-1} \boldsymbol{\theta}, \sigma^2), \quad (1)$$

$$\theta_j \mid P, \pi \sim \pi \delta_0 + (1 - \pi) P, \text{ and} \quad (2)$$

$$P \sim \text{DP}(\alpha P_0), \quad (3)$$

where δ_0 denotes a point mass at zero and P_0 is a truncated normal base measure supported on the positive real line. This prior serves two roles. First, the point mass at zero performs variable selection by allowing some transformed coefficients to be exactly zero. Second, the Dirichlet process prior clusters nonzero coefficients, so multiple basis functions may share the same coefficient value. This clustering reduces effective model complexity and allows the fitted curve to adapt toward simpler shapes when supported by the data. In the original BNMR paper, positivity of the nonzero transformed coefficients is what guarantees monotonicity; in our decreasing version, the same positive-support prior is retained, but monotonicity now follows through the modified matrix A_{dec} .

We completed the model with priors on the remaining parameters, including a Gaussian prior on the intercept term θ_0 , a Beta prior on the spike probability π , and a variance prior on σ^2 , following the BNMR formulation of [18]. Posterior inference proceeds through the BNMR Gibbs sampling scheme, which uses a Polya urn representation to integrate over the Dirichlet process clustering structure and update the transformed coefficients subject to positivity constraints. The resulting posterior sample provides inference on both the regression function $f(\cdot)$ and its uncertainty.

For our draft application, the BNMR hyperparameters were selected by leave-one-draft-year-out cross-validation. Let $g = (\phi_0, \sigma_0, M)$ be the hyperparameter combination where ϕ_0 controls the base prior variance, σ_0 controls the prior scale of the intercept, and M is the Bernstein basis order. We evaluated a combination of values for each of these hyperparameters and judged performance on the same cross validation measures described below for comparing across methodologies.

This decreasing-BNMR formulation preserves the main advantages of the original method: smooth monotone estimation, automatic coefficient selection, clustering of basis effects, and a reduced effective parameter space. The only structural change is the sign-reversal in the non-intercept rows of the constraint matrix, which changes the shape restriction from monotone increasing to monotone decreasing and thereby aligns the method with the substantive expectation that draft value declines with draft position.

3.2 Evaluation Metrics

To evaluate how well are our models fit the data we considered two metrics root mean squared error and a weighted root mean squared error. For both of these methods we used a k-fold cross validation approach where we used each draft year to be a fold. Thus we use a leave-one-year-out cross validation (LOYO-CV). That is, to evaluate the performance of each approach we will train each model on nine of the tens seasons of draft performance data, then assess how well the approach predicts the held out season. Since our data cover the 2009 to 2018 National Hockey League drafts, we have 10 folds.

Root Means Squared Error Our first evaluation metric is a root mean squared error (RMSE).

$$\text{RMSE}_d = \sqrt{\frac{1}{n_d} \sum_{i \in V_d} (y_i - \hat{y}_i)^2}.$$

Here y_i and $\hat{y}_i$ are the observed response and predicted response, respectively. V_d is the set of players in the validation/holdout set, d , and n_d is the number of players in that set. For our data $d = 2009, 2010, \dots, 2018$. The RMSE is a traditional way to evaluation performance of a statistical model. The resulting quantity represents the average predictive error for a given model.

Weighted Root Mean Squared Error To emphasize the importance of early draft selections, a weighted version of the RMSE was the other metric we considered.

$$\text{wRMSE}_d = \sqrt{\frac{1}{n_d} \sum_{i \in V_d} w_i (y_i - \hat{y}_i)^2}.$$

The weights, w_i , we chose were the standard deviation of the values for y_i at each pick k . As above, V_d is the set of players in the validation/holdout set, d . In this way we systematically emphasize predicting the draft picks with the highest variability which are generally the earliest selections.

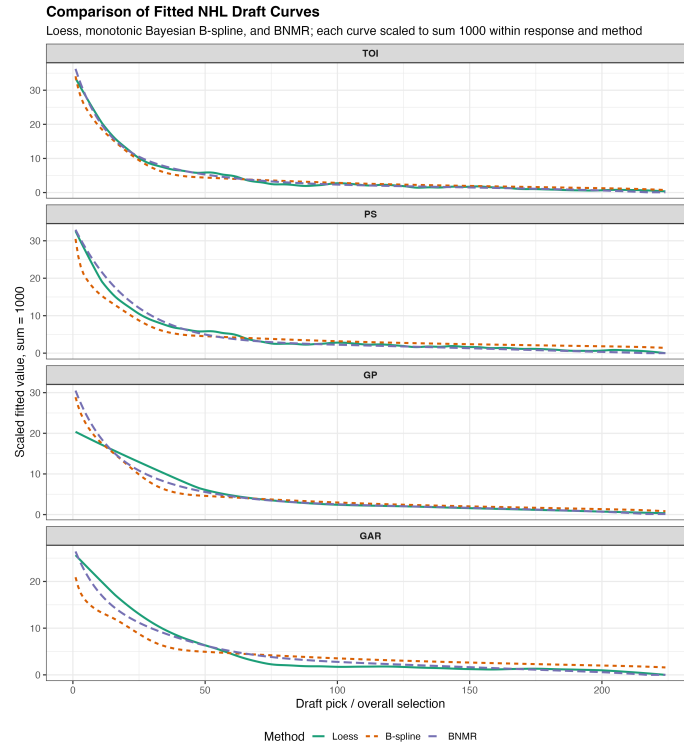


Fig. 2. Comparison of draft value curves across responses and methods. All curves scaled to sum to 1000.

4 Results

In this section, we discuss the results from applying the curve estimation methodologies of the previous sections to data from the 2009 to 2018 National Hockey

League drafts. Each of these methods is applied to data on each player taken in those drafts, their draft selection number, and their cumulative performance in the seven years after they were drafted.

We begin by looking at plots of the DPVC curves on all of the years of data. We next compare the results for our three methodological approaches to the four responses. Figure 2 has plots for the DPVC of the best methodology for each chosen based upon a hyperparameter study using RMSE and wRMSE with greater emphasis on the wRMSE. To have comparable values for the plots, we scaled each predicted DPVC to have the same sum, 1000. For the LOESS models, the best spans were 0.45 for GP, 0.10 for both PS and TOI, and 0.15 for GAR. For BNMR, tuning was performed over the Bernstein basis order and prior hyperparameters using the same LOYO cross-validation framework. For the monotonic Bayesian B-spline model, predictive performance was summarized using both RMSE and wRMSE across the held-out draft classes. details about the final model. For the monotonic Bayesian B-spline model, we used a fixed cubic B-spline specification with internal knots at selections 5, 15, 30, 50, 100, 150, and 200. Monotonicity was enforced through a decreasing coefficient construction based on an initial coefficient and cumulative nonnegative step parameters, with relatively larger prior scales on the first four steps and smaller prior scales thereafter to allow greater flexibility early in the draft and stronger smoothing later. Predictive performance was then summarized under the same LOYO cross-validation framework using both RMSE and wRMSE across held-out draft classes.

From the graphs in Figure 2, we can see some differences across the responses. Starting with GP, we see that curve based upon the LOESS approach differs from the B-spline and the BNMR methods. In particular, the LOESS curve assigns less value to earlier picks than the other two methods. Next, looking at the plots for PS, we see that the three approaches yield roughly the same curve though the B-spline curve is slightly steeper among early draft selections. The LOESS curve for PS exhibits some bumpiness and non-monotonicity for later draft selections. The TOI curves from our three methodologies are the most similar among the response variables. Again the LOESS curve shows some non-smoothness beginning around the 50th selection.

Table 1. Average LOYO prediction error by response and method.

Response	Method	Mean RMSE	Mean wRMSE
GP	LOESS	95.02	933.89
GP	B-spline	93.26	924.83
GP	BNMR	89.91	977.06
PS	LOESS	10.30	34.19
PS	B-spline	9.88	32.87
PS	BNMR	9.84	36.53
TOI	LOESS	1680.36	72373.00
TOI	B-spline	1600.00	69331.26
TOI	BNMR	1591.68	76252.77
GAR	LOESS	47.70	319.03
GAR	B-spline	46.41	311.49
GAR	BNMR	46.16	337.92

Moving to the cross-validation study, we see the results in Table 1. For each response and each curve estimation methodology, we calculated the 10-fold cross validated average RMSE and wRMSE across the holdout years. The methodology with the smallest RMSE and wRMSE are given in **bold**. We seem to see remarkable similarity in these cross validation results. BNMR consistently does the best for RMSE and B-spline consistently outperforms the other methods for wRMSE. LOESS is always the worst for RMSE but had middle performance for wRMSE suggesting that it may do well modeling some of the earlier picks.

The RMSE for GP is smallest for the BNMR approach but the wRMSE has its minimum with the B-spline approach. Similar results hold for PS, TOI and GAR. We also note that these curves exhibit similar performances; no method outperforms another by more than 10%. Figure 3 plots the BNMR methods on the same scale for comparison. The scaling in that figure makes the total of all draft picks the same for each response.

Additionally, for each response, we produced 90% credible intervals using BNMR for each pick and those bounds can be found in Tables 3, 4, 5 and 6. For those tables, we make the first overall pick worth a value of 100. Note that these intervals are for the mean value of the response for a particular pick.

5 Discussion

Quantifying a value to draft picks is a valuable tools for decision makers contemplating possible trades of these picks. A draft pick value chart (DPVC) assigns a currency or value to each pick in the draft and allows for comparable value to be found when trades comprised of picks are evaluated. In the paper above, we have proposed new methodological approaches (B-spline and BNMR) to the estimation of DPVC and evaluated how those methods compare to a LOESS approach. Since our focus was on using player performance to inform our DPVC

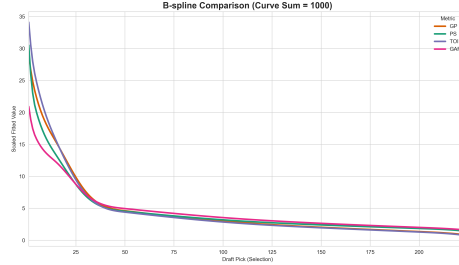


Fig. 3. Comparison of B-Spline Curves for GP, PS, TOI and GAR with identical sums of 1000

Using leave one year out cross-validation, we evaluated the out of sample performance for each of these methods by comparing their average RMSE and average wRMSE.

In addition to the new methodologies, we specifically chose methods that would produce inferential intervals. We believe that acknowledging the uncertainty in the estimated values of a DPVC is something that is important and that has been overlooked in the literature on DPVCs.

The methodology that performed the best in our cross-validation study was the Monotonic Bayesian B-spline model slightly outperforming the other methods on wRMSE and performing similarly on RMSE. We prioritized the wRMSE because it more heavily weights the values early in the draft which are the most valuable to teams. Thus, we suggest using the B-spline approach to DPVC construction. The choice of response is an important one. Looking again at Figure 3, we see that the TOI curve give the largest value to the first pick in the draft while GAR gives the lowest. The PS B-spline curve there seems to have the steepest slope. The three traditional metrics — GP, PS and TOI — have very similar curves, while GAR tends to place relatively more value on picks later in the draft. All of these methods have an substantial change in slope between about the 35th and 40th pick. This change, or ‘elbow’, reflects the data as we can see in Figure 4.

