Testing by betting when the bets are real: an e-value audit of 84 pre-registered market-efficiency hypotheses

Roman Prigodskii
Independent researcher
lioneldassy@gmail.com

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 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.

1 A null hypothesis with money behind it

Testing by betting (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. However, here it is the situation itself.

We audit VERTEX, 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.

The prior analysis of this pool (Prigodskii, 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

[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 (`8f9fcbd`) precedes every artifact the scan produced (`6b5fbd9`) by nine minutes

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

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. It measured each 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 $1{,}787$. That report is the reason for this one: the instrument was known to be inadequate before its answers were read.

2 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 bouts *in calendar order* is a non-negative test martingale, so its terminal value is an e-value and, by Ville's inequality (Ville, 1939), $\Pr_{H_0}(\sup_t M_t \geq 1/\alpha) \leq \alpha$: 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 ($\lambda_t = 0$ off it), 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 (§7).

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 below is given at both prices, ***fair*** and ***real*** — a factor of five over 300 bouts, and seven on the segment below, 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 below is therefore an e-value against that composite null — the posted pair itself, nothing estimated inside it.

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 §6 covering the truth at *every* t in 97.4% of them.

Where the model actually stands. 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.

3 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 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-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

in a public repository. 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.

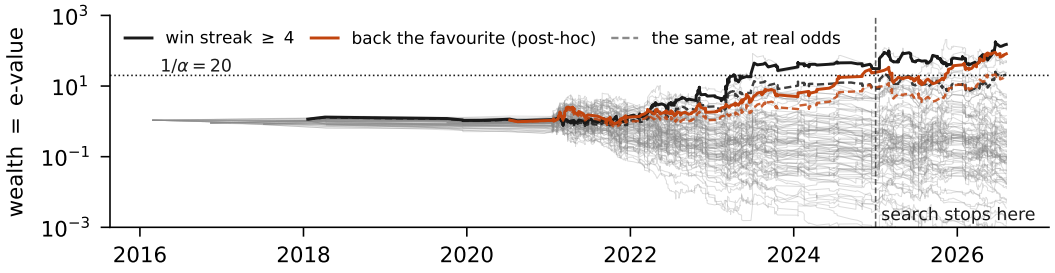

Figure 1: Wealth over calendar time for every pre-registered segment (grey) and the two the analysis singled out (coloured); dashed, the same bets at the book's real prices. At the vertical rule the search stopped and the replication window began — a fixed-n analysis must start a new test to its right,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

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.

4 "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.

5 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.

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 $1/\alpha$.

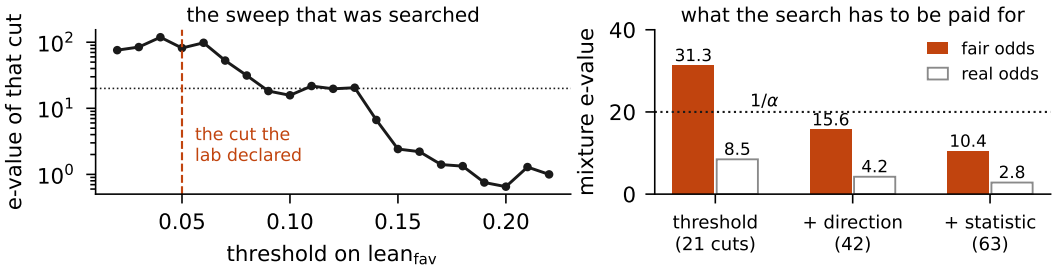

Figure 2: Left: the e-value of each threshold on the model's lean toward the book's favourite; the declared cut at 0.05 was chosen after this shape was visible. Right: a mixture over progressively more of the freedom the search had. At fair odds the finding pays for the threshold but not the direction;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

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.

6 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.

7 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 §4) and the predictable plug-in, which is no choice at all, settles at $c = 0.70$, $e = 63.1$ (below both, 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.

[2] Both are idealised in the same way: the simulated model deviates from the price only by the edge, so its per-bout contrast is four times quieter than a real one's (0.065 against 0.27 nats). §1's $\sim$89,900 is this same arithmetic at that realised dispersion, which is the whole of the gap. These numbers claim the ratio only.

References

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.

R. Prigodskii. Where does the model beat the book — 87 slices, one survivor. Technical report, VERTEX MMA, 2026. The registry — one file, the scan's input — is at `https://github.com/romanprigodskii/vertex-mma/blob/8f9fcbd3ccd4eb65f9aa06d7615de0e3363914f7/scripts/simulation/scripts/lab_edge_registry.py`; the scoring code and full write-up sit beside it.

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.