Testing by betting when the bets are real: an e-value audit of 84 pre-registered hypotheses about a deployed forecaster

Anonymous Author(s)
Affiliation
Address
email

Abstract

E-values are motivated by a gambler betting against a null hypothesis. We report an audit where the gambler is not a metaphor: the null is a bookmaker's posted price, the stake is a stake, and the wealth process *is* the e-process. The subject is a deployed UFC bout-outcome forecaster and a pre-registered search for pockets where it beats the closing line: 84 hypotheses, 1,787 priced bouts, first analysed with paired log-loss and Benjamini–Hochberg. Re-asked as bets, three things change. (i) The 84 slices are overlapping subsets of one pool, so BH's guarantee is not established; its one favourable discovery fails under both Benjamini–Yekutieli and e-BH. (ii) A segment the fixed-n analysis reported as *failing to replicate* ($p = 0.30$) is, on the wealth scale, still growing in the replication window and 100 bouts short of $1/\alpha$ at fair odds, 253 at the book's: optional continuation converts a refutation into a schedule. (iii) A post-hoc threshold rule, paid for by a mixture martingale, survives the search it actually ran ($e = 31.3$) and dies twice: when the direction it could also have searched is priced in (15.6), and when the bookmaker's margin is (8.5). We report the ladder, not the winner. The three failure modes we price here — overlapping evaluation slices, a follow-up window read as a refutation, a threshold chosen after the sweep — are ordinary habits of empirical ML, and the market only makes their cost legible.

1 Introduction

Empirical machine learning has a set of standard habits for reading a table of results. We slice a held-out pool by subgroup, task, or difficulty, and report where the model wins. We charge the slicing with Benjamini–Hochberg (Benjamini and Hochberg, 1995). We run the promising slice again on a later window and call the second result a replication or a failure to replicate. When a threshold was chosen after looking at a sweep, we say so in a limitations section. Each habit is defensible in isolation. Each also has a sequential alternative that makes its cost a number rather than a caveat.

This paper is a case study in swapping all four at once, on a study that had already been completed and written up under the standard habits. The setting is unusual in a way that turns out to matter: the null hypothesis is a price, so the cost of being wrong is denominated in the same units as the evidence.

Testing by betting. The e-value literature (Shafer, 2021; Grünwald et al., 2024; Vovk and Wang, 2021; Ramdas et al., 2023) introduces its central object through a gambler whose capital cannot grow in expectation if the null is true. The gambler is usually a device — a construction that makes the mathematics vivid. Here it is the situation itself. We audit a deployed UFC bout-outcome forecaster

Submitted to 40th Conference on Neural Information Processing Systems (NeurIPS 2026). Do not distribute.

against a betting exchange's closing line. The null is a number a counterparty posted with capital at risk. A stake against it is an actual stake, settled at an actual price, and the wealth it accumulates is an e-process with no translation step in between.

What we did. A prior technical report on this pool (Anonymous, 2026) declared 86 candidate segments — filters on pre-fight covariates, each with a stated mechanism for why a good market would be wrong there — in a registry written before any of them was scored, then measured each with a composition-adjusted paired log-loss contrast and charged multiplicity with Benjamini–Hochberg. We re-ask all 84 pre-registered hypotheses (§3) as bets and report what changes.

Contributions.

1. **An e-process with no simulation step** (§4). Because the null is a posted price, we can give every figure twice: at *fair* odds, against the de-vigged probability, and at *real* odds, against the two decimals the book actually offered. The second is an e-value against a composite null that estimates nothing — the posted pair itself. The gap between the columns is what the bookmaker's margin costs, and on this pool it is a factor of five over 300 bouts.
2. **A multiplicity correction the original analysis was not licensed to make** (§5). The 84 slices are overlapping subsets of one pool, so BH's independence-or-PRDS condition is not established, and the dependence is too irregular for PRDS to be plausible. The single favourable discovery survives BH and dies under both Benjamini–Yekutieli and e-BH.
3. **A refutation re-read as a schedule** (§6). A segment reported as failing to replicate at $p = 0.30$ is, on the wealth scale, growing in the replication window at 82% of its discovery rate and 100 bouts short of $1/\alpha$. Optional continuation distinguishes "this was refuted" from "this is seventeen months of data away" — a distinction a fixed-n design cannot draw.
4. **A price for a post-hoc choice** (§7). The largest effect in the original analysis used a threshold chosen after its sweep was visible. A mixture martingale over the thresholds the search could have taken turns that admission into a quantity, and we report the ladder of increasingly honest mixtures rather than the rung that clears.
5. **A sample-size calculation for the sequential instrument** (§8), and the cost of anytime-validity on this pool's headline number.

