SPORTINDEX

PROOF Protocol — Summary (v2.3)

Pre-registered validation of an NBA pre-match model against the market · November 2026 · v2.3 replaces v2.2 · frozen 14 November 2026

1. Question

Does the deviation between an NBA pre-match model's probability and the market probability at the decision time forecast realised returns at prices that were actually obtainable, net of the bookmaker margin and execution costs, out of sample? A null or negative answer is an acceptable outcome; the protocol is designed to make that answer credible either way.

2. Data

- Historical NBA game results plus historical odds with source timestamps from a commercial provider. Candidate sources are compared in Phase 0 on Pinnacle coverage, timestamp precision and history depth.
- Choice of price source, fixed now: candidate sources are ranked first by the depth of timestamped Pinnacle NBA history, then by timestamp precision. Where two sources both have a decision-time price for the same game and their margin-free probabilities differ by more than 0.01, the game is excluded. The source used for each season is recorded.
- Market reference: the Pinnacle price at the decision time, defined in advance with a fixed maximum staleness. If Pinnacle coverage is inadequate, a median-of-books price, defined in advance, is used instead. The Pinnacle close is recorded for every game and is used in the confirmation test (section 7).
- Price history for the market correction: the correction (section 5) needs only prices and outcomes, not the model, so it is fitted on the longest history available — at least eight seasons where possible, closing prices being sufficient for this purpose — and refitted walk-forward.
- Official NBA injury reports with timestamps, where available, to record the injury-report state at each decision.
- Player-level and officiating data: public box-score data (minutes, line-ups, player contributions), the time-stamped injury and player-status record, rest and travel including back-to-back games, and the officials assigned to each game, each used only as known at the decision time. Each enters the candidate model only under the component rule of section 3. Licensed player-tracking data are not used in this test; they may enter a later cycle with its own pre-registration and its own held-out seasons.
- Markets: moneyline is the primary test. Totals are a second pre-registered test with a separate model, included only if Phase 0 confirms timestamped Pinnacle coverage for totals; the Holm correction then spans both tests. Spreads are reported as a diagnostic only. For totals, when the line moves between the decision time and the close, CLV is computed by converting the closing price to the decision-time line through a fitted distribution of total points; CLV on games where the line did not move is a sensitivity check.
- Phase 0 is a data-validation step with a written coverage report; a measure of how much prices move from the decision time to the close; the measured spread sd(d) of a simple baseline model on the development seasons; the sample-size calculation of section 3; a check of timestamped Pinnacle coverage for NCAA basketball and the NHL at the chosen provider, for the data-extension rule of section 3; and a go / no-go decision before any modelling.

3. Pre-registration

- Hypotheses, metrics, success thresholds, the market-probability method, cost assumptions, model specifications, the bet side and the sample-size target are written down and time-stamped before the test seasons are touched.
- Only two models are evaluated: a simple baseline and one candidate. No model selection on the test seasons.
- Bet side: the home team, for every game. The side the model favours is reported as a diagnostic only. The models are trained and scored on all games.
- Sample size: the target is the size of the true edge, not a share of d. τ is the smallest standard deviation of the true edge, in probability points, worth detecting: on a typical game the true home-win probability differs from the calibrated market probability by about τ (pre-registered, default 0.02). The games needed for 80% power are $N_{80} \approx 1.56 \cdot \mathrm{sd}(d)^2 / \tau^4$, multiplied by 1.27 for the Holm correction, where sd(d) is measured on the development seasons. A noisier model therefore needs more games, not fewer. In simulation the formula is slightly conservative, by about 5–12% for τ of 0.02 to 0.03. If N_{80} exceeds the games available in the test window, the data-extension rule below applies.
- Data-extension rule: if the NBA games in the test window are fewer than N_{80}, then (1) the NBA history is extended to earlier seasons with timestamped Pinnacle prices (2018-19 and before; 2019-20 stays excluded), the two earliest seasons becoming the development period and the held-out seasons unchanged; (2) if still short, $\tau = 0.03$ is applied and N_{80} recomputed — the only permitted change to τ, applied mechanically; (3) if still short, at most one additional league is added as a separate pre-registered test with its own model, chosen from the Phase 0 coverage report before any test season is evaluated; (4) if still short, the result is no-go. Leagues are never pooled into one regression, and the Holm correction spans all tests.
- Component rule: a transform, correction or feature enters the frozen specification only if it improves out-of-sample log loss on the development seasons. Nothing is kept on prior belief alone.