Figure 4 has plots for each of the DPVCs for each of our responses along with 90% credible bounds for those curves. All three of these curves are scaled so that their sums are identical to allow for direct comparisons. The curves in that figure are based upon all ten years of player data and the credible intervals are quite tight to the posterior mean curve and nearly indistinguishable on this figure. Another way to evaluate these curves is to look at the value of the first pick relative to the last pick. (We used 224 for the last pick since that is the number of picks in the 2026 NHL draft.) Returning to Tables 3 through 6, we can see that the ratio of values for the first to last picks are approximately 35, 22, 45 and 13 for

For our model comparisons, we analyzed a subset of reasonable hyperparameters and chose specific values based upon a leave-one-year-out cross-validation.

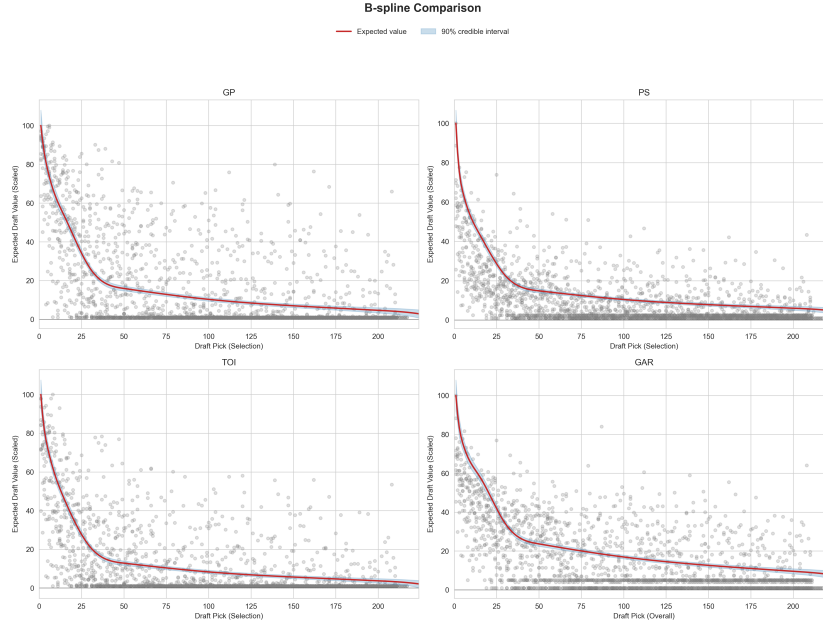


Fig. 4. Plots of B-Spline DPVC curves 90% Credible Intervals and Pick 1 assigned a value of 100.

However, other choices might be worth investigating. One such choice was the spline knot locations for the B-spline approach. We selected 5, 15, 30, 50, 100, 150, and 200 to have more flexibility in the shape of the DPVC among earlier draft picks. The robustness of these choices could be studied. In our cross-validation we leaned on the weighted root mean squared error as our outcome of choice. The weights that were chosen were the standard deviations of each metric at a given pick. One possible future iteration might be to smooth those standard deviations to produce weights. Further, it might be worth considering a different approach to cross validation by taking data from years $t - k, \dots, t - 1$ and predicting year t .

The credible intervals are relatively narrow since we have chosen to make these inferential intervals for the mean value of the DPVC at each pick. Given the amount of data we have for estimation of the curve, the width of those intervals is, perhaps, not surprising. That said, we chose 90% intervals expecting the intervals to be wider. Choosing a wider credible level, for example, 95% or 99%, would certainly be reasonable particularly for the comparison of multiple intervals. A alternative would be to produce posterior predictive intervals for a single pick. These intervals would likely be substantially wider.

One assumption of our approach is that the DVPC should be modeled as a smooth curve without discontinuities. One critique of DPVCs in hockey is that first round picks are given opportunities that other players are not. For instance,

a first round pick may be given additional minutes or games in the NHL to help justify their selection. Consequently, there is a question about whether or not the data supports the hypothesis of a ‘jump’ or discontinuity between the first and second rounds of the NHL draft. An extension of the methodology here could be used to answer that question.

Returning to the 2024 trade between the San Jose Sharks and Buffalo Sabres, we can use the resulting bounds to get a better picture of the values from this trade. The values in Table 2 that are given are the 90% confidence bounds, (L=lower bound , U=upper bound) for the value of each pick using the B-spline approach for a particular response, either GP, PS, TOI or GAR. For the Sabres, we can add the values for the bounds to get a total value interval. For example, looking at GP the interval for the Sharks is 57.59 to 63.42 and the interval for the Sabres is 67.43 to 76.18. Since these intervals don’t overlap, we can say that the Sabres got more value from this trade. We would reach the same conclusions if we used the DPVCs based upon point shares (PS), time on ice (TOI) and games above replacement (GAR). It is important to note that these bounds form inferential intervals for the average or mean value of a given pick, averaged over multiple seasons. As such, they provide more information than single values. However, all DPVCs should be taken as averages (across years). Each draft year will vary from any one specific year and teams value players differently. Thus, it is reasonable for teams to deviate from a DPVC if they believe there is a player available to be drafted whose value differs from that provided by a DPVC.

Table 2. Evaluation of the 2024 Trade between San Jose Sharks and Buffalo Sabres

Metric	Sharks	Sabres
Draft Picks	11	14 + 42
GP	(57.59, 63.42)	$(52.37, 57.29) + (16.06, 18.89) = (67.43, 76.18)$
PS	(51.58, 56.46)	$(45.37, 49.86) + (12.87, 15.35) = (58.24, 65.21)$
TOI	(47.75, 52.42)	$(43.12, 47.09) + (14.91, 17.13) = (58.03, 64.22)$
GAR	(60.81, 66.36)	$(57.22, 62.16) + (23.97, 26.72) = (81.19, 88.88)$

Here we have presented methods for generating a DPVC based upon player performance. It is certainly possible to apply the methodology that we suggest, the Monotonic Bayesian B-spline, on other performance metrics or on other valuation approaches. There are other approaches to estimation of DPVC that ensure monotonicity and may be worth investigation. See [18] for some additional possible approaches. Finally, the methodology here is flexible and can accomodate data from other responses, as well as other sports.

Acknowledgments. We thank the three anonymous reviewers for their comments that added substantially to this manuscript and we acknowledge the support of the School of Data Science at the University of North Carolina at Charlotte.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. National Hockey League Players Association: The NHL/NHLPA Collective Bargaining Agreement , <https://www.nhlpa.com/cba/>
2. Brill, R.S., Wyner, A.J.: Exploring the discrepancy between NFL draft expected value curves and the observed trade market , <https://www.stat.cmu.edu/cmsac/conference/2024/assets/pdf/Brill24.pdf>
3. Gelman, A., Rubin, D.B.: Inference from Iterative Simulation Using Multiple Sequences. *Statistical Science* **7**(4), 457 – 472 (1992). <https://doi.org/10.1214/ss/1177011136>, <https://doi.org/10.1214/ss/1177011136>
4. Hoffman, M.D., Gelman, A.: The no-u-turn sampler: Adaptively setting path lengths in hamiltonian monte carlo. *Journal of Machine Learning Research* **15**(47), 1593–1623 (2014), <http://jmlr.org/papers/v15/hoffman14a.html>
5. James, G., Witten, D., Hastie, T., Tibshirani, R.: *An Introduction to Statistical Learning: with Applications in R*. Springer Texts in Statistics, Springer, New York, second edition edn. (2021), <https://www.statlearning.com/>
6. Kubatko, J.: Calculating Point Shares , https://www.hockey-reference.com/about/point_shares.html
7. National Football League: Contract Language, <https://operations.nfl.com/inside-football-ops/nfl-operations/nfl-free-agency/contract-language/>
8. LLC, S.R.: Hockey-reference.com - hockey statistics and history., <https://www.hockey-reference.com/>
9. Lopez, M.: The making and comparison of draft curves, <https://statsbylopez.com/2016/06/22/the-making-and-comparison-of-draft-curves/>
10. Luszczyszyn, D.: By the numbers: Revisiting the true value of a draft pick, <https://www.nytimes.com/athletic/1798463/2020/06/12/by-the-numbers-revisiting-the-true-value-of-a-draft-pick/>
11. Massey, C., Thaler, R.: The loser’s curse: Overconfidence vs. market efficiency in the National Football League draft. <http://ssrn.com/abstract=697121> (August 2010)
12. Moreau, R.: *Valuation of NHL Draft Picks using Functional Data Analysis*. Master’s thesis, Simon Fraser University (2020)
13. Schuckers, M.E.: Draft Pick Value Chart 2016, <https://www.statsportsconsulting.com/2016/05/26/1658/>
14. Schuckers, M.E.: What’s An NHL Draft Pick Worth? A Value Pick Chart for the National Hockey League, https://myslu.stlawu.edu/~msch/sports/Schuckers_NHL_Draft.pdf
15. Schuckers, M.E.: An Alternative to the NFL Draft Pick Value Chart Based upon Player Performance. *Journal of Quantitative Analysis in Sports* (2011)
16. Stuart, C.: The draft value chart: Right or wrong? <http://www.pro-football-reference.com/blog/?p=527> (2008)
17. Tulskey, E.: NHL draft: What does it cost to trade up?, <https://www.broadstreethockey.com/post/nhl-draft-pick-value-trading-up/>
18. Wilson, A., Tryner, J., L’Orange, C., Volckens, J.: Bayesian nonparametric monotone regression. *Environmetrics* **31**(8), e2642 (2020). <https://doi.org/10.1002/env.2642>, <https://doi.org/10.1002/env.2642>, <https://onlinelibrary.wiley.com/doi/abs/10.1002/env.2642>

19. Younggren, J., Younggren, L.: Evolving hockey, <https://evolving-hockey.com/>

Short papers and community notes

Defining Physicality and Visualizing Skaters' Ability to Play Through Contact

Louis Boulet¹[0009-0000-3432-2604]

¹ Hockey Data Analyst, LB-Hockey, Ottawa ON, Canada
lbhockey.social@gmail.com

Abstract. With physicality being an often-overlooked skill throughout hockey analytics, we look to define what it means for teams and players to be successfully physical in the modern NHL using publicly attainable data. This is done by identifying six high-contact situations encompassing game states or locations that imply physical contests and require an ability to maneuver through direct pressure. Combining these allows us to quantify a skater's ability to play through contact and visualize it with a tool called the "Battle Grid" that is made freely available with this piece. With this, we look to expand the possible analysis in the public sphere by making the dashboard accessible to hockey fans within it. Finally, we explore how this evaluation regarding layers of physicality and the resulting Battle Grid can enrich our case studies with two examples.

Keywords: Hockey Analytics, Physicality, Play Through Contact, High-Contact Situations, Playoff Success, Player Development.

1 Inspiration

The NHL has trended back to putting increased emphasis on heavy play in the last few years. Look no further than the Florida Panthers' recent dominance, or absence of short defensemen selected in the latest drafts [1-2]. Despite this, advancements in the hockey analytics public sphere regarding physical play do not appear to have seen as much growth as other areas of the game. Most of the focus tends to be set on surface-level metrics, such as hits, or certain proxies to estimate physical potential, like player size. But the reality is much deeper than that. It is not exclusively about having an imposing frame or leaving your post to hit everything that moves.

The large majority of hockey audiences do believe that meaningful physicality plays a massive role in both team success and player development. So, our goal with this paper is to add an extra layer to the public's ability to analyze physical play through new metrics, visualization, and case studies. We want to demonstrate that physicality requires awareness, aim, and most importantly, purpose. So how do the best of the best hone this skill, and where do they most apply it?

