Test Everything, Publish Both Results: A Protocol That Cuts Backtest False Positives from 37% to 0.4%

Alex Vidovich
Independent researcher · alexvidovich.com
Working paper — August 2026 (rev. 3)

Abstract. This paper delivers four artifacts for practitioners of systematic strategy research. First, a taxonomy of recurring backtest failure modes, each with a testable diagnostic signature. Second, a Monte Carlo study of both error types across common research behaviors. Under a pure null with twenty variants per idea, tuning on the full sample falsely accepts 36.9% of candidates (a closed-form result: selecting on a reused holdout is the *same* quantity, since a holdout consulted for selection is not a holdout), a disciplined single holdout achieves its nominal 2.3%, and a trial-count-deflated hurdle achieves 0.4% — with the popular fold-consistency heuristic shown to be near-vacuous (87% acceptance) because selection contaminates every fold. Under an era-locked alternative — a real in-sample effect confined to one regime — the entire standard stack fails (deflation passes 95%), motivating our third artifact: leave-best-era-out (LBEO), a pre-registrable criterion that collapses era-locked acceptance to 1.2% at a null size of 0.04%, and whose measured cost is power — 7% detection of a stable SR 0.5 on eight years of dailies, versus deflation's 28%, both falling further when trial variants are realistically correlated. Severity, in other words, is purchased with power at exactly the effect sizes that survive costs, and we quantify the price rather than deny it. Fourth, the burial-record template that converts each rejection into durable research infrastructure, including an operational counting rule for the quantity deflation depends on and presentations omit: the number of trials. All results are simulation- and backtest-based; nothing here constitutes investment advice.

Keywords: backtesting, negative results, multiple testing, cross-validation, overfitting, pre-registration, research process.

JEL classification: C12, C15, C52, C58, G11, G14.

---

1. Introduction

A textbook account of quantitative research describes a pipeline: hypotheses enter, evidence accumulates, and a small number of robust effects emerge. What the textbook omits is the ratio.

Over several years we tested several dozen distinct families of ideas against a fixed out-of-sample protocol: conditioning schemes drawn from market microstructure, alternative ranking constructions, regime filters, exit overlays of every conventional flavor, signals derived from the options market, and timing overlays motivated by the academic literature. Nearly all of them died. Some died loudly — inverting their sign out of sample. Most died quietly, indistinguishable from noise at any confidence level a referee would accept.

This paper is about the deaths: how to arrange them, how to record them, and why a well-kept record of them is among the most defensible assets a systematic research program produces. The framing is deliberately opposite to the industry's instinct. Practitioner writing overwhelmingly documents survivors — entries, exits, annotated equity curves — implying research processes with hit rates no honest program has achieved. Academic finance has a documented file-drawer problem; practitioner finance has a worse one, because no editor even nominally requires the drawer to be opened.

We make four contributions. First, a taxonomy of four recurring *shapes of death* (Section 5), each with a diagnostic signature computable from standard backtest output — to our knowledge the first attempt to organize backtest failure modes into a reusable classification. Second, a Monte Carlo quantification (Section 4) of *both error types* — size and power — across common research behaviors, on data that is synthetic by construction and therefore free to be reported in full; the experiments also demonstrate, uncomfortably, that a popular fold-consistency heuristic is near-vacuous once selection is accounted for. Third, a pre-registrable criterion — leave-best-era-out (Section 4.3) — that operationalizes the era-concentration diagnostic of Section 5.2 and whose measured trade-off between severity and power we report rather than hide. Fourth, a record-keeping layer (Section 6) — burial records, a graveyard, and a drawer for the uncomfortable category of ideas that validate statistically and are rejected anyway — including an operational trial-counting rule (Section 3.3) for the deflation arithmetic that the literature leaves undefined in practice.

A remark on what this paper withholds, and what it does not. No parameter values, universe definitions, or performance figures from our proprietary research appear here; figures describing our own pipeline are schematic. The Monte Carlo of Section 4, by contrast, is fully specified and reproducible — the discipline the paper recommends is priced into the paper itself.