What this paper is not. Betting the model against the book over the whole priced pool gives $e = 3.6 \times 10^{-4}$ at fair odds and 3.1×10^{-8} at real ones. This paper is not about beating a market. It is about what can be honestly claimed inside a search that mostly failed — and about how much of that honesty a fixed-n instrument is structurally unable to express.

2 Background: testing by betting in one page

We state only what the later sections use. Readers who know e-values can skip to §3.

E-values and Ville's inequality. An *e-value* for a null H_0 is a non-negative random variable E with $\mathbb{E}_{H_0}[E] \leq 1$. By Markov, $\Pr_{H_0}(E \geq 1/\alpha) \leq \alpha$, so $e \geq 20$ is a rejection at $\alpha = 0.05$ for a single pre-specified hypothesis. The interpretation is the one the name suggests: E is the factor by which a gambler's capital multiplied while betting against H_0 at odds H_0 implies are fair. A *test martingale* is a process (M_t) with $M_0 = 1$, $M_t \geq 0$, and $\mathbb{E}_{H_0}[M_t \mid \mathcal{F}_{t-1}] = M_{t-1}$. Ville's inequality (Ville, 1939) strengthens Markov to hold over the whole path,

$$\Pr_{H_0}(\sup_t M_t \geq 1/\alpha) \leq \alpha,$$

which is the property that does the work below: the process may be stopped, extended, or peeked at, at any time, without spending anything.

Optional stopping and optional continuation. Because the guarantee is time-uniform, a null that fails to be rejected by time t can simply be carried forward. There is no second test to pay for, and nothing is discarded. This is what separates "we did not reject" from "we reject", and, in §6, "it failed to replicate" from "it has not arrived yet".

Optional skipping. If membership in a subset is $\mathcal{F}_{t-1}$-measurable — decidable before the outcome — then betting only inside it, with $\lambda_t = 0$ outside, preserves the martingale property. Slicing a pool by a pre-fight covariate is therefore free; slicing it by anything the outcome touches is not.

Multiplicity under arbitrary dependence. e-BH (Wang and Ramdas, 2022) controls FDR for K hypotheses using e-values under *arbitrary* dependence: sort them, and reject the k largest for the largest k satisfying $e_{(k)} \geq K/(\alpha k)$. Nothing about the correlation structure enters. This is the price and the point: the threshold is high, but the guarantee does not have a condition that has to be argued for.

Paying for a search with a mixture. Any convex combination $\sum_j w_j M_t^{(j)}$ of test martingales with weights w_j fixed in advance is itself a test martingale (Robbins, 1970; Howard et al., 2021). If the $M^{(j)}$ index the choices a search could have made, the mixture is a valid e-value for the search, and the winner's evidence is deflated by exactly the prior mass placed elsewhere. The maximum over j is not an e-value; the mixture is.

Confidence sequences. The same construction inverted gives an interval valid at every n simultaneously (Waudby-Smith and Ramdas, 2024), rather than at the one n an analyst stopped at.

3 The audit setup

The pool. We audit a deployed UFC bout-outcome model against a betting exchange's closing line: $4{,}141$ walk-forward out-of-fold bouts scored between 2016 and 2026 by models refit at 32 quarterly origins; $1{,}787$ across 225 events carry an archived closing price. Two properties no simulation supplies:

The null is adversarial and external. H_0 is not an assumption of convenience — it is a number a counterparty posted with capital at risk, and a bettor who rejects it collects. Let q_t be the de-vigged closing probability: the two closing decimals stripped of their overround by the power method, which from raw implied probabilities $r = 1/d$ solves $r_a^k + r_b^k = 1$ and sets $q = r^k$. Under H_0, q_t is the true conditional probability of the outcome $y_t \in \{0, 1\}$.

Predictability is enforced by the calendar. The model probability p_t comes from a fit that saw only bouts strictly before its origin, and q_t is posted before the fight; both are predictable by construction. This is the condition §2 needs, and here it is supplied by the world rather than by an assumption.