2 What Are the High-Contact Situations?

On the ice, there are plenty of zones synonymous with grittier play. But the necessity for playing through and preventing progress against contact are equally based on circumstance as much as location, they are not just limited to grinding out pucks along the boards. This means that when we are looking to establish these “high-contact situations”, we are seeking regions and game states that call for an increased level of physical willingness and battling prowess. To help in narrowing down our selection of these situations, we have come up with the following criteria:

1. Physical: implies direct contact with an opposing player or occurs in an often highly contested area of the ice
2. Helpful: encompasses successful plays that improve the team’s position in the game
3. Quantifiable: can be quantified using publicly attainable datasets

For an example of a play this rubric aims to capture, we can turn to one of the league’s best over the last two decades in Sidney Crosby. The Pittsburgh captain is known for his ability to shrug off contact and progress offensive plays despite physical pressure. His assist in game 2 of the 2010 playoffs against Ottawa provides an excellent visual representation of these skills [3]. Looking at this sequence, he recovers a puck along the boards, wards off the opposing player behind the net, then finally escapes his pursuer and passes it off to his defenceman for a goal. This fits all three of our established criteria. It is physical as Crosby protects the puck against an aggressive defender low in the offensive zone. It is helpful since the play begins as a contested battle for the puck in the corner of the ice, and ends with a goal.

It is also quantifiable, as both the initial starting point where Crosby gains possession of the puck gets marked down as a “forecheck recovery”, and the resulting pass is recorded as a “low-to-high shot assist” within the AllThreeZones Project [4]. This is one of the two data sources used for this study. It provides passing information, offensive context, transition metrics, and forechecking data for roughly 300-to-500 games per season dating back to 2020-21 (384 in 2024-25 to be exact). While it may offer a more nuanced look into certain areas of the game, we should remember this portrays only a somewhat-representative sample of the season rather than its true entirety.

The other source at hand is the NHL’s own stats API. Although it remains officially undocumented by the league, a user by the name of “Zmalski” has provided a thorough GitHub page listing its known endpoints [5], which has been used by various analysts for NHL projects in recent years. This API provides access to statistics commonly used in fan discussions such as hits, blocked shots, penalties, shot locations, and more.

With all of this in mind, we have identified six different high-contact situations that isolate the type of plays we are looking for. The situations are as follows:

1. Proactive Contact
2. Transition Contests
3. Forechecking Engagement
4. Forecheck Offence

5. Down-Low Playmaking
6. Net-Front Presence

Although there are more situations we would like to include, this seemed to be the limit with the data available to us. The above situations are actively measured and have attainable data through the public sphere. Including shot pressures (instances when a defensive player forces the puck carrier into taking a quick shot of lesser quality) would add a welcome defensive dimension to this analysis. The same can be said for general puck-protection instances, where a player successfully uses his body and stick positioning to maintain position against direct pressure, to measure even more progress against contact all over the ice. Unfortunately, these do not get tracked at a large enough scale publicly for us. Hockey analytics companies almost certainly provide these types of metrics, but due to their price points they fall out of the “attainable” reach for the average fan.

2.1 Proactive Contact

The first high-contact situation falls more into the “when” bucket rather than the “where”; capturing instances all over the ice where a skater proactively makes contact with an opposing carrier in some way to impede their progress. This can be directly applied to the carrier through a hit and stick-check, or indirectly by obstructing a shooting lane, thus resulting in a blocked shot. Penalties drawn are also included here as they are often the product of a player successfully inviting contact and forcing their opponents’ hand (whether purposefully or not). These are more commonly-found peripherals: secondary counting metrics often present in boxscores that highlight individual actions in a game outside of the usual offensive contributions (goals, assists, shots).

Metrics Chosen: hits, takeaways, blocks, penalties drawn

2.2 Transition Contests

Up next are pressured plays in transition. These are instances where one team either hinders the opposition, or progresses through such a hurdle, while attempting to move the puck up the ice. When against the puck, this takes the form of zone entry targets where the defender rapidly closes space on the carrier to deny possession past the blue line. Meanwhile, examples with the puck are mirrored where the carrier is subject to pressure but successfully maneuvers through it for a successful breakout, with the puck exiting his defensive zone.

Metrics Chosen: entry denials, exits under pressure

2.3 Forechecking Engagement

The following couple of situations have been separated into their own sections but have a similar focus: the forecheck. Here, players chase dumped pucks deep into the

offensive zone with hopes to pressure the opposing team and pin them back. This subsection encapsulates the initial engagement off of forecheck commitments, including puck recoveries, pressures onto opposing defencemen's dump retrieval attempts, and disrupted zone exits. Prototypical checking forwards will especially thrive in these metrics, as their crash-and-bang style of play is well-represented with the common battles here.

Metrics Chosen: forecheck recoveries, forecheck pressures

2.4 Forecheck Offence

If possession is gained off those forechecks, we can turn to more offensive-leaning skills. This transition from applying pressure to capitalizing within the chaos is the main focus of this category, as we isolate for scoring chance creation after forecheck involvements. Whether that is with shots or shot assists, this prioritizes the blending of physical leverage, puck control, and processing speed to reward skaters who possess the ability to quickly shift between skills in a suffocating environment.

Metrics Chosen: shots off the forecheck, shot assists off the forecheck

2.5 Down-Low Playmaking

Maneuvering contact is also important during offensive zone cycles. One of the conclusions from Ryan Stimson's Passing Project back in 2015 is that low zone passes from behind the net generate goals more frequently than other passing sequences [6]. Offensive puck-movement trends follow this finding. The goal-line playmaker has become an important commodity in today's game. Bumper plays relying on touches between the boards and the net headline multiple powerplay strategies. Cross-crease threads through the lower-slot are the endgame for a plethora of set plays. Overall, those passes originating from low in the zone require comfort in one's ability to wade through tight spots consistently, and even more importantly, they improve teams' scoring states.

Metrics Chosen: low-to-high passes, behind-the-net passes

2.6 Net-Front Presence

Lastly is the type of offence most synonymous with grit and smartly navigating through crowded spaces: the net-front game. It is at the core of Chris Kreider, Zach Hyman, and Brady Tkachuk's playing styles. That is thanks to a powerful mix of body positioning and balance, along with advanced situational awareness and puck skills. Their volume-of and finishing-on rebounds, tips, and deflections are what is being used to quantify this. Defencemen are not forgotten here, as a mirrored version of this score is used for them. Their performance is measured by how efficiently they can suppress chances and rebounds in-close. This can be achieved through establishing strong positioning, boxing out the opposition in front of his goalie, or tying up their sticks.

Metrics Chosen (Forwards): rebounds, tips, and deflections

Metrics Chosen (Defencemen): on-ice chances against around the crease (the negative is taken since fewer chances against is better, so that the eventual z-score follows the same format as all other categories of a higher score indicating better performance).

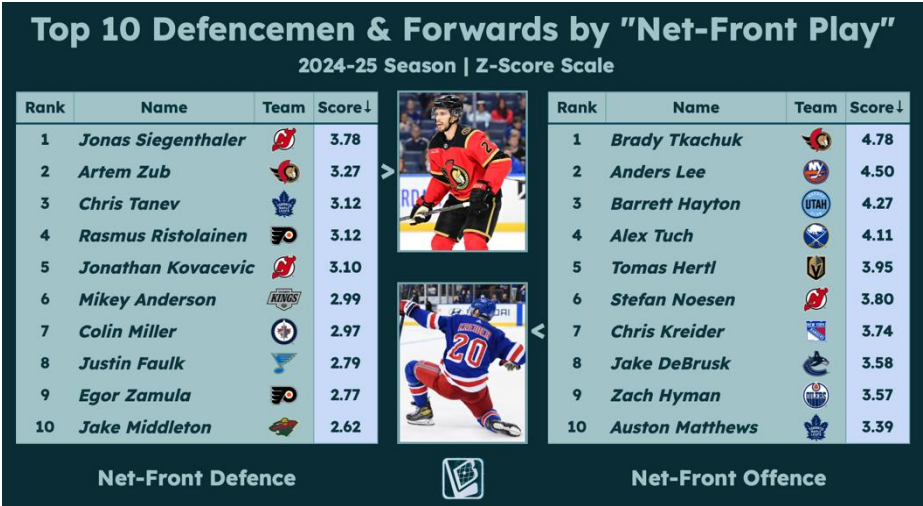


Fig. 1. Above are the top 10 players by position for the 2024-25 season according to their net-front play. Results are shown as z-scores (centralized over the positional league average and scaled over the group's standard deviation). The higher scores for forwards follow what would be expected given that individual metrics are used to compute them, which tend to be more extreme in contrast to the on-ice metrics used for the scores of defencemen here.

3 Building the Battle Grid

Using the tracked event counts at 5v5 outlined in each section, we can measure the rate at which every skater performs them relative to their 5v5 ice time, giving us scores for each of the six high-contact situations. We also set a time on ice threshold for player inclusion in our data of 225 5-on-5 minutes. From testing onto the past five years, this drops approximately the bottom 20% of skaters in a season and weeds out low sample size performances that may carry scores of high variability.

Next is bringing them all together into one encompassing statistic. To accomplish this, we first split the data into forwards and defencemen due to the difference in both their roles and measurements within our data (notably for net-front play). Then, our final situation metrics are converted into z-scores as described in the figure caption above to put them all on one standardized scale.

To ensure we prioritize the skills that are more important for team success and less reliant on random chance, we need to apply weights to our six metrics. We thus derive two weights per position for each of our high-contact situations based on the league-wide trends in these areas: skill repeatability and play potency.

In measuring how repeatable our situations are from year-to-year, we can look at every player's z-score from one season to the next and take the Pearson correlation coefficient. By doing this for every player in our sample dating back to 2020-21 (the first season in our sample), we can obtain a value at forward and defence for each of our six measurements. These will act as our repeatability weights.

For potency, we would like to measure how strongly these physical plays impact success on the ice. There are options for target variables here such as player WAR (wins above replacement), goal differential, and win probability added. The player's on-ice expected goal share (xGF%) has been chosen in this study as it offers a relatively simple way to correlate our physicality metrics with play-driving. It should also be noted that 5v5 xGF% is utilized here since all of our high-contact situations are also measured in this state. We then compute the partial correlation (using ordinary least squares regression residuals) between each metric and this established play-driving target to get our potency weights.

We now have a repeatability factor and a potency factor for each high-contact situation per position (forward and defenceman). Then we multiply the player z-scores in each situation by their positional cohort's repeatability and potency factors, before summing the six weighted z-scores together to arrive at our final one-number metric which we have dubbed "Play Through Contact". It captures skaters who are often willing to engage in high-contact situations and consistently succeed when pressured with or against the puck. We can visualize this new statistic and all of its components with a tool titled the "Battle Grid" [7].