2. The base rate has a literature

The mortality rate that surprises practitioners has been quantified by academics, and the quantification deserves to be better known outside the academy.

Harvey, Liu, and Zhu (2016) catalog the factor literature and conclude that, once the accumulated volume of testing across the profession is accounted for, the conventional t-statistic hurdle of two is far too permissive: a newly proposed factor should clear three or more, and by that standard a large fraction of the published record is likely false. Hou, Xue, and Zhang (2020) make the point with brute replication: of 452 published anomalies, 65% fail to clear $|t| \geq 1.96$ when microcaps are controlled with NYSE breakpoints and value weights, and 82% fail the multiple-testing hurdle of $|t| \geq 2.78$. Bailey, Borwein, López de Prado, and Zhu (2014) give the practitioner's version a sharper instrument: given enough configurations, a strategy with no skill can be tuned to an arbitrarily attractive backtest, and the probability of exactly this manufacture rises mechanically with the number of trials — the quantity backtest presentations omit. Their companion work formalizes the probability of backtest overfitting (Bailey et al., 2017); the deflated Sharpe ratio (Bailey and López de Prado, 2014) restates a backtested Sharpe after charging it for the trials

behind it, the non-normality of returns, and the sample length. Harvey and Liu (2015) offer an alternative "haircut" with the same moral. A Sharpe ratio unaccompanied by its trial count is not a statistic; it is an advertisement.

We note that this literature is contested. Jensen, Kelly, and Pedersen (2023) find high replication rates for factor research under consistent methodology and a Bayesian reading of the evidence, and Chen and Zimmermann (2022) reproduce the bulk of published cross-sectional predictors from open source. The disagreement is partly definitional — what counts as the same factor, the same universe, the same hurdle — and we do not adjudicate it here. Our claim needs only the uncontroversial core on which both sides agree: unpriced multiple testing inflates reported performance, and the inflation is largest precisely where trial counts are largest and least reported — which describes practitioner backtesting better than it describes the refereed literature either side is arguing about. McLean and Pontiff (2016) add the dynamic half of the picture: post-publication returns decay, so even genuine effects are wasting assets — which is why the recycling of old ideas under new names, a phenomenon our record-keeping is designed to price (Section 6), is not a harmless curiosity.

The implication that matters for practice: high mortality under a severe, fixed protocol is the *expected* outcome of honest testing, not evidence of a failing program. The converse claim — that high survival implies a weak program — requires care, and we defer it to Section 8.

3. A rejection-first protocol

Our response to the arithmetic above was institutional rather than statistical: a protocol whose architecture makes the seductive path — one more parameter nudge, one more filtered regime, one more "sanity re-run," each decided after seeing the data — expensive. The statistical components are standard and we compress their description accordingly; the protocol's ancestry runs from the data-snooping reality checks of White (2000) and Sullivan, Timmermann, and White (1999) through the machinery assembled in López de Prado (2018). Arnott, Harvey, and Markowitz (2019) propose a backtesting protocol in the same spirit — economic motivation first, multiple testing priced, research culture treated as the binding constraint. Our delta relative to their protocol is the subject of Sections 5 and 6: they specify how to test; we specify how to *record*, and what the accumulated record of failures is worth. The insistence on ex-ante commitment follows the preregistration argument made general by Nosek et al. (2018).

3.1 Units of analysis

Four terms are used precisely throughout. An *idea family* is an economic hypothesis ("overnight information is underpriced at the open"). A *variant* is a concrete parameterization of a family. A *candidate* is the variant put forward for validation — typically the variant that exploratory work suggested was best, which is exactly why it must be charged for its siblings. A *trial* is any variant whose performance was observed by a human decision-maker before the candidate was fixed. These distinctions are load-bearing: the deflation arithmetic of Section 3.3 operates on trials, the protocol evaluates candidates, and the graveyard of Section 6 buries families.