The registry. The prior analysis (Anonymous, 2026) declared 86 candidate segments, each with a stated mechanism for why a good market would be wrong there, in a registry[1] written before any of them was scored. Two pairs are the same filter under different lenses, so the pre-registered family is 84; three more, added once results were visible, are post-hoc there and excluded here — $84 + 3$ is the 87 its title counts.

Why re-audit at all. The original analysis measured each segment with a composition-adjusted paired log-loss contrast, cluster-robust by event, and charged multiplicity with BH — an instrument whose own power section reports 10–15% against a 3 percentage-point edge, 0–1% *once BH multiplicity is charged*, and $\sim$89,900 bouts needed. There are $1{,}787$. The main problem with that report was that the instrument was known to be inadequate before its answers were read. That is what led to this one.

4 The wealth process

For a bout with de-vigged closing probability q and outcome y, a predictable stake $\lambda \in (-1/(1-q),\ 1/q)$ buys the increment $M = 1 + \lambda\,(y - q)$, and $\mathbb{E}_{H_0}[M] = 1$. The product over a segment's

[1] Pre-registration here means author-controlled version history, not third-party registration: the registry is one inspectable source file, the scan's input, whose commit precedes every artifact the scan produced by nine minutes in a public repository. (The repository, file, and commit identifiers are withheld for double-blind review and will be given in the camera-ready.) That fixes the hypotheses before they were scored. It does not fix them before any contact with a pool already months in use. No timestamp available to us could.

Table 1: Every e-value the paper reports, at both prices. *Fair* settles against the de-vigged probability q; *real* settles at the two decimals the book posted, and is an e-value against the composite null of §4, with nothing estimated inside it. For a single pre-specified hypothesis the threshold is $1/\alpha = 20$; the 84-hypothesis e-BH cutoff at $k{=}1$ is 1680. The row marked † is reported but not claimed — it is measured on the window that selected it.

	window	n	e (fair)	e (real)
The model against the book, everywhere				
whole priced pool	2016–2026	1,787	3.6×10^{-4}	3.1×10^{-8}
Win streak ≥ 4 — the one segment BH liked (BH $q = 0.016$)				
discovery window†	<2025	198	31.39	8.54
replication window	2025–26	110	4.81	2.48
full pool, 5-seed median	2016–2026	308	86.37	12.30
full pool, 5-seed range	2016–2026	308	56.2–150.9	8.39–21.18
Back the favourite the model over-bids — a post-hoc rule, priced				
max over thresholds‡	2016–2026	281	118.99	—
mixture, 21 thresholds	2016–2026	281	31.28	8.47
+ the fade direction (42)	2016–2026	281	15.64	4.23
+ the other statistic (63)	2016–2026	281	10.43	2.82

‡Not an e-value — the maximum over a searched family is exactly what the mixtures below it pay for.

bouts *in calendar order* is a non-negative test martingale, so its terminal value is an e-value and, by Ville's inequality, it may be stopped at any time. Every segment filter is a function of pre-fight covariates, so membership is $\mathcal{F}_{t-1}$-measurable and restricting the product to a segment is optional skipping, preserving the martingale property. The growth-optimal stake under the alternative "the model's p is the truth" is $\lambda^\star = (p-q)/[q(1-q)]$, the edge deflated by the market's own variance. We report fractional Kelly at $\lambda = \frac{1}{2}\lambda^\star$ and, because that fraction is a free choice, a predictable plug-in (Waudby-Smith and Ramdas, 2024) that learns the multiplier from past bouts only (§11).

The margin is a supermartingale, not an excuse. Write $f = \lambda q$ for the fraction of wealth the stake commits to the backed side and d for the decimal the book offers; a fair book would offer $d = 1/q$. Settling at d gives $\mathbb{E}_{H_0}[M] = 1 - f(1 - q\,d) \leq 1$, since $q\,d < 1$ whenever the book charges an overround (for $\lambda < 0$, read d_b and $1-q$): the deployable process is a valid e-value, uniformly conservative, short of the statistical one by what the margin costs. On this pool the mean overround is 1.0482 and the shortfall 0.00525 nats per bout, so every e-value we report is given at both prices, *fair* and *real* — a factor of five over 300 bouts, and seven on the streak segment, whose stakes run larger.