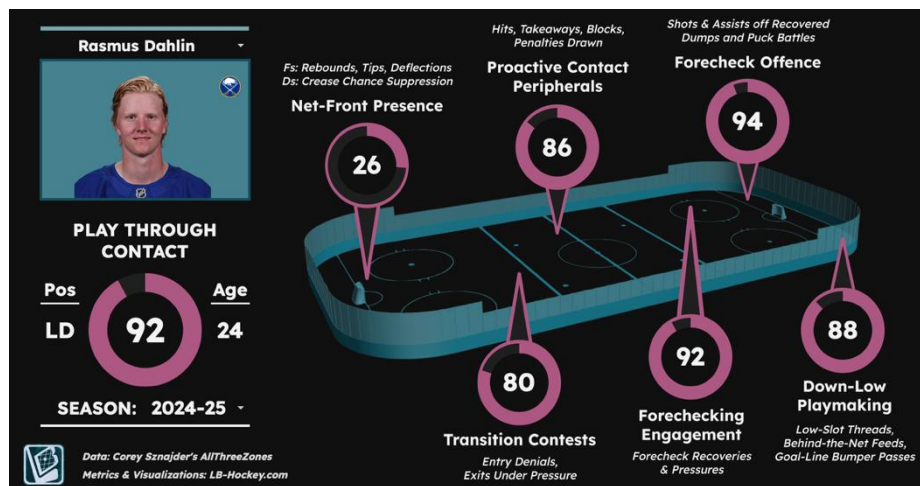


Fig. 2. Rasmus Dahlin's 2024-25 Battle Grid (all scores shown as percentiles among defencemen for that season)

The dashboard outlines each of the high-contact situations we have defined, as they are placed at common locations on the ice where these tend to occur. This further helps interpretation by highlighting frequent battle spots, pressure points, and generally

physical zones on the ice. These categorical measures, along with the final Play Through Contact score, are shown in percentiles by position (forward or defenceman) for the selected player and season, with data since the start of 2020-21 available for selection.

4 How Can It Supplement Our Analysis?

4.1 For Team Success in the Playoffs

The playoffs are generally perceived to host a tougher and more physically-draining game. Open ice is seldom found, loose pucks are constantly contested, and pressure gets applied in the blink of an eye. With this faster rate at which leeway shrinks, the necessity for a skater to compete for every inch of space becomes imperative.

Player size has often been used as a proxy for estimating physical potential and grit, which should result in increased playoff triumph. That intuition is indeed backed up with some solid data, as David Johnson's research shows [8-10]. But looking at recent Cup-winning trends, there may be more to unravel, and that is where our new metrics can help.

As mentioned in the opening paragraph, the recent Panthers' dynasty heavily relies on a suffocating style of play [11-12]. Mainstream media has talked ad nauseam about the intensity of their forecheck, but it does not end there for Florida. Their forward group not only thrives at hounding pucks along the boards, but also creates chances within that chaos unlike any other team's. Meanwhile, their defencemen consistently outplay opposing forechecks, rarely allowing dumps to be recovered, let alone chances off of those. Even in transition and open ice their shutdown pressure is prominent, as the back-to-back Cup winners held the best entry denial and entry chance defence rates league-wide in 2024-25 [13].

While it has grown with the addition of Seth Jones, Florida's defensive corps sported just below median height and weight amongst playoff teams during their runs. This places them as the biggest size anomalies for Cup finalists over the last handful of years (along with the 2022 Avalanche) in Johnson's aforementioned study.

But their physicality shines in what we have built here. More than half their roster ranks at the 82nd percentile or better in our Play Through Contact metric for 2024-25.



Fig. 3. Top 10 Florida Panthers players ranked by this “Play Through Contact” metric for the 2024-25 season. Scores are displayed in percentiles by position (meaning forwards and defence-men are separated).

This includes the league #1 and #3 in Matthew Tkachuk and Aleksander Barkov, 2025 Conn Smythe trophy winner Sam Bennett at 98%, and 8 total players in the league’s top 90. Their entire core fits this mould; from their scorers to the two-way forwards and top-end defensemen.

As for their placement this season, it is not a surprise to see them missing the playoffs. The Panthers led the league in various injury-related metrics for 2025-26, whether by total man games lost, injured cap hit, average minutes of injured players, or lost positive wins above replacement [14]. Being hit with the injury bug to this degree will greatly hinder any roster.

However, with Florida relying so heavily on the forecheck to generate offence like we mentioned earlier, losing Aleksander Barkov for the entire season along with Matthew Tkachuk and Brad Marchand for half of it, strips away skaters who rank very highly in our high-contact situations and are integral to the team’s physical system. With that being said, it is no surprise to see a team falter when they are without the specific pieces who enable the team’s style to succeed.

So already at the team-level, this new data seems to help bridge the stats and eye test further by filling in the gaps of physicality and adding another level of contact play to our realm of analysis.

4.2 In Player Development

As for its purpose in player development, Will Smith’s evolution during his rookie and sophomore outings offers up a compelling case study.

For the first half of his inaugural season, San Jose fans pointed out that the rookie appeared to be playing tentatively: pucks never stayed on his stick very long, his defensive zone involvement was negligible, and he rarely applied pressure to opposing

carriers facing him. All around, he seemed to avoid most physical elements of the game and seldom proactively set out to create opportunities where he would have to rely on applying contact in any way.

In his defence, that tends to be one of the toughest steps in a player's transition to the NHL. The league has a reputation of demanding far more physicality and pace from its players. Compound this speed and grind of the game with the schedule density, and it can inhibit levels of exertion and exhaustion that can require a very long time to adapt for new players.

In Will's case specifically, he was still maximizing almost everything outside of this area. He thrived on plays in open pockets of space and that are less associated with playing through contact. For example, his creation on entries was strong, the passing was already above the league median player, and if he had time to settle and release pucks, his finishing was first-line-worthy [15].

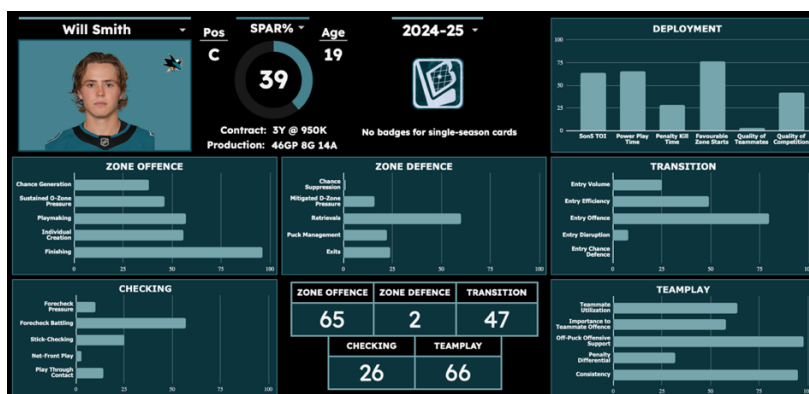


Fig. 4. Will Smith's Single-Season LB-Hockey Card [15] for 2024-25 (as of February 1st, 2025)

His greatest skill came where he would be the least susceptible to immediate pressure: when his teammates held the puck in the offensive zone. Smith was consistently able to get open as a high-danger pass recipient leading to shots; not only creating chances for himself but also acting as a decoy, which provided his linemates even more space for themselves. This culminated in him attaining a top-25 "Off-Puck Offensive Support" rating league-wide [15].

This is especially impressive given it all seemed to be done a fair distance away from the lower slot, as indicated by his very low Net-Front Play. Speaking of which, all of his high-contact situation scores were below the league median for forwards, and even around replacement-level in most. This would corroborate the feeling from fans that the Sharks rookie seemed to panic when time and space shrank around him.

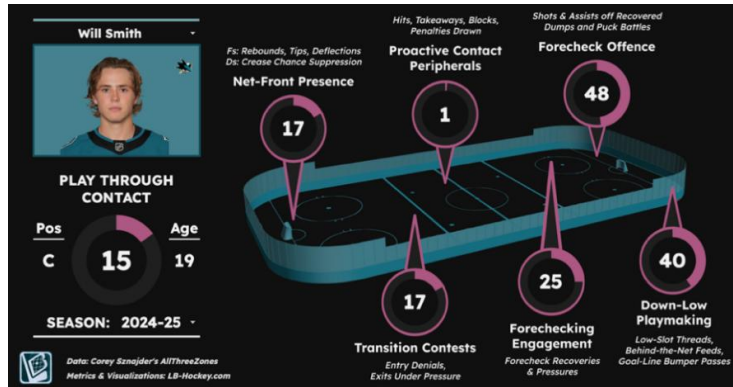


Fig. 5. Will Smith's Battle Grid for the first half of 2024-25 (37/74 games)

Shortly after the calendar turned and Will began the second half of his rookie season, things started starkly improving for him. Offensively, Smith jumped from 6 goals 8 assists in the first half, to 12 and 19 respectively. He carried that momentum into 2025-26 with a similar pace, putting up 24 goals and 35 assists. The story is similar when looking at AllThreeZones's [4] "Microstat Game Score": a metric that weights individual events a skater performs to approximate possession-advancing value. The Sharks rookie sported a 0.52 game score for the first half of his season, which nearly doubled in the second to 0.99, and then remained steady around 1.0 during his sophomore year.

These two measures indicate that Will seemed to have made some big improvements to his game that unlocked greater offensive production and on-ice impact around the halfway mark of his rookie season. So, what could those changes be? His offensive statistics were already strong at that point as we established with Figure 4, whether that was in his passing involvement, finishing ability, scoring sense metrics, or transition offence. An area that does see sparkling improvements in this exact period of time however, is his contact game.

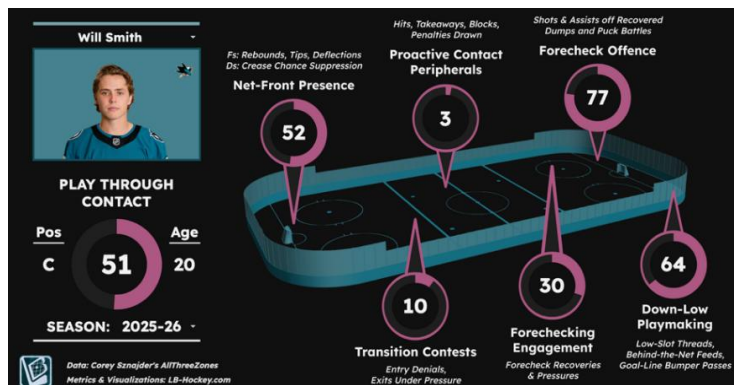


Fig. 6. Will Smith's Battle Grid for 2025-26

Looking at his 2025-26 Battle Grid, Smith has improved in nearly all of the high-contact situations, and taken big leaps in plenty of them. The ones regarding the creation of offence in contested areas see the most notable jumps (Net-Front Presence, Forecheck Offence, Down-Low Playmaking). These three have gone from ranking at a fairly low level, to legitimate strengths in his toolkit.

This data would seem to tell us that halfway through his rookie year, Will started to wade deeper into the offensive zone and invite contact more frequently, allowing him to hold on to pucks a little longer and utilizing his gravity over the opposition's defence. As a playmaker, this opened up more options than previously available when he off-loaded pucks at a far too rapid rate, resulting in a higher Down-Low Playmaking score. On the scoring side, he generated more around the often-contested low-slot, pulling up his Net-Front Presence metric. And lastly, he began to apply this increased physical willingness on the forecheck to create chances for both teammates and himself, raising his Forecheck Offence rank.

Putting this all together, Will Smith engaged and succeeded in physically demanding situations of the game far more frequently. Blending that newly-found involvement with his already-present offensive skills benefitted his point production greatly as we established earlier. But most importantly, the leap in Microstat Game Score indicates that refining these contact-adjacent skills and tendencies have also resulted in a significant increase to his all-around impact on the game. It just goes to show how important adding this extra physical dimension to a skater's arsenal can be for their evolution into an impactful NHLer, no matter their core style of play.

Overall, this new data seems to open the door for a more refined evaluation of player development, especially in terms of acclimating to contact at the NHL level and potential future improvements in that space.