4. Evaluation design

- Six seasons, 2020-21 to 2025-26; 2019-20 is excluded (interrupted and completed in a closed bubble). Development: 2020-21 and 2021-22. Primary test window, walk-forward: 2022-23 to 2025-26 — fit on all past seasons, predict the next. Held out in

code until the method is frozen, then scored once: 2024-25 and 2025-26. Earlier seasons' results, which need no odds, may be used for training.

- All features computed as-of the decision time; no information after the decision timestamp. Tests in the code fail on any later information.
- Decisions taken at a fixed pre-close time, at a price that was actually available at that time.
- All games in the sample are scored, not only those the model would bet on.

5. Market probabilities (removing the margin)

- Probabilities are normalised so that they sum to one.
- A favourite-longshot correction is always applied as a smooth calibration: a three-parameter logistic regression of outcomes on the log-odds x of the normalised probability and its cube, logit $p^* = a + b \cdot x + c \cdot x^3$. Its form is fixed in the pre-registration; its parameters are fitted only on the price history of section 2 before the season being predicted and refitted walk-forward. A smooth fit uses all games at once instead of about 950 per decile; the cubic term lets it follow a bias concentrated in the tails (strong favourites and long shots).
- Why three parameters: in the simulation study of the v2.2 test (1,000 synthetic histories per setting, eight seasons of price history), the two-parameter calibration passed a model with no information 11.4% of the time under a tail-shaped bias, against a 5% target; the three-parameter form passed 6.8%, with power unchanged (68.2% against 68.7% at sd(d) = 0.03).
- Diagnostic only: the decile check of normalised probability against outcomes is published but not used. Sensitivity checks: the two-parameter logistic calibration, the decile correction and an isotonic calibration. The Shin, power and goto_conversion methods are not used.

6. Metrics

- Primary, all games: for each game, d = p_model − p_market at the decision time, and r = the realised return of a one-unit bet on the home team at the price obtainable at that time, net of costs. The primary test is a regression of r on d across all games, with the calibrated market log-odds of the home team as a control; the hypothesis is a positive slope on d. Standard errors are clustered by game day (CR1, t with G−1 degrees of freedom). The slope must also be significant under a wild cluster bootstrap by game day (Rademacher weights, null imposed, 999 draws); both are required (section 7).
- Confirmation at the close: the same regression with r computed at the Pinnacle closing price. A positive slope at the decision-time price that does not hold at the close indicates information that reached the model before it reached the market, not model skill.
- Secondary: realised ROI of the pre-registered betting rule (bet when d exceeds a fixed threshold, default 0.03), with a bootstrap interval; log loss and Brier score against the market at the decision time and at the close; net CLV, E = p_close × O_obtained − 1; the result by season and with each season left out in turn; the result split by games with and without late injury-status changes after the decision; for reference only, model edge and realised return by odds band of the side at the decision time — below 1.50, 1.50–1.91, 1.91–2.50 and above 2.50. The odds-band results are not a condition.

7. Success criteria

- Pass: the slope on the decision-time return is positive and significant at the one-sided 5% level, under both the CR1 t-test and the wild cluster bootstrap, after Holm correction across the two models (and, if totals are included, across the two markets); the slope at the close is positive with one-sided $p < 0.10$; the sign holds under the sensitivity calibrations, without the correction, and with each test season left out in turn.
- Fail: the slope estimate at the decision-time price is zero or negative.
- Inconclusive: any other result, including a positive but not significant slope, and a pass at the decision-time price that is not confirmed at the close, which is reported with the note "probable information leakage".
- The ROI of the betting rule is published alongside every result but is not a condition. The sample-size calculation of section 3 states the smallest effect the test can detect.
- If the result is not a pass, it is reported as such and the project does not proceed to capital.