The *real* column needs no de-vig. q enters only the stake, which is free, so it need not enter the null: $\mathbb{E}_\pi[M] \leq 1$ for every true π in $[\,1 - 1/d_b,\ 1/d_a\,]$, the band no bet at these two prices can exploit, non-empty because the book charges an overround. Every *real* figure we report is therefore an e-value against that composite null — the posted pair itself, nothing estimated inside it.

Table 1 collects every e-value in the paper, at both prices.

Validation before use. Under a true null — prices and stakes held, only the outcome resampled from q — 40,000 replications give $\mathbb{E}[e] = 0.985 \pm 0.018$ per segment, $\Pr(\sup_t e_t \geq 1/\alpha)$ of $0.029/0.067/0.151$ at $\alpha = 0.05/0.10/0.20$, none of the 84 above its own α, e-BH rejecting at all in 74 of the 40,000 runs, and the confidence sequence of §8 covering the truth at *every* t in 97.4% of them.

5 The multiplicity charge BH was not licensed to make

The 84 slices are overlapping subsets of one pool of 1,787 bouts, scored by one model against one book: a single upset moves dozens of them together. BH's FDR guarantee (Benjamini and Hochberg, 1995) requires independence or PRDS (Benjamini and Yekutieli, 2001), and neither is established here. The dependence is also too irregular for PRDS to be plausible: across the 3,486 segment pairs, membership correlations run from -0.49 to $+0.84$, half of them negative, and 83 pairs share no bouts at all. The classical repair is Benjamini–Yekutieli, which charges $\sum_{i \leq 84} 1/i = 5.01$. e-

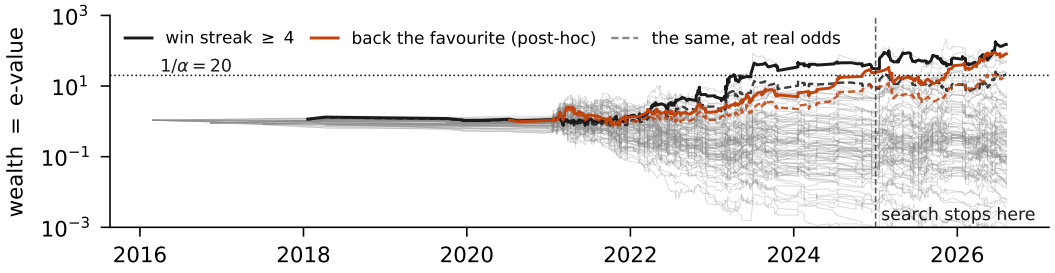

Figure 1: Wealth over calendar time for every pre-registered segment (grey) and the two the analysis singled out (coloured); dashed, the same bets at the book's real prices. At the vertical rule the search stopped and the replication window began — a fixed-n analysis must start a new test to its right, a test martingale need not fix its n.

BH (Wang and Ramdas, 2022) is valid under *arbitrary* dependence: its threshold for k rejections is $K/(\alpha k)$, whatever the correlation structure. On the discovery window the lab searched — 1,191 bouts, the same 84 hypotheses — BH rejects four, one in the model's favour and three against it. No back-fitted or decoy-padded registry would produce this. Benjamini–Yekutieli and e-BH (cutoff $e \geq 1680$ at $k{=}1$), the two procedures whose guarantee holds here, reject none. The favourable one was a fighter carrying a win streak of four or more, at BH $q = 0.016$.

It would be wrong to read this as e-values buying less. The same segment's e-value on the full pool is 86.4 at fair odds — the five-seed median — and for a *single* pre-specified hypothesis $e \geq 20$ already rejects at $\alpha = 0.05$; at real odds it is 12.3, which falls short. What is expensive is not the machinery but the charge for asking 84 questions of a pool this size — the original analysis's own lesson in visible currency.

6 "Failed to replicate" and "not yet" are different findings

The original analysis froze its discovery window, selected the streak segment, and ran a fresh fixed-n test on 2025–26. It returned $p = 0.30$ and was reported, correctly under its own logic, as a failure to replicate — the only move a fixed-n design has, since the second window buys a second test and nothing carries over.

On the wealth scale, restarted at 1 because the segment was chosen on the left, the same 110 bouts read differently. Log-growth in the replication window is $+0.0143$ nats per bout against $+0.0174$ in the window that selected it — 82% of the discovery rate, the shrinkage one expects when a slice is chosen on the data that flatters it, but the same sign and order. The e-value accumulated there is 4.81 — real evidence, short of $1/\alpha = 20$. At that rate, crossing needs 210 bouts and 110 have arrived — the e-process is valid whenever stopped, but the crossing date is only a forecast. The finding is not refuted; it is 100 bouts short, and at about 71 a year that is seventeen months of multiplying (Figure 1). At real odds the window pays 2.48, the margin taking 0.0060 of the 0.0143, and the wait is another 253 bouts: three and a half years.