References

1. Berube, J.: The Decline of Small Defensemen in the NHL, <https://hockeyprospect.com/the-decline-of-small-defensemen-in-the-nhl-draft/>, last accessed 2026/03/22
2. Therrien, M.: Draft 2025: not even one defenseman under six feet drafted, <https://dose.ca/2025/06/29/draft-2025-not-even-one-defenseman-under-six-feet-drafted/>, last accessed 2026/03/22
3. Annotated Play Through Contact – Crosby Assist (2010 Playoffs Game 2 vs OTT), <https://www.youtube.com/watch?v=s1lr3g3cM9Y>, last accessed 2026/03/23
4. All-Three-Zones Project Home Page, <https://allthreezones.com>, last accessed 2026/03/23
5. Zmalski's NHL API Documentation GitHub Page, <https://github.com/Zmalski/NHL-API-Reference>, last accessed 2026/03/23
6. Stimson, R.: 2015 Passing Project Data Release: Volume I, <https://www.allaboutthejersey.com/2015/5/20/8627957/2015-passing-project-data-release-volume-i>, last accessed 2026/03/24
7. LB-Hockey • Battle Grid (Google Sheets, "Viz" tab), https://docs.google.com/spreadsheets/d/1U-etDC1LC4nsj6_WH0d2frHrktNicXCjrvV_opC0pdU/, last accessed 2026/03/24

8. Johnson, D.: Player Size and Playoff Success – Part II, <https://hockeyanalysis.com/2023/11/26/player-size-and-playoff-success-part-ii/>, last accessed 2026/03/25
9. Johnson, D.: Playoff Predictions using Defensemen Height and Weight, <https://hockeyanalysis.com/2024/04/17/playoff-predictions-using-defensemen-height-and-weight/>, last accessed 2026/03/25
10. Johnson, D.: Are Bigger Defensemen better at defending in the Playoffs?, <https://hockeyanalysis.com/2025/05/27/are-bigger-defensemen-better-at-defending-in-the-playoffs/>, last accessed 2026/03/25
11. Kezwer, V.: How the Florida Panthers have become a playoff powerhouse, <https://www.sportsnet.ca/nhl/article/how-the-florida-panthers-have-become-a-playoff-powerhouse/>, last accessed 2026/05/20
12. Kelly, M.: Forecheck, backcheck, paycheck. The Florida Panthers, <https://x.com/MikeKellyNHL/status/1935332255917887707>, last accessed 2026/05/20
13. Boulet, L.: The Areas That Will Define Every Round 2 Series, <https://lb-hockey.com/2025/05/09/the-areas-that-will-define-every-round-2-series/>, last accessed 2026/03/25
14. NHLInjuryViz: Final NHL injury summary for the 2025/26 regular season, <https://bsky.app/profile/nhlinjuryviz.bsky.social/post/3mjpprhlqqk2j>, last accessed 2026/06/18
15. LB-Hockey Single-Season Player Cards, <https://lb-hockey.com/player-cards-single-season/>, last accessed 2026/03/26

Net Man-Games Lost as a Coaching and GM Evaluation Metric in the NHL

Quinton J. Krueger^{1,2}[0009-0008-9340-2119], Toby Velte¹, Doug Plocki¹, and Nick Carone¹

¹ Probility AI, Minnesota, USA

² Northwestern University, Evanston, IL, USA
quinton@probability.ai

Abstract. NHL front offices often treat injury burden as bad luck. We argue it is partly organizational, and that the right unit is relative: a team’s man-games lost against its opponents’. On a 17-season corpus with 2024–25 as the focal year, we introduce Net_MGL and its cap-weighted counterpart Net_MGL_Money and evaluate coach and GM tandems on both. The single-season signal is weak, but across 430 full-length team-seasons qualification falls monotonically with burden (odds ratio 0.470 per SD). A crossed random-effects decomposition places the persistent component at the organizational level: about a fifth of the variance in season-long burden is stable franchise and front-office structure, while the head coach’s share is near zero and neither travels between teams nor responds to coaching changes. At the game level a team wins about 1.5 percentage points more often per net man-game of health advantage, an effect that survives within-roster controls yet is small against tandem-quality differences of 6.4 win-rate points. Net_MGL surfaces a health dimension that records obscure and, for long-tenured tandems, a durability screen for hiring.

Keywords: Ice hockey · NHL · Coaching evaluation · Net_MGL · Man-games lost · Injury management · General manager

1 Introduction

Every NHL coach talks about health, yet when coaches are formally evaluated, health rarely appears explicitly: injuries are treated as exogenous shocks, subtracted out before judgment is rendered on tactics. Cap accounting has made that framing hard to sustain. NHL teams tallied 14,566 days on injured reserve in 2022–23, costing approximately \$370 million against the cap [4]; the closest peer-reviewed estimate put league-wide salary lost to injury near \$218 million per year [3]. A team’s injury profile is an organizational outcome, associated with workload, deployment, and acquisition choices, and it deserves to be measured.

Absolute counts are an imperfect target: a healthy team may still lose if its opponents are healthier. What matters is relative health position on a given night, and the NHL’s hard cap makes that framing tractable in a way the NBA’s luxury tax and MLB’s

revenue sharing do not. This paper develops that idea into a pair of metrics and asks what they reveal about the tandems managing rosters.

Related work. The economic burden of NHL injuries is documented at league scale [3,4] and in recent epidemiology of incidence, positional risk, and schedule-density effects [5,6]; Luu et al. [7] showed machine learning outperforms logistic regression for next-season player-injury prediction. That literature treats injury as an outcome to be predicted or costed at the player level. What it does not do, and what this paper does, is treat a team’s season-long relative injury position as an organizational quantity and ask how much of it persists with the franchise, the general manager, or the coach.

2 Net_MGL and Net_MGL_Money

For each NHL game, both teams typically carry players on injured reserve; the imbalance between the two lists is informative. The framing is deliberately distinct from our playoff-forecasting companion work [1,2], which treats man-games lost as one feature among 632; here relative injury position is the quantity of interest. We define the per-game Net_MGL from team t ’s perspective as:

$$Net_MGL_{t,g} = MGL_{opp(t),g} - MGL_{t,g} \quad (1)$$

where $MGL_{t,g}$ is the man-games lost on team t ’s roster entering game g and $opp(t)$ is that game’s opponent. A positive value means team t dressed from a healthier roster; summing across the 82 games gives a seasonal figure of merit.

Net_MGL treats each missing player equally, the wrong aggregation in a hard-cap league. Net_MGL_Money therefore weights each man-game lost by the player’s cap hit, so a team losing a \$10M forward for 20 games against an opponent’s \$900K rookie registers a large negative value even when the count is nearly even.

2.1 Computation

For each game g and team t we count roster players absent for a confirmed injury rather than a deployment decision: IR/LTIR membership is the starting signal, cross-referenced against public injury reports to separate injuries from healthy scratches; day-to-day designations count only when the game is missed. Diminished-capacity play is invisible (Sect. 7.1). Traded players attribute to the roster carrying them at game g , cap-weighted at that date. Playoff games are excluded. Neither metric corrects for travel, schedule, or style; every absence is tagged by position group (Sect. 3.2, Sect. 6). The underlying injury panel is proprietary to Probility AI; an anonymized team-season extract sufficient to reproduce the aggregate results is available to academic researchers on request.

3 Relative Health and Playoff Qualification

Across 2024–25, man-games lost and injury cost correlate strongly ($r = 0.750$, $n = 32$), but the count does not separate qualifiers: Toronto (333 man-games) and Colorado (328) qualified under heavy burdens while the most-injured team, San Jose (338), missed. Net_MGL fares no better on single-season evidence: 9 of 16 playoff teams posted a positive value, as did 8 of 16 non-playoff teams. The relationship is a corpus-level pattern, quantified next.

3.1 Statistical Validation Across the 17-Season Corpus

The quartile analysis uses all 522 team-seasons; the inferential models use the 430 full-length team-seasons after excluding the abbreviated 2012–13, 2019–20, and 2020–21 campaigns.

Quartiles. By season man-games lost (boundaries 157, 220, 292), playoff qualification declines monotonically across quartiles: 64.1%, 58.6%, 54.3%, and 37.4% from lowest to highest burden, with non-overlapping Wilson 95% intervals at the extremes. The highest-injury quartile qualifies at little more than half the rate of the lowest.

Regression. Regressing qualification on standardized man-games lost and injury cost across the 430 team-seasons gives an injury odds ratio of 0.470 per SD ($p < 10^{-4}$) and a cost odds ratio of 1.29 ($p = 0.16$): burden is a meaningful negative predictor, while cost once burden is controlled reads as a roster-quality signal. The relationship holds on the continuous outcome as well: with season fixed effects and team random effects, each one-SD increase in relative health is worth +0.031 in points percentage ($p \approx 5 \times 10^{-15}$), roughly five standings points.

3.2 Where the Burden Persists

Method. The decompositions use crossed random-effects models (REML) with team, head coach, and GM as crossed factors, which shrink each factor toward zero in proportion to its noise, with adjusted R^2 and a within-season label-permutation null reported alongside. Measures are standardized within season; retired players carried on the cap are excluded, the conservative choice, while LTIR absences are retained as genuine injuries (roughly 71% of game-days lost). A game-by-game cap-weighted net CHIP series [4] feeds the cap-weighted variant.

The locus of the signal. An OLS of season man-games lost on coach-and-GM identity indicators explains 53.4% of variance, against 19.4% for team identity, but with nearly two hundred parameters the fit is mostly mechanical. Adjusted R^2 falls to 0.187, and

shuffling coach labels within season yields a null R^2 averaging 0.26 and reaching 0.36, placing the observed coach-only 0.357 at the ceiling of the null.

The crossed random-effects model recovers the genuine signal. The organizational component (team and GM identity, confounded over multi-year regimes) is 19.7%, almost all front-office regime; the head coach is 1.0%; the residual is 79.3% (Table 1; Fig. 1). The residual is a ceiling on chance rather than chance itself, still containing player injury histories and schedule load. Burden is therefore not pure luck: about a fifth of its variance is a stable organizational property, and it belongs to the franchise and front office, not the head coach.

The position-resolved rows hold a finding of their own: the organizational component is largest for defensemen (27.2%) and smallest for goaltenders (9.7%), where the head-coach share, though under a tenth, is highest (6.8%), consistent with the shut-down-selection reading in Sect. 5.

Two design-based tests isolate the coach more cleanly. Across the 46 coaches employed by two or more franchises, the correlation between a coach's standardized burden at one team and at the next is -0.09 , indistinguishable from zero. Replacing a coach within a team produces no shift either: the 149 within-team transitions involving a coaching change show the same year-over-year swing in standardized burden as the 246 that retained the coach (mean absolute change 0.873 versus 0.942; Welch $p = 0.37$). Once prior-season burden is controlled, a change has no detectable effect (coefficient 0.00, $p = 0.96$). Within-pairing persistence is likewise absent: a naive average of per-tandem lag-1 autocorrelations is -0.32 , but series averaging 4.6 seasons carry an expected downward bias near -0.29 , so corrected persistence is essentially zero. We read Net_MGL as a season-level descriptor and an organizational context variable, not a measure of stable per-person skill.

