How Much Tail Prediction Could We Have Detected? A Power Accounting for a 1,450-Test Search

Charlie Yan*

August 2026

Abstract

We searched eight feature families—intraday microstructure, skew structure, quote geometry, event anatomy, funding spreads, cross-asset carry, GARCH/EVT states, and sector-rotation dispersion—for predictors of the left tail of a short-volatility options book: roughly 1,450 tests against a fixed pre-registered target panel, with adversarial re-verification of every candidate. Zero survived the family-wise bar. This paper argues that such a null is nearly uninterpretable unless the searcher also reports what the search *could* have found, and supplies that accounting. The honest unit of the tail sample is not the 1,237 rows of the target panel (1,168 of which carry a complete ten-day forward window) but approximately fifteen distinct loss episodes; after deflating for overlapping horizons, the search had 80% power at the Bonferroni-adjusted threshold only against true tail information coefficients of $|IC| \geq 0.44$ at the ten-day horizon and ≥ 0.32 at five days, while event-level classifiers were certifiable only above an area under the ROC curve of 0.83. The best graded candidates sat at 0.21–0.23—half the certifiable threshold. The correct conclusion is therefore a bound: any true separator in the tested families is weaker than these thresholds, weaker than anything the positioning and flow literature sustains, or was not among the tested features. Two structural findings sharpen the bound and survive it. First, a census of the fifteen episodes shows only two began from a calm state; the other thirteen were preceded by elevated generic risk states that carry conditional *means* three to thirteen times baseline in every year of the sample—so the states that forecast the tail also forecast the premium, and de-sizing on them forfeits more than it saves. Second, protection was 4–16% more expensive precisely when candidate separators fired: the market charges for the same information at the moment it becomes available. We propose that tail-signal searches publish this power accounting as standard practice, and offer the ratio of tail-IC to mean-IC (≥ 2) as a pre-registration gate that removes the premium-confound false positives which otherwise dominate such searches.

1 Why a null needs a power statement

Negative results in quantitative finance are published rarely and read sceptically, and the scepticism is usually well founded: "we looked and found nothing" is consistent with there being nothing to find, with looking in the wrong place, and with looking with an instrument too blunt to see what is there. Only the first is a scientific claim. Distinguishing them requires the searcher to state, before or alongside the result, the smallest effect the search could have certified.

*Draft v0.3 (journal-length), August 2026. All statistics SCREEN-basis: information coefficients and event classification. No performance claims appear in this paper. The detectability table and Exhibit 1 are produced by a checked-in script (`p3_power_table.py`), not by in-text arithmetic. Code, pre-registration documents and a data-availability statement: https://github.com/charlieyanhx/tail-power-accounting.

For tail prediction this matters more than usual, because the effective sample is far smaller than the nominal one in two compounding ways. Overlapping multi-day targets deflate independent observations by roughly the horizon; and the events of interest—the days a short-volatility book actually loses badly—are rare by construction. A study reporting a four-figure observation count and a Bonferroni correction over a large test ledger sounds rigorous and may in fact be nearly powerless. This paper supplies that accounting for one such search and argues it should be routine.

2 The search

The target panel fixes, per trading day: forward five- and ten-day downside semivolatility, the worst single-day loss in units of trailing volatility, a tail-event indicator, and—critically —the forward *mean* over the same window, as an explicit confound target. Features were drawn from eight families and tested as information coefficients against the tail targets at horizons of five and ten days, with family-wise error controlled by Bonferroni across a declared ledger of roughly 1,450 tests, and with every near-miss re-derived independently by a second implementation before being graded.

The economics matter for interpreting the bar. The book being protected is short volatility, so the payoff to a true separator is de-sizing before losses. The cost of a false one is forfeited premium—and in this sample the richest sub-period is the *non-acute* portion of stress episodes, which is precisely when a naive risk-state signal would have the book standing down. A tail signal must therefore not merely predict losses; it must predict them in a way that is separable from the premium it would cause the book to forgo. Section 5 turns that requirement into a testable gate.

3 The power accounting

Overlap deflates the sample. The target panel holds 1,237 daily rows, of which 1,168 carry a complete ten-day forward window. Overlapping h-day targets leave roughly n/h independent observations: at $n = 1{,}168$ usable daily rows, $n_{\text{eff}} \approx 116$ at $h = 10$ and ≈ 233 at $h = 5$.

The detection threshold. Under the Fisher transform, $\text{se} \approx 1/\sqrt{n_{\text{eff}} - 3}$. With a Bonferroni critical value $z_\alpha = 4.14$ at $\alpha = 0.05/1{,}450$ and $z_{0.8} = 0.84$ for 80% power,

$$|IC|_{\min} \;\approx\; \tanh\!\Big(\frac{z_\alpha + z_{0.8}}{\sqrt{n_{\text{eff}} - 3}}\Big). \tag{1}$$

Horizon	n_{eff}	$\lvert IC\rvert_{\min}$, Bonferroni + 80% power	$\lvert IC\rvert_{\min}$, nominal 5%
10 days	≈ 116	≈ 0.44	≈ 0.26
5 days	≈ 233	≈ 0.32	≈ 0.18

Table 1: Smallest detectable tail information coefficient, from Equation (1).