We report this segment's pre-split wealth (31.4 fair, 8.54 real) but do not claim it: that is the window that chose it. The effect is stable across the pool's five walk-forward seeds — fair 56.2 to 150.9, real 8.4 to 21.2, medians 86.4 and 12.3 — a spread q-values had no natural place to put. Only the most favourable seed clears $1/\alpha$ at real odds, and it is the one decomposed above.

7 Pricing the freedom a post-hoc rule used

The largest effect in the original analysis is a direction rather than a segment: where the model gives the book's favourite *more* than the book does, flat-staking that favourite returned $+11.5\%$ over 281 bets across 163 events — declared post-hoc there: the cut at $\text{lean} \geq 0.05$ came after seeing the sweep positive from 0.02 to 0.22, peaking at 0.13, and a fixed-n framework has no principled penalty for that.

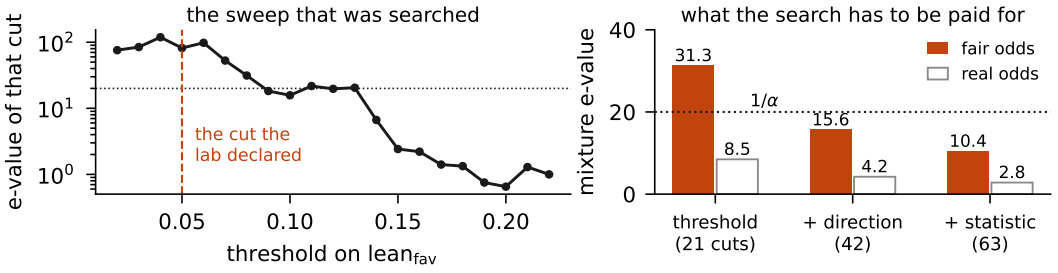

Figure 2: Left: the e-value of each threshold on the model's lean toward the book's favourite; the declared cut at 0.05 was chosen after this shape was visible. Right: a mixture over progressively more of the freedom the search had. At fair odds the finding pays for the threshold but not the direction; at real prices, neither.

A mixture martingale has one, with weights fixed before any cut won (Robbins, 1970; Howard et al., 2021): any convex combination of the thresholds' wealth processes is a martingale, and the search is paid for by the prior mass not on the winner. The maximum over thresholds, 119, is not a valid e-value. The uniform mixture over the 21 cuts is $e = 31.3$, which clears $1/\alpha$ (Figure 2). Widening it to the 21 cuts the search could also have taken in the opposite direction gives 15.6, below $1/\alpha$. Adding the other disagreement statistic on the same frame gives 10.4. Those are fair odds; at the prices the book offered, the same mixtures are 8.5, 4.2 and 2.8 — a bettor paying the margin never clears $1/\alpha$.

We regard the ladder, not any rung, as the result: it turns "this was post-hoc" from a caveat into a quantity. The finding absorbs the threshold search at fair odds, and neither the direction it could have taken nor the margin it would have paid.

8 What the sequential instrument needs

A growth rate makes planning a calculation rather than a verdict. Against a true 3 pp edge simulated on this pool's own prices, the half-Kelly bet grows at 0.00166 nats per bout, so expected log-wealth reaches $\log(1/\alpha)$ in $\sim$1,800 bouts — a 59% chance of having crossed by then, which is weaker than a power guarantee. Matched at 80% over 20,000 simulated runs, the sequential test needs 2,385 bouts to the paired log-loss test's 5,346;[2] at 5 pp, 864 to 1,902. It is roughly twice as cheap in fights. The reason is not that it is more powerful per observation; it is that a paired contrast on a slice discards most of what each bout carries.

Anytime-validity is not free. On the directional rule the cluster bootstrap gives a 95% ROI interval of $[+2.2\%, +20.4\%]$, valid at the one n the analysis stopped at — chosen after watching the number. The hedged betting confidence sequence (Waudby-Smith and Ramdas, 2024), valid at every n, gives $[-5.3\%, +22.9\%]$ and does not exclude zero.