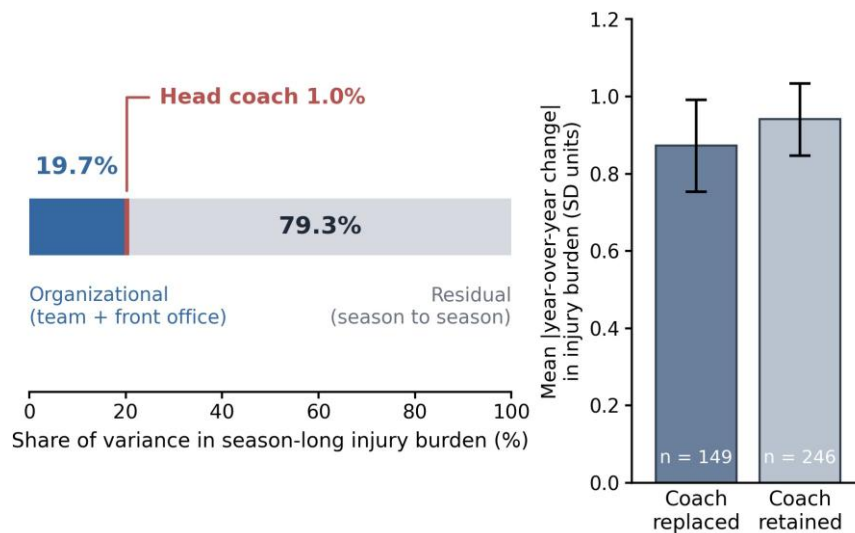


Fig. 1. Left: variance in season-long injury burden partitioned into organizational, head-coach, and residual components (within-season standardized). Right: mean absolute year-over-year change in burden when the coach is replaced versus retained, with 95% CIs.

Table 1. Variance shares (%) in season-long man-games lost from the crossed random-effects model, overall and by position group; rows sum to 100.

Injury burden measured as	Organization (team + GM)	Coach	Residual
All man-games	19.7	1.0	79.3
Forward man-games	19.3	0.6	80.1
Defense man-games	27.2	2.3	70.5
Goaltender man-games	9.7	6.8	83.5

4 Evaluating Coach and GM Tandems

Fig. 2 ranks the 32 tandems active in 2024–25 by cumulative seasonal Net_MGL, count and cap-weighted. Labels follow each franchise’s end-of-period pairing, so a handful of mid- or post-season transitions carry the incoming personnel. The panels correlate positively but imperfectly (Spearman $\rho = 0.72$), and 19 of 32 teams shift at least one quartile under cap weighting, justifying both panels. Net_MGL itself refines raw own-team man-games lost rather than departing from it: its rank correlation with own-team totals is -0.99 , so the larger reranking comes from cap weighting, not the opponent adjustment. That near-unity correlation is a fair caution: as a season-level ranking device, the opponent adjustment adds little beyond the raw count. The value of the relative framing lies elsewhere, in the game-level reading of Sect. 6, where health enters as a two-sided quantity at puck drop, in the cap weighting that drives the larger reranking, and in the organizational decomposition of Sect. 3.2.

Four archetypes emerge. Tampa Bay (+130), Winnipeg (+65), and New Jersey (+51) were **healthy and qualified**. Seattle (+117), the Rangers (+87), and Boston (+68) led the eight teams that **stayed healthy yet missed**, cases that point scrutiny at tactics or roster talent, not the medical staff. Toronto (−139) and Colorado (−125) headlined the seven that **qualified despite heavy deficits**, Edmonton reaching the Cup Final at −11. San Jose (−124), Columbus (−112), and the Islanders (−95) were **injured and missed**. A tandem qualifying from a clear deficit is generating value the record understates; one missing despite an advantage warrants scrutiny of tactics, not the training room.

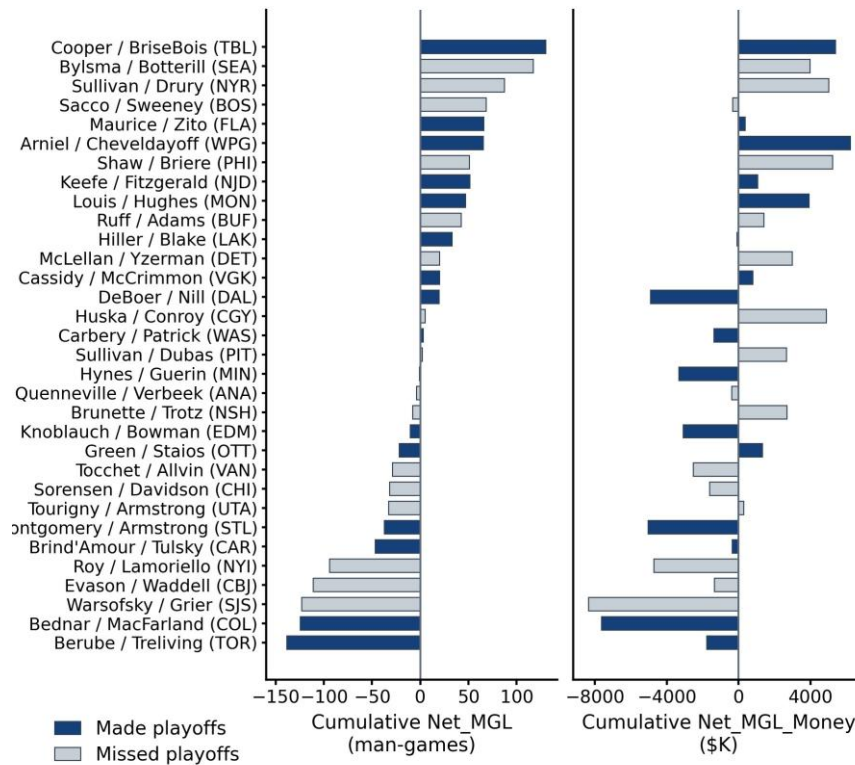


Fig. 2. Coach + GM tandem evaluation for 2024–25. Left: cumulative Net_MGL (man-games). Right: cumulative Net_MGL_Money (cap-weighted \$K). Dark bars = made playoffs; light = missed.

5 Goaltender Injury Trajectories

Season-end totals are not the only signal: the shape of the goaltender-injury curve is itself diagnostic. Across 398 full-length team-seasons, cumulative curves for playoff and non-playoff teams separate from roughly game 15, non-playoff teams finishing at 1.34× the playoff rate. The game-82 gap is significant under a Welch test (17.97 versus 24.14 incidents, $\Delta = 6.17$, 95% CI 1.17–11.17, $p \approx 0.016$) where the mid-season gap is not ($\Delta = 1.86$, $p \approx 0.14$). An alternative reading is shutdown selection: losing teams designate goaltenders to IR more readily than contenders, who push starters through minor injuries. Splitting absences into spells supports that reading as the larger component, long late-season spells carrying a +3.38 man-game gap ($p = 0.015$) against +0.82 from short spells ($p = 0.033$), though the residual gaps are not zero. The finding is correlational; still, a team entering the second half with a healthy primary netminder sits in the league's favorable tail on this dimension.

6 Relative Health and Game Outcomes

If relative health matters, it should matter one game at a time. On a panel of 40,592 team-games, 2008–09 through 2024–25, we compute the same opponent-differenced quantity for each game and model whether the team won; the estimated 54.4% home win rate matches the league baseline, a check that outcomes are recorded correctly.

Relative health predicts winning on a clean monotone gradient (Fig. 3). Even matchups win exactly 50.0%; teams five or more man-games healthier win 59.3%, teams five or more short win 40.7%. A linear probability model gives +1.54 percentage points per net man-game (95% CI 1.36–1.72; $p \approx 1 \times 10^{-61}$). The effect is not team quality in disguise: with team-season fixed effects, comparing the same roster on healthier and sicker nights, it is +1.22 points, positive in all 17 seasons. In aggregate, a sustained two-to-three man-game edge is worth three to four wins, the margin between a playoff berth and the lottery, and the mechanism behind the modest relationship of Sect. 3.1.

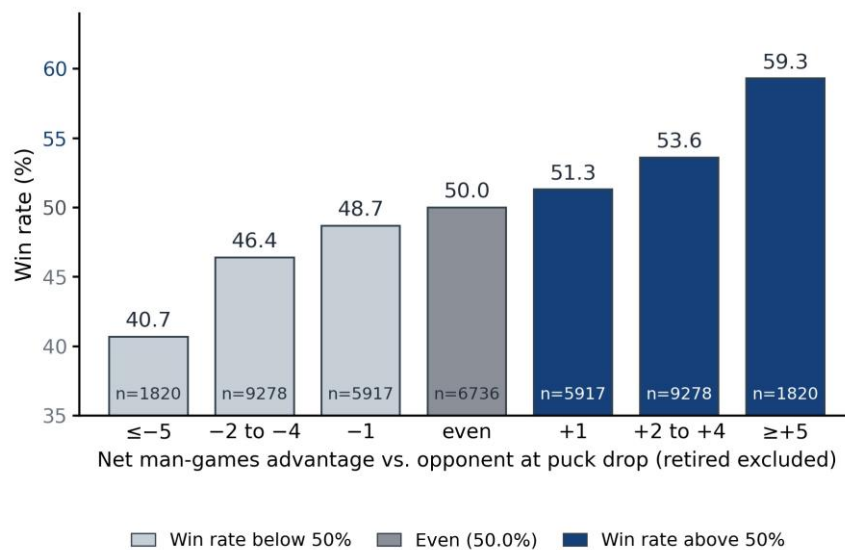


Fig. 3. Win rate by net man-games advantage at puck drop, 40,592 team-games 2008–09 through 2024–25 (retired players excluded). Even matchups win exactly 50.0%; each net man-game of health advantage adds roughly 1.5 percentage points. Navy bars sit above 50%, gray below; axis labels share the coloring.

The game level also settles whether cap weighting predicts outcomes better than the count. It does not: entered together, the count carries the relationship ($\beta = 0.139$, $p \approx 1 \times 10^{-15}$) while cap-weighted cost adds nothing significant ($p = 0.07$); Net_MGL_Money keeps its value for flagging acquisition asymmetries (Sect. 4).

Position resolves the differential further: forwards move win probability +1.55 points per man-game, defensemen +1.47, goaltenders +1.74. A goaltender man-game

is worth about 1.1 times a skater's, not the categorical premium conventional wisdom assigns, so the Sect. 5 separation reads better as shutdown selection than unique cost at the point of play. Own and opponent counts enter equal and opposite, and per-position weighting barely reorders teams ($\rho = 0.999$).

6.1 Tandem Win-Rate Residuals

The game-level model also yields a read the seasonal metric cannot. Holding relative health and home ice fixed, tandems still differ in win rate by about 6.4 percentage points (Fig. 4; SD across the 41 tandems with at least 300 games together); relative health explains essentially none of that spread. The benches above expectation are the recognized elite of the era (Cassidy/Sweeney, Cooper/BriseBois, Trotz/MacLellan, Brind'Amour/Waddell, Hitchcock/Armstrong); the largest under-performers are rebuild-era pairings. Sampling error runs near 2.9 points at the 300-game floor, narrowing toward 1.9 for the longest-tenured pairings, so the genuine between-tandem component is near 5.8 points, a reliability of about 0.8: tandems differ materially, though adjacent rank order is not separable. Relative health is a real input to winning, but a secondary one.