3.2 The statistical layers, briefly

Combinatorial folds. The sample is partitioned into temporal groups and every admissible assignment of training and testing groups is evaluated — combinatorially purged cross-validation (López de Prado, 2018, ch. 12). A candidate faces a distribution of paths rather than one path; performance concentrated in a single regime reveals itself as exactly that. Fig. 1 is schematic. A caution that Section 4 makes quantitative: the naive form of this idea — "profitable in most folds" — is defeated by selection and audits almost nothing; the operational value of fold-level output lies in calibrated criteria (placebo-referenced thresholds, and the leave-best-era-out rule of Section 4.3), not in raw fold-profitability counts.

Purging and embargo. Observations whose labels overlap fold boundaries are removed from training, with an embargo buffer against serial dependence (López de Prado, 2018, ch. 7). Label leakage across the train-test frontier is, in our experience, among the most common manufacturers of false confidence in short-horizon work — we state this as an impression from our own funerals, not as a measured ranking.

Ex-ante kill criteria. Before the first fold runs, the acceptance region is written down: the fraction of folds that must be profitable, the significance hurdle (set with Harvey-Liu-Zhu in view, i.e. materially above two), and the maximum overfitting probability tolerated. After the folds run, there is nothing to discuss. The criteria were chosen by people who did not yet know the answer — the only disinterested parties in the room.

Placebo controls. Randomized variants — scrambled orderings, shifted calendars, randomly overridden decisions — that preserve a candidate's mechanical structure while destroying its claimed information content. A candidate must separate from its own placebos; the placebo family also supplies the empirical null. Section 4's second experiment shows why this layer cannot be waived even when the others pass.

3.3 Deflation, and an operational trial-counting rule

Surviving statistics are deflated for the trials behind them (Bailey and López de Prado, 2014). The literature specifies the correction given the trial count but is largely silent on how to *count* — is a re-run under a new seed a trial? a variant abandoned after ten minutes? We use the following rule and propose it as a default:

> **A trial is any variant whose realized performance was observed before the candidate was frozen.** Seed re-runs of an unchanged variant do not increment the count (they estimate the same object); any change to features, parameters, filters, universe, or horizon does, *including variants abandoned early* — observation is the event, not effort. Exploratory work done before the family's pre-registration is counted by reconstruction from the research log, rounding up; work that cannot be reconstructed is bounded by the log's session count. The count is recorded in the family's burial or adoption record and is never revised downward.

The rule's virtue is not precision — reconstruction is approximate — but incentive-compatibility: because rounding is always upward and revision only upward, the researcher's laziness inflates the charge against their own candidate rather than the reverse. The arithmetic of Section 2 does not care which trials one would prefer to forget.

4. What the machinery is worth: a Monte Carlo of both error types

The claims of Section 3 can be evaluated on data where the truth is known by construction. We simulate families of K = 20 variants, each variant a series of T = 2,016 daily returns (eight years); the *candidate* is the variant with the best full-sample t-statistic, mirroring the researcher who submits the best thing exploration found. All rules are evaluated on the same families; analytic values are reported where closed forms exist, with simulation confirming them to within Monte Carlo error. Design details and dependence caveats are in Appendix B; the simulation code is public (Appendix B).

4.1 Size: the pure null

Zero-alpha variants (i.i.d. Gaussian). Every acceptance is false. Rates, analytic [simulated, n = 20,000]:

Rule	**False acceptance**
Full-sample tuning: accept candidate if t > 2	**36.9%** [37.0%]
Reused holdout: select best-of-K on the 30% holdout, accept if its t > 2	**36.9%** [36.8%]
Clean holdout: select on 70% train, evaluate once on untouched 30%, t > 2	**2.3%** [2.0%]
Fold consistency: candidate profitable in ≥ 80% of the 15 CPCV splits	— [**87.8%**]
Deflated hurdle: candidate t > to + 1.645, to from Bailey–López de Prado for K = 20	**0.4%** [0.4%]