9 Related work

The game-theoretic foundation of testing by betting is due to Shafer (2021) and Ramdas et al. (2023), building on Ville (1939); Grünwald et al. (2024) develop the same object as safe testing and Vovk and Wang (2021) as the calibration and combination calculus for e-values. Our multiplicity treatment is e-BH (Wang and Ramdas, 2022), the dependence-free counterpart to Benjamini and Hochberg (1995), whose condition Benjamini and Yekutieli (2001) both diagnosed and repaired. The mixture device we use to price a search is Robbins' method of mixtures (Robbins, 1970), in the time-uniform form of Howard et al. (2021); the betting confidence sequence and the predictable plug-in stake are from Waudby-Smith and Ramdas (2024).

[2] Both are idealised in the same way: the simulated model deviates from the price only by the edge, so the paired contrast it produces is far quieter than a real one (sd $= 0.065$ nats). That is why 5,346 sits so far below the $\sim$89,900 quoted in §3, which prices a different estimand — a fixed-n test of a segment against the pool, at the dispersion such contrasts actually realise — and which we have not recomputed here. We claim the ratio within a row, never the level, and never a comparison across the two.

What we add is not a method. It is a setting in which the abstraction is literal — the null is a posted price, the stake is a stake — and a side-by-side re-analysis of a completed study whose fixed-n conclusions are on record. In most applications the null is a modelling choice the analyst makes and then tests against; here it is posted by a counterparty with capital at risk, so it is a null no one chose for its convenience. The bookmaker's margin then puts a floor under how much of the evidence is actually collectible, which is a discipline a self-chosen null does not impose.

10 What transfers to ML practice

The sport is incidental. Three of the habits this audit replaced are near-universal in empirical ML, and the market's contribution is only that it makes their cost payable in a visible currency.

Overlapping evaluation slices break BH. A per-subgroup, per-task, or per-difficulty-bucket breakdown of one model on one held-out set is a family of overlapping subsets of one pool, and a handful of hard examples moves many cells together. That is the structure of §5: correlations from -0.49 to $+0.84$, half of them negative. BH is the default correction in this situation and its condition is rarely checked. Here, checking it cost the study its only favourable discovery. e-BH needs no such check.

A follow-up window is not a replication test. Re-running a promising slice on newer data and reporting a p-value is standard, and it forces a binary verdict on a quantity of evidence that may simply be incomplete. On the wealth scale the same 110 observations say "growing at 82% of the discovery rate, 100 short of the threshold" (§6). Both readings are correct under their own instrument; only one of them tells you what to do next.

Post-hoc thresholds can be priced instead of confessed. Choosing a cutoff after seeing the sweep is ordinary — and ordinarily handled by a sentence in the limitations. A mixture over the choices the search could have made converts that sentence into a deflation factor (§7), and the interesting result is that the finding survives the threshold search and not the direction it could equally have taken. The ladder is the honest object, because the question "how much freedom did you actually have?" has more than one defensible answer, and the mixture makes each answer's price explicit.

The same table, in a setting with no bookmaker. Make the first of these concrete. One model, one held-out set of a few thousand examples, and a results table that breaks accuracy down by input length, by language, by domain, and by a hardness flag — twenty or thirty cells, each with a stated reason to expect a gap, each charged with BH. That is the shape of §5 exactly. The cells are not disjoint: a long, low-resource, hard example sits in four of them at once, so a handful of such examples moves many cells together, and any cell built from a conjunction of attributes is nested inside each of its parents. Crosscutting attribute slicing produces the correlation structure of §5 — ours ran from -0.49 to $+0.84$ with half the pairs negative — by construction rather than by accident. Nothing in such a table establishes independence or PRDS, and the condition is rarely checked before the correction is applied.

Re-running it as a wealth process costs one design decision. Each example needs a stake that is predictable, which means the examples need an order fixed before their outcomes are read. Our calendar supplied one for free; a static held-out set does not, so the analyst has to choose — collection order, or a seed committed in advance — and defend the choice, because reordering after seeing the scores destroys exactly the property the guarantee rests on. That is a real cost. What it buys is a correction whose validity does not depend on a condition nobody checked.

A caution on transfer. Beyond the ordering cost just described, one thing does not carry over at all: the *real* column has no analogue without a counterparty. There is no margin to pay, so the pessimistic bound that disciplined every headline number in this paper — the column that killed the post-hoc rule outright — would have to be replaced by something else, and we do not know what.