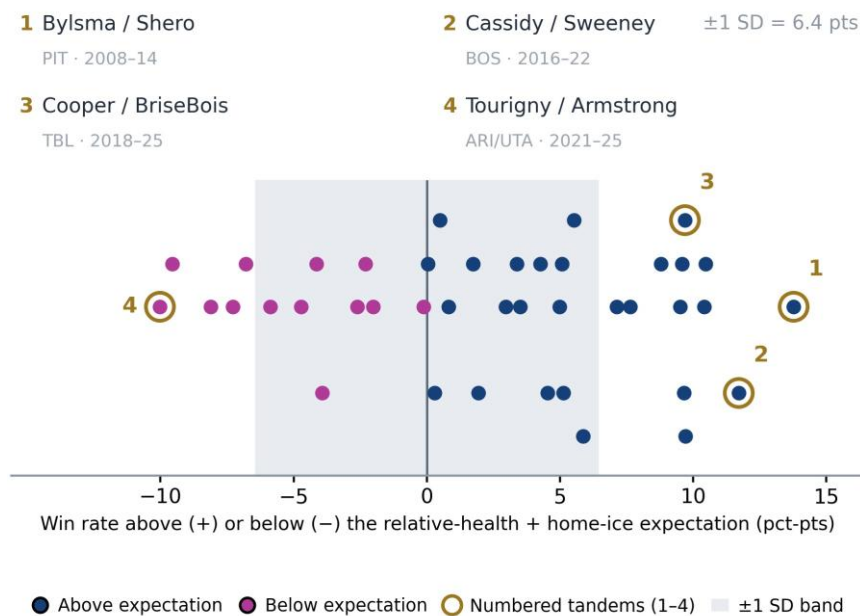


Fig. 4. Win-rate residuals for the 41 coach and GM tandems with at least 300 games together, relative to the expectation set by relative health and home ice. Navy = above expectation, magenta = below; the band marks $\pm 1 \text{ SD}$ (6.4 points).

Fig. 5 widens the view to all 228 coach and GM combinations, bucketed by tenure. The spread contracts as tenure grows, from 9.9 points among single-season pairings to 4.6 among the seven benches with six or more seasons together, so extreme residuals are usually small-sample variance; the lanes stack by tenure, and the funnel narrows visibly from the partial-season lane at the bottom to the long-tenured lane at the top, more games together meaning a tighter spread. Among qualified pairings, Bylsma/Shero hold the highest residual (+13.9) and Tourigny/Armstrong the lowest (−10.1). For an owner, the reading is a durability screen: a positive residual held across six or more seasons is the hardest signal to manufacture with luck, a short-tenure extreme should be discounted heavily, and a negative residual deserves context, since rebuild-by-design pairings read poorly by intent. Paired with Net_MGL_Money, a bench that wins under a heavy cap-weighted burden is showing the depth management that transfers best to a new job.

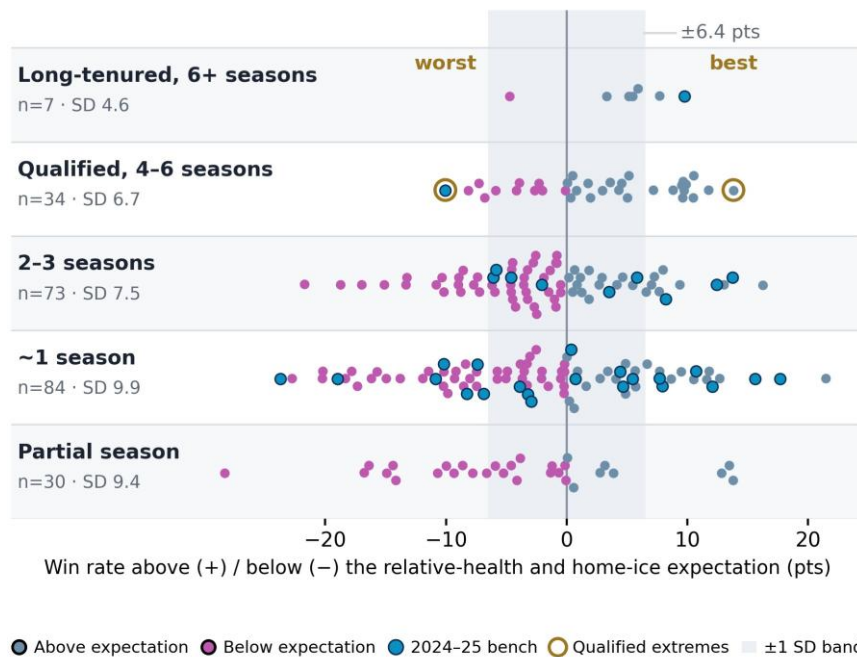


Fig. 5. Win-rate residuals for all 228 coach and GM combinations since 2008–09, in five tenure buckets labeled with sample size and SD. Navy = above expectation, magenta = below; cyan marks 2024–25 benches, gold rings the qualified extremes.

7 Discussion and Practical Use

Net_MGL is a descriptive metric of relative health position, not a complete model of coaching value; three uses follow. In tandem evaluation, the metrics give context to raw records: a tandem qualifying from a negative Net_MGL is generating value the record understates. In selection, candidates can be compared on the historical Net_MGL of teams they have led, with Sect. 3.2 in mind: persistent differences are organizational, a coach's record does not carry to his next team, and one bad season for a stable tandem is weak evidence of a durable problem. In GM portfolio review, Net_MGL_Money flags acquisition asymmetries that counts leave invisible.

The results support a measured conclusion. Relative health predicts qualification only weakly at the single-season level, and a tandem's year-to-year fluctuations look like noise once its franchise baseline is accounted for. What survives is organizational: about a fifth of the variance in burden is stable team-and-front-office structure, while the head coach's share is near zero, does not travel, and does not respond to coaching changes. The game level makes the mechanism concrete: about 1.5 points of win probability per net man-game, real in every season yet dwarfed by tandem-quality differences. The most defensible use of Net_MGL is as a guard against over-reading injury luck in either direction, the contrast Colorado and Seattle draw in 2024–25, and as context that isolates the cases worth a closer look.

7.1 Limitations and Future Work

Three limitations remain. The framework is correlational; the within-team coaching-change comparison in Sect. 3.2 is the closest design-based test the data permit. The metric is binary on availability and misses diminished-capacity play, which would require tracking data. The tandem residuals are era-spanning descriptive quantities of uneven tenure; partial pooling would sharpen them. The 79.3% residual is not yet partitioned, and its largest identifiable piece is likely player-level injury-proneness, which a player-season random effect could isolate from chance; whether goaltender shutdown selection is an organizational skill needs usage data the public record does not carry.

Acknowledgments. The authors thank Probility AI colleagues for discussions that shaped the exploratory-analytics framing from which Net_MGL emerged.

Disclosure of Interests. All authors are affiliated with Probility AI, which builds predictive models of athlete injury and performance for agents, teams, players, and betting markets; its proprietary dataset was used in this study. No other competing interests are declared.

References

1. Krueger, Q.J.: Who's In by Game 15? Multi-Model Per-Game Forecasting of NHL Playoff Qualification. 5th International Ice Hockey Analytics Conference – LINHAC 2026, 36-47 (2026).
2. Krueger, Q.J.: Multi-model NHL playoff prediction system: daily incremental learning with traditional machine learning, recurrent neural networks, transformers, and Bayesian deep learning. M.S. Thesis, Northwestern University (2025)
3. Donaldson, L., Li, B., Cusimano, M.D.: Economic burden of time lost due to injury in NHL hockey players. *Injury Prevention* **20**(5), 347–349 (2014). <https://doi.org/10.1136/injuryprev-2013-041016>
4. NHL Injury Viz (LW3H): team injury breakdowns and game-by-game net CHIP charts, 2008–09 through 2025–26, compiled from CapFriendly and TSN. <https://nhlinjuryviz.blogspot.com/p/index-page.html> (2025), accessed 2025
5. Martin, C., Lieb, A., Tokish, J., et al.: Initial versus Subsequent Injury and Illness and Temporal Trends Among Professional Hockey Players. *International Journal of Sports Physical Therapy* **19**(2), 215–226 (2024). <https://doi.org/10.26603/001c.92309>
6. Ehlen, Q.T., Costello, J.P., Wagner, J.D., et al.: The Effects of Modified Game Schedules on Injury Rates in the National Hockey League. *Cureus* **15**(10), e46898 (2023). <https://doi.org/10.7759/cureus.46898>
7. Luu, B.C., Wright, A.L., Haeberle, H.S., et al.: Machine Learning Outperforms Logistic Regression Analysis to Predict Next-Season NHL Player Injury: An Analysis of 2322 Players From 2007 to 2017. *Orthopaedic Journal of Sports Medicine* **8**(9), 2325967120953404 (2020). <https://doi.org/10.1177/2325967120953404>

Introducing a Gaussian Rink Control-Based Expected Threat Framework to Hockey

Michele Sezgin and Tyler Quinn

Carnegie Mellon University

Abstract. Public hockey analytics often study shot-based measures, but most actions are not shots. We implement an Expected Threat (xT) framework incorporating rink control via two-dimensional Gaussian distributions. Our model, xT_{rc} , evaluates player impact by separating value into interpretable components such as passing and carrying, providing deeper and more granular insights than traditional metrics.

1 Introduction

Mavrik Bourque of the Texas Stars was named the AHL’s Most Valuable Player for the 2023–24 season, but traditional public metrics provide limited insight into exactly what made him so dominant. Public hockey analytics have largely focused on shot-based measures such as expected goals (xG), despite the fact that a large majority of actions in hockey are not shots. Aggregated on-ice xG-based metrics are useful, but they require a large sample size to be accurate, and offer limited context about how scoring chances are created.

To address this limitation, we implement an expected threat (xT) framework that assigns value to nearly all even-strength actions by estimating the value of every game state and assigning credit for changes in that value to the responsible player. The model incorporates rink control in high-danger areas via nested models for xG and probability to shoot. This approach produces an overall measure of player impact, but also separates total value into interpretable components. For a player with high hockey IQ like Bourque who consistently creates value via creative passing (see Figure 1), we can directly observe this value by finding the expected threat value gained from his passes. This framework can be used to gain actionable insight into overall player quality, stylistic profiles, and role fit, supporting decision-making for both coaching staff and front offices.

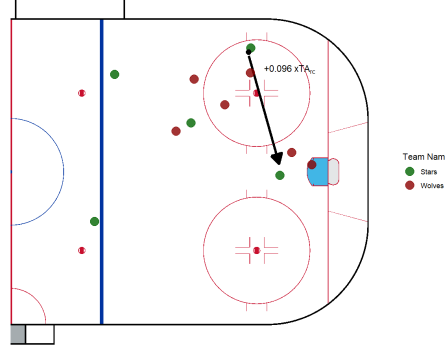


Fig. 1: Example of a high value pass made by Mavrik Bourque

2 Calculating Expected Threat

We start by adapting the algorithm outlined by Singh (2019), originally used for soccer analysis. First, we split the ice into a grid of two hundred 10' x 8.5' boxes (longer in the x, or lengthwise, direction). Then, we create a simple expected goals model by calculating the percentage of shots that go in from each box b ,

$$xG_b = \frac{goals_b}{shots_b}.$$

We also compute how likely a player is to shoot from a given location, xS_b :

$$xS_b = \frac{shots_b}{events_b},$$

where an event can be one of: controlled breakout, pass, dumpin, dumpout, carry, puck protection, or shot. Using these two quantities, we can now calculate the quality of a shooting position in box b , given by

$$xG_b * xS_b.$$

If a player does not shoot, he will move the puck via carrying or passing to another location on the ice (which can be the same as his current location). Note that this probability to move, which we will call m_b , is equal to $1 - xS_b$. In order to evaluate the quality of a box as a position to carry or pass from, we calculate the empirical probability that a player will move from a given box to any other box on the ice and multiply these probabilities by the value of the other boxes. We let $T_{b \rightarrow b'}$ denote the probability that a player will move from box b to box b' , and define the value of box b as the expected threat of box b , xT_b :

$$xT_b = xS_b * xG_b + m_b \sum_{b'=1}^{200} (T_{b \rightarrow b'} * xT_{b'}).$$

The first term is the shooting location value of box b , while the second term is the passing/carrying value of box b , $\sum_{b'=1}^{200} (T_{b \rightarrow b'} * xT_{b'})$, scaled by the probability a player moves instead of shoots, m_b .