Event classification is harsher still. Treating the problem as classification rather than correlation, with $n_1 \approx 15$ loss episodes against $n_2 \approx 1{,}150$ ordinary days, the Hanley–McNeil standard error of an AUC near 0.75 is ≈ 0.073 (Hanley and McNeil, 1982); certifying at the same threshold and power requires AUC ≥ 0.828. For calibration: that is a classifier that ranks a randomly chosen loss episode above a randomly chosen ordinary day more than four times in five, on daily data, out of sample. Rather than assert that no such feature exists—a claim we cannot support without a

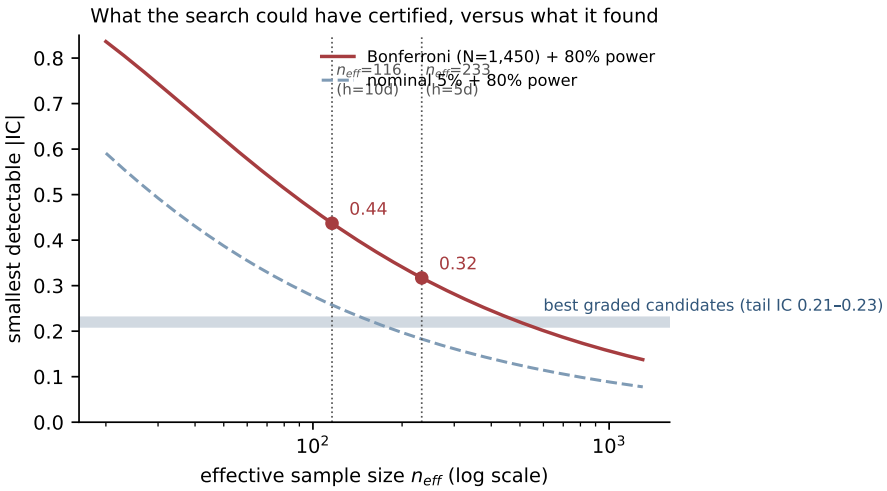

Figure 1: Exhibit 1. The detectability frontier against effective sample size, with the search's two operating points and the band occupied by its best graded candidates. The candidates sit at roughly half the certifiable threshold: the instrument could not have resolved effects of the size that plausibly exist.

systematic review—we note the implied effect size: sustained daily-horizon information coefficients in the positioning and flow literature are typically an order of magnitude below the correlation such an AUC implies, and our own best-documented positioning feature caps at $|IC| \approx 0.08$. A reader who knows of a feature at this level has, by the same accounting, identified something the search would indeed have found.

What the null therefore means. The best graded candidates in the search—a quote-geometry feature and a sector-rotation dispersion measure, each with tail IC in the 0.21–0.23 range, a coherent mechanism, and the correct sign out of sample—sit at about half the certifiable threshold. The search was structurally unable to certify effects of the magnitude that a realistic tail predictor might have. The publishable conclusion is the bound, not the slogan: *any true separator in these families is weaker than* $|IC| \approx 0.44$ *at ten days, or was not among the features tested.* That is a considerably weaker claim than "nothing predicts the tail", and it is the one the data support.

4 Two structural findings that survive the null

The killer census. We catalogued the fifteen worst episodes and asked, for each, what the candidate features looked like on the preceding day. Only two began from a genuinely calm state; the remaining thirteen were preceded by elevated generic risk features. This sounds encouraging until one examines the same states' conditional means: they are three to thirteen times baseline, in every year of the sample. The states that forecast the tail also forecast the premium. This does not by itself condemn a de-sizing rule: what matters is the *ratio* of tail loss avoided to premium forfeited, and a state whose tail scales faster than its mean would still pay to act on. The question is therefore quantitative, and in our data the ratio is unfavourable—every walk-forward overlay we tested has an improvement interval containing zero. We state that as the absence of a demonstrable gain rather than as a demonstrated loss: our intervals are wide enough to admit a small improvement we lack the episodes to resolve. What the census establishes is narrower but sufficient for the search's

purpose—detection was never the binding problem, so better detection is not the lever.

Stress is pre-priced. Protection—put wings, downside skew—was 4–16% more expensive conditional on candidate separators firing. The market charges for the same information the separator carries, at the moment it carries it. This is the mechanism behind the previous paragraph rather than a separate fact, and it is the reason we regard the ceiling on this family of research as economic rather than statistical: a public, contemporaneous signal of elevated risk is priced into the instruments one would use to act on it.