8. After the backtest

- Forward paper trading with time-stamped decisions logged before the close, scored with the same code as the backtest.
- Capital gate, fixed now: no capital is committed unless the backtest result is a pass and forward paper trading under the frozen model and betting rule, at prices obtainable at the decision time, meets all of the following, judged once at least 300 bets over at least three months of the regular season are recorded: (1) mean net CLV positive with one-sided $p < 0.10$, standard errors clustered by game day; (2) the closing-line move — the market probability of the home side at the close minus at the decision time — regressed on d over all paper-traded games has a positive slope with one-sided $p < 0.10$, clustered by game day; (3) realised return net of costs minus net CLV, over the bets, is not significantly negative (one-sided t above −1.645, clustered by game day). Rule (3) is checked again at the end of the regular season, and capital is withdrawn if it fails then. Realised return is not a pass condition: over about 300 bets its noise is some fifty times the edge being tested. In a simulation of 2,000 paper-trading seasons per case, a true net edge of 2% passed this gate 94% of the time and a return-based gate 59%; a model at exactly zero net CLV passed 8.3%, within the 10% target. Profit itself is tested by the primary backtest test over thousands of games, which must pass first; that test, not the gate, is the protection against closing-line value that does not turn into return. The gate is reported as pass or fail and is not re-run with changed settings.
- Independent reproduction of the full pipeline from the repository by a third party.

- Results are published on a single dashboard: slope with interval, decision, betting-rule ROI, measured sd(d) and power, results by odds band, results by season, robustness checks, confirmation at the close.

9. Open questions for reviewers

- Resolved in v2.3 — sample size: the target is now the true edge τ in probability points, so a noisier model needs more games; the formula was checked by simulation.
- Resolved in v2.2 — market correction: with about 950 games per decile the decile correction carried roughly ±1.6 points of noise, close to the edge sought. Following external review, a smooth calibration replaces it, and the market log-odds enter the primary regression as a control.
- Resolved in v2.3 — seasons: 2019-20 is excluded; 2020-21, played largely without crowds, is a development season and is not scored in the test.
- Should the market correction be refitted on a rolling window within a season rather than once per season, given that the market reprices after every public signal?
- Is one-sided $p < 0.10$ the right threshold for the confirmation at the close, or should it match the primary threshold?
- Does a second pre-registered test on totals add power or dilute it?
- v2.1 (November 2026): the market correction is applied unconditionally and fitted on the longest price history; the bet side is fixed to the home team; the betting-rule ROI becomes a reported measure rather than a pass condition; the sample size is set by formula from the measured sd(d), with a no-go rule; a confirmation test at the closing price is added; totals may enter as a second pre-registered test; a component rule is added. Made after a simulation study of the test's size and power (1,000 synthetic seasons per setting) and reviewer feedback, before any test season was evaluated.
- v2.2 (November 2026): public player-level and officiating data, time-stamped, are allowed in the candidate model under the component rule, and licensed player-tracking data are excluded from this test; a data-extension rule is added (earlier NBA seasons first, then at most one additional league as a separate test, then no-go); results by odds band are reported for reference. Following external review: the decile market correction is replaced by a smooth two-parameter logistic calibration, with deciles kept as a diagnostic, and the market log-odds enter the primary regression as a control; an odds floor of 1.91 was considered and dropped, so the test covers all games. All changes were made before any test season was evaluated.
- v2.3 (November 2026): after a simulation study of the v2.2 primary test on 1,000 synthetic six-season histories per setting with eight seasons of price history, the market calibration gains a cubic term in the log-odds. The two-parameter form passed a no-information model 11.4% of the time under a tail-shaped favourite-longshot bias; the three-parameter form passed 6.8%, with power unchanged. The two-parameter form becomes a sensitivity check. Following contractor review: the sample-size target becomes the true edge τ in probability points ($N_{80} \approx 1.56 \cdot \mathrm{sd}(d)^2 / \tau^4$), checked by a second simulation; a capital gate on forward paper trading is fixed in advance. Both studies are published at sportindex.ai/simulation.html. Freeze decisions: seasons fixed (2019-20 excluded; development 2020-21 and 2021-22; test window 2022-23 to 2025-26); a pass requires both the CR1 t-test and a wild cluster bootstrap by game day — in simulation this brings the no-edge pass rate on the walk-forward window from 5.7–6.5% to 5.0–5.8%, at a cost of about one to three points of power; a mechanical fallback to $\tau = 0.03$ if the test window stays short of N_{80} after earlier seasons are added; a fixed rule for choosing the price source; CLV on totals when the line moves; the capital gate tests net CLV and the closing-line move, with realised return only as a consistency check. Made before any test season was evaluated.

Vladimir Roslyakov, Founder & CEO, SportIndex, Inc. · vr@sportindex.ai