Notice that we need to use the xT of other boxes to calculate xT_b . We follow Singh (2019)'s algorithm, initializing $xT_b = 0$ for every box and evaluating this formula iteratively until convergence. With the first iteration we consider only the shooting value of the box, the second we consider "move, then shoot", the third "move, move, shoot", and so on. With the data given, the algorithm converged in eight iterations. As noted by Singh (2019), this leads to an interesting interpretation, allowing us to understand xT as the predicted probability of scoring within the next 8 events, starting from box b .

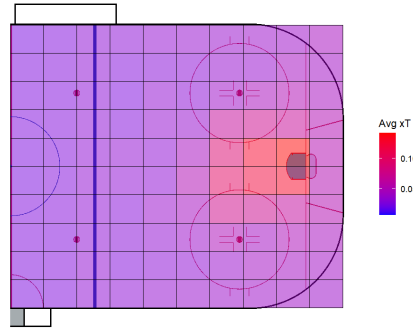


Fig. 2: Expected threat of offensive zone boxes is greatest closest to the net front.

It is clear that the position of other players on the ice affects the puck carrier's decision making. We utilize a two-dimensional Gaussian space control approach outlined by Fernandez et al. (2018) in order to investigate the impact of net-front traffic on xS and xG . Under this framework, each player on the ice has influence over some area of the rink. They have high influence very close to them, and this influence decreases as the point of interest gets farther away. Fernandez et al. (2018) use a two-dimensional Gaussian distribution to approximate a player's influence. Each player's distribution of influence over the rink is computed by starting with a standard two-dimensional Gaussian distribution centered at their current position, which is then stretched and shifted based on their velocity.

We calculated these distributions for each player in each relevant event using the tracking data and summarized them to be relevant inputs to xS and xG models. Relevant events are relevant shooting probability model events, as defined previously. For all relevant events, we drew a triangle between the location of the puck and both posts. We then filtered to only boxes that are at least partially contained within this triangle, exclusive of the box the puck is currently in, in order to calculate net-front rink control. For each box, we integrate each player's rink control distribution, bounded by the box. We then sum all of the

offensive player integrated rink control values across all boxes and subtract the defensive player integrated rink control values across all boxes. Finally, we pass this quantity to the expit function to bound it between zero and one, so it is more interpretable. This final quantity is the offensive triangle rink control for a given event. Figure 3 illustrates our rink control model in action. We choose to visualize the offensive triangle rink control broken down by box, for a more intuitive understanding of how individual players influence team rink control (though the final quantity will just sum across these boxes).

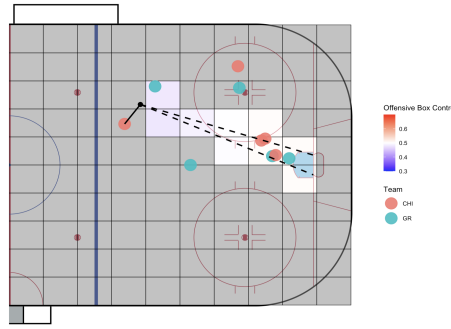


Fig. 3: Example of a shot from Chicago with pressure and traffic in front of the net. The offensive rink control is less than 50% in the boxes immediately in front of the shooter, because a defensive player is pressuring the shooter. The more contested zones directly in front of the net are slightly greater than 50%, indicating a slight offensive advantage in those boxes.

We additionally include the total number of players present in the triangle in the models, because we anticipate that the overall amount of traffic in front of the net also matters for these estimates. Further, we wanted to make the position-based portion of the models fully continuous as opposed to the relatively arbitrary boxes from the simpler models we used before. The xG enhanced with rink control is estimated via a generalized additive model (GAM):

$$xG_{rc} = \text{expit}(\text{te}(x, y, \text{triangle rink control}) + s(\# \text{ of players in triangle})),$$

with tensor product splines capturing the relationship between goal probability and our covariates. We use nonlinear tensor product splines (te) between the x and y coordinates of the shot, along with the aforementioned triangle rink control. We also account for the number of players in the triangle via a smoothing spline (s).

Similarly, xS enhanced with rink control is modeled with a GAM

$$xS_{rc} = \text{expit}(\text{te}(x, y, \text{triangle rink control}) + s(\# \text{ of players in triangle}) + s(xG_{rc})),$$

where we include the shot's estimated xG_{rc} via a nonlinear spline.

With this, the expected threat of the i^{th} event enhanced with rink control can be defined as

$$xT_{rc,i} = xS_{rc,i} * xG_{rc,i} + m_b \sum_{b'=1}^{200} (T_{b \rightarrow b'} * xT_{b'}),$$

where $xS_{rc,i}$ is the rink control-enhanced shot probability of the i^{th} event, and $xG_{rc,i}$ is the rink control-enhanced goal probability of the i^{th} event. The summation term from the original expected threat algorithm remains the same. In order to reduce computational cost, we did not calculate rink control for events not present in shooting or scoring datasets (i.e. events that are not part of the aforementioned relevant shooting probability list), which means we could not make rink control-enhanced xG or xS predictions for these events. Regardless, we still wanted to predict shot and goal probabilities using a fully continuous approach for all events. To do this, we fit the same GAMs to relevant shooting and scoring event data, just without triangle rink control or number of players in the triangle. Because these models only require x and y coordinates, which are present for all expected threat events, this can be used to make predictions for all events. This allows for more fine-grained shooting and scoring probabilities than the original boxed approach.

3 Analysis

We return to Mavrik Bourque to illustrate the effectiveness of xT_{rc} for player evaluation. In our selected example play, Bourque makes an excellent pass from a low danger position near the half wall and successfully finds a teammate in the slot for a one-timer shot (Figure 1). Our implementation of xT_{rc} correctly identifies this as a very valuable pass, as the xT_{rc} when his teammate Logan Stankoven receives the pass is significantly higher than the xT_{rc} when Bourque releases the pass. We can aggregate many events like this over the course of the season to evaluate players. Who is creating value with their passes, and who is not? Which players turn the puck over most often in high xT_{rc} situations, wasting great scoring chances? To answer these questions, we define the expected threat added of the i th event as:

$$xTA_{rc,i} = xT_{rc,i+1} - xT_{rc,i}$$

Where we let i index the events in a given possession sequence, with $i + 1$ being the next event (e.g. the target of a pass receiving said pass). If event i is a shot, the xTA_{rc} value is $xTA_{rc,i} = xG - xT_{rc,i}$. This means that it is possible to earn negative xTA_{rc} from a shot. If a player is in a particularly threatening situation but a bad shooting position, it may be more optimal to move the puck than shoot. If a non-shot action ends a possession sequence, we set the $xT_{rc,i+1}$ to zero, and each loose puck recovery that begins a possession sequence is assigned $xTA_{rc,i} = xT_{rc,i}$.

We chose to aggregate xTA_{rc} over the 2023-24 AHL season by summing each player's xTA_{rc} for each play. We considered other aggregation metrics, but deemed summation to be the best. If a player consistently creates value for his team while playing heavy minutes and fighting off fatigue throughout the long season, he should be rewarded. Figure 4 illustrates the top 10 players by total xTA_{rc} in the dataset.

Rank	Player Name	Team	# Relevant Events	xTA_{rc}
1	Mavrik Bourque	TEX	2351	15.79
2	Justin Sourdif	CLT	2083	15.07
3	Jeff Kubiak	BRI	2034	13.08
4	Rasmus Asplund	CLT	2038	12.69
5	Tanner Laczynski	LV	1830	12.67
6	Marco Kasper	GR	2034	12.10
7	Mason Jobst	RCH	2203	11.75
8	Brian Pinho	BRI	1954	11.62
9	William Trudeau	LAV	2504	11.49
10	Justin Pearson	CLE	1705	11.46

Fig. 4: xTA_{rc} Leaderboard

Note Mavrik Bourque at the top of the leaderboard. While analysis using xTA_{rc} was not necessary to know that he had an excellent season since he led the AHL in points, it is still useful in understanding how he generated value. Figure 5 shows how different types of plays made by the top five xTA_{rc} skaters contribute to their overall 2023-24 AHL xTA_{rc} . We see that Bourque generates much of his positive xT_{rc} via loose puck recoveries, passes, and carries, for example. Additionally, because we only considered 5v5 events in our analysis, our model provides evidence that he was not merely racking up points on the power play. As a sanity check for the appropriateness of xTA_{rc} , we also observe several players in the top 10 xTA_{rc} players that have "graduated" and now play as full-time NHL contributors for their clubs (including Bourque).

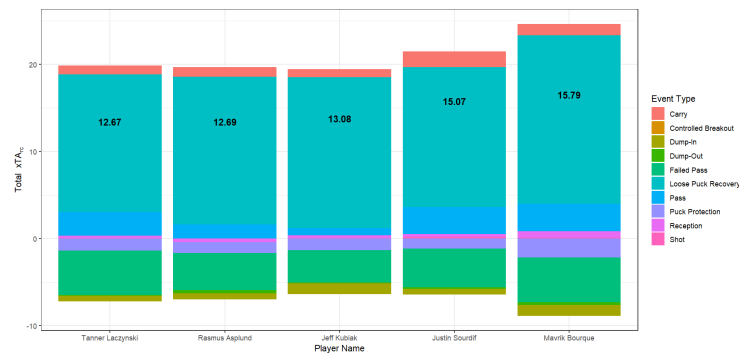


Fig. 5: Top 5 in xTA_{rc} , broken up into components

4 Discussion

Our implementation of expected threat using hockey data is informative for evaluating player quality, stylistic profiles, and role fit in order to support decision-making for both coaching staff and front offices. However, it would be ideal to use more informative models to estimate $T_{b \rightarrow b'}$. For example, one could use the rink control status of the entire ice as a feature in a potentially more complex model, allowing the other players on the ice to influence where the puck carrier is most likely to move the puck. Additionally, turnovers are often much worse in true value than the 0 we assigned— the probability for the offense to score on that possession is in fact 0 with a turnover, but with it comes some probability that it leads to a goal for the opposing team, so in reality the value of a turnover should be negative. We do attempt to somewhat account for this by assigning credit to the defensive player who recovers the puck, giving him $xTA_{rc,i} = xT_i - 0$ for a loose puck recovery at event i . However, given more time, we could look at more specific defensive impacts (e.g. blocked shots reducing the xT of the offense's position). In addition, the current defensive impact is entirely captured in the loose puck retrieval category, but this also includes offensive rebounds and usual puck retrievals. That is also why this category seems disproportionately large in the stacked bar chart above— it is really several categories in one.

Thus, for future work, we would like to look at more categories and event types on both offense and defense in order to find very specific strengths and weaknesses of players. Additionally, we only investigated 5v5 events for the purposes of this project, but it would be interesting to apply this methodology to special teams too. Many of these limitations can be addressed, and we are forced to leave them for future work due to time constraints. Regardless, expected threat is an overall powerful method of valuing non-shot actions that can be used to quantify more complex and specific player impacts than goals or points can alone.

References

- Fernandez, J., Barcelona, F. C., & Bornn, L. (2018). Wide Open Spaces: A statistical technique for measuring space creation in professional soccer.
- Singh, K. (2019). Introducing Expected Threat (xT). Retrieved February 19, 2026, from <https://karun.in/blog/expected-threat.html>