11 Limitations

The stake and the de-vig are both free choices, and the result holds under every setting we tried. Fractional Kelly at $0.25/0.5/1.0$ gives $e = 26.0/150.9/115.3$ on the streak segment's most favourable seed (the seed of §6) and the predictable plug-in, which is no choice at all, settles at $c = 0.70$, $e = 63.1$ (below both its neighbours, having paid to learn c); every variant clears $1/\alpha$. Under Shin's de-vig those fair-odds figures read 120.6, 4.36, $34.0/17.0/11.3$ against 150.9, 4.81, $31.3/15.6/10.4$; the proportional split, which flatters this rule, would *raise* it to $56.3/28.2/18.8$.

Boosting e-BH (Wang and Ramdas, 2022) would recover power but needs each e-value's null distribution, which a composite martingale null does not supply; pre-specified weights would have been the better route here.

Three contaminations are inherited from the pool and sized there — an age correction shipped on replication-window data, hyperparameters tuned on a window the pool later scores (122 of 1,191 discovery bouts), a rating not fully point-in-time (21.9% of bouts) — and removing any moves the numbers unfavourably. And this is one sport, one book, one model: the method transfers, the numbers do not.

12 Open questions we would value feedback on

This is a re-analysis, and several of its choices are ones we would like to argue about rather than defend.

1. **How wide should the mixture be?** §7 reports three nested families because we could not find a principled place to stop. The widest family we can imagine is not the same as the freedom the search actually had, and the narrowest is self-serving. Is there a defensible rule for reconstructing a search's true support after the fact, or is the ladder the honest terminal object?
2. **Is the *real* column the right kind of conservatism?** Betting at the posted decimals gives an e-value against a composite null with nothing estimated inside it, which is attractive. It also confounds "the market is efficient" with "the margin is large", and we are unsure whether reporting both columns clarifies that or hides it.
3. **What is the honest way to report a crossing date?** §6 says a segment is 100 bouts short at the current growth rate. That rate is itself estimated on the same wealth path, so the forecast is not anytime-valid even though the process is. We report it as a forecast; we suspect there is a better convention.
4. **Does the pre-registration argument hold?** The footnote in §3 states what a commit history can and cannot establish. We would like to know whether reviewers find that distinction sufficient, or whether an audit of a pool already in use is simply not pre-registrable by any available means.

References

Anonymous. Where does the model beat the book — 87 slices, one survivor. Technical report, 2026. Prior work by the present authors; the registry — one file, the scan's input — together with the scoring code and full write-up sits in a public repository whose URL is withheld for double-blind review and will be given in the camera-ready.

Y. Benjamini and Y. Hochberg. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B*, 57(1):289–300, 1995.

Y. Benjamini and D. Yekutieli. The control of the false discovery rate in multiple testing under dependency. *Annals of Statistics*, 29(4):1165–1188, 2001.

P. Grünwald, R. de Heide, and W. Koolen. Safe testing. *Journal of the Royal Statistical Society: Series B*, 86(5):1091–1128, 2024.

S. R. Howard, A. Ramdas, J. McAuliffe, and J. Sekhon. Time-uniform, nonparametric, nonasymptotic confidence sequences. *Annals of Statistics*, 49(2):1055–1080, 2021.

A. Ramdas, P. Grünwald, V. Vovk, and G. Shafer. Game-theoretic statistics and safe anytime-valid inference. *Statistical Science*, 38(4):576–601, 2023.

H. Robbins. Statistical methods related to the law of the iterated logarithm. *Annals of Mathematical Statistics*, 41(5):1397–1409, 1970.

G. Shafer. Testing by betting: a strategy for statistical and scientific communication. *Journal of the Royal Statistical Society: Series A*, 184(2):407–431, 2021.

J. Ville. *Étude critique de la notion de collectif*. Gauthier-Villars, Paris, 1939.

V. Vovk and R. Wang. E-values: calibration, combination and applications. *Annals of Statistics*, 49(3):1736–1754, 2021.

R. Wang and A. Ramdas. False discovery rate control with e-values. *Journal of the Royal Statistical Society: Series B*, 84(3):822–852, 2022.

I. Waudby-Smith and A. Ramdas. Estimating means of bounded random variables by betting. *Journal of the Royal Statistical Society: Series B*, 86(1):1–27, 2024.