5 A pre-registration gate for tail research

The dominant false-positive class in our search was not noise but *confounding*. Features predict the tail because they predict volatility, and volatility predicts both tails and premium. Multiplicity correction does not address this: a confounded feature can be genuinely, robustly, replicably correlated with the tail and still be useless, and Bonferroni will pass it if the correlation is strong enough.

We therefore propose registering, in advance, the requirement

$$\frac{|IC_{\text{tail}}|}{|IC_{\text{mean}}|} \;\geq\; 2,$$

computed against an explicit forward-mean confound target constructed on the same panel and horizon. In our search this single gate removes every premium-confound candidate while retaining the (uncertifiable but mechanistically coherent) near-misses—a better trade than raw multiplicity correction, which removes both. We flag the obvious circularity: the threshold was set by inspection on the same search whose candidates it is then reported to separate, so that separation is a description of our data, not an out-of-sample validation of the number 2. What transfers is the *form* of the gate—ratio to a named confound target, registered before results—not our particular threshold. The gate is cheap: it requires one additional target column and one additional ratio per test, and it forces the researcher to name, before seeing results, the conditional-mean confound that would otherwise be discovered late and rationalized.

This connects to a live debate. Moreira and Muir (2017) showed that volatility-managed portfolios improve Sharpe ratios; Cederburg et al. (2020) and subsequent work show the improvement is fragile out of sample and sensitive to implementation. Recent work on regime probabilities and the horizon profile of tail risk documents the same confound we encounter: apparent tail predictability entangled with conditional means. Our contribution to that debate is not another test but an ex-ante instrument for separating the two—and the observation, from the census, that in our data the entanglement is not a nuisance to be controlled away but the dominant feature of the problem.

6 Relation to prior work

The multiple-testing literature in finance (Harvey and Liu, 2015; Harvey et al., 2016) establishes that a declared ledger and a family-wise correction are necessary; our point is that they are not sufficient for a *null*, which additionally requires a power statement. Outside finance this is standard: clinical trials report the effect size a study was powered to detect, and a null without one is not publishable. Ioannidis (2005) makes the general case that low-powered studies produce uninformative results in either direction. The finance-specific complication is that effective sample size is not the row

count: overlapping horizons and rare events deflate it, sometimes by an order of magnitude, and reporting the nominal n obscures exactly the quantity a reader needs. Hanley and McNeil (1982) give the classification analogue we use. On the economics, Bates (2000) and the crash-prediction literature more broadly have long found tail timing elusive; our census suggests one reason that is under-emphasized—not that warnings are absent, but that they arrive attached to compensation.

7 The gate is not vacuous: features that pass it exist

Two strands of recent work show the confound our gate targets is pervasive rather than peculiar to our book, and that the gate nonetheless admits real candidates. On the confound: risk-neutral tail-risk measures are found to *positively predict* excess market returns, and tail exposure is priced in the cross-section — so a feature correlated with tail risk is, by default, also correlated with expected return, which is exactly the entanglement that makes raw tail ICs uninterpretable. On the other side, recent work on factor crowding reports a feature that *predicts the probability of extreme losses even when it fails to predict mean returns* — precisely the object our ratio is designed to isolate, and evidence that the gate screens rather than merely excludes. A gate that no real feature could pass would be a restatement of pessimism; this one has a documented example on the passing side.

8 Limitations

One book, one underlying, 5.7 years, approximately fifteen episodes: the bound is only as large as the sample, and a longer sample or a cross-section of books would tighten it substantially—indeed the most useful extension of this paper would be the same accounting run on a panel of strategies, where episodes accumulate across books rather than only across time. The power calculation uses normal approximations and n/h effective-sample deflation, both conventional and both approximate; a block-bootstrap power study would be more exact and, we expect, no more encouraging. Features outside the eight families remain untested—notably initiator-signed per-strike option flow with opening/closing attribution, which we could not adequately proxy from the data we own, and which is the search's declared blind spot. Finally, the tail-to-mean gate we propose is calibrated on one program's experience; the threshold of 2 is a judgment, not an estimate, and we would welcome its recalibration by anyone with a wider sample of confounded candidates.

References

Bates, D. (2000). Post-'87 crash fears in the S&P 500 futures option market. *Journal of Econometrics* 94(1–2).

Cederburg, S., O'Doherty, M., Wang, F., Yan, X. (2020). On the performance of volatility-managed portfolios. *Journal of Financial Economics* 138(1).

Hanley, J., McNeil, B. (1982). The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology* 143(1).

Harvey, C., Liu, Y. (2015). Backtesting. *Journal of Portfolio Management* 42(1).

Harvey, C., Liu, Y., Zhu, H. (2016). . . . and the cross-section of expected returns. *Review of Financial Studies* 29(1).

Ioannidis, J. (2005). Why most published research findings are false. *PLoS Medicine* 2(8).

Moreira, A., Muir, T. (2017). Volatility-managed portfolios. *Journal of Finance* 72(4).