Four observations. *First*, the top two rows are the same number, and this is an identity, not a coincidence: the t-statistic's null distribution does not depend on sample length, so selecting the best of K on a reused holdout is distributionally identical to selecting on the full sample — 1 − Φ(2)^K = 36.9%. **A holdout used for selection is not a holdout**; it is a smaller full sample. *Second*, the clean-holdout row shows that a disciplined single split already achieves nominal size. The popular narrative that "single splits overfit" is imprecise: *reuse* overfits; the split is innocent. What the deflated hurdle buys beyond a clean holdout is not size — both are honest — but the ability to keep testing after the holdout has been spent, which is the realistic condition of a living research program. *Third*, the fold-consistency rule is near-vacuous: the best-of-K candidate is mildly lucky *everywhere*, so it clears "profitable in most folds" 88% of the time under a pure null. A rule of this form, un-augmented, audits nothing. *Fourth*, the false-acceptance rate of naive tuning is increasing and unbounded in K: 10.9% at K = 5, 36.9% at K = 20, 90.0% at K = 100 — and real

exploratory research is closer to K = 100 than K = 20, making 36.9% a *lower bound* on the practice the first row describes.

4.2 The era-locked alternative

The second experiment is not a null: each variant receives a genuine drift confined to one randomly chosen sixth of the sample, scaled so a typical candidate shows a full-sample t ≈ 2.5 — the statistical silhouette of an edge that existed once, in one regime (Section 5.2). Deflation, asked its own question — "is this effect real in this sample, net of search?" — answers correctly: yes. The failure is in the *generalization* a practitioner would infer, so we report generalization-failure rates [n = 10,000]:

Rule	Era-locked acceptance
Naive t > 2	100%
Deflated hurdle	**94.9%**
Fold consistency (≥ 80% of splits)	95.5%
Deflation ∧ folds	91.7%
Clean holdout	2.5%
Leave-best-era-out (§4.3)	**1.2%**

The result that surprised us — and that we report *because* it is uncomfortable for the machinery this paper defends — is the fourth row: the deflation-plus-folds stack, the naive reading of "CPCV plus DSR," passes 92% of era-locked candidates. Selection contamination defeats the simple fold rule exactly as it did in Section 4.1: the best-of-K candidate is lucky everywhere, which drags its off-regime folds above zero. Meanwhile the humble clean holdout — temporally out-of-sample by construction — catches era-locking almost perfectly. The moral is precise: *deflation and fold-profitability rules answer "real, net of search?"; only genuinely out-of-sample evaluation answers "stable across time?"* — and these are different questions, neither reducible to the other.

4.3 Leave-best-era-out: a pre-registrable era diagnostic

The clean holdout cannot be reused, and a living program will reuse it. We therefore propose a criterion that captures the holdout's temporal honesty while remaining computable on the full sample: **leave-best-era-out (LBEO)** — drop the candidate's single best temporal group, recompute the deflated t-statistic on the remaining 5/6 of the sample, and require it to clear the same hurdle scaled by √(5/6). LBEO operationalizes the calendar-concentration signature of Section 5.2 as a pre-registered kill criterion: a candidate whose case rests on one era loses that era and dies; a temporally stable candidate loses only √(5/6) of its t.

Measured performance: null size **0.04%** (conservative — dropping the best group removes the selection winner's largest luck deposit); era-locked acceptance **1.2%** versus the standard stack's 92%. LBEO is what the fold-profitability rule wanted to be.

4.4 Power: what severity costs

Every rate so far measures Type I error; a protocol that rejects everything scores perfectly and is worthless. The same rules were run against temporally *stable* true alpha at annualized Sharpe ratios 0.5, 0.75, and 1.0 (implied full-sample t of 1.41, 2.12, 2.83). True-acceptance rates [n = 10,000]:

Ann. SR	Clean holdout	Deflated hurdle	LBEO
0.50	11.2%	**28.2%**	6.9%
0.75	20.4%	**80.5%**	36.0%
1.00	32.6%	**99.7%**	83.0%

Three honest readings. *First*, the deflated hurdle is respectably powered at SR ≥ 0.75 but detects a stable SR 0.5 only 28% of the time — and SR 0.5 is exactly the effect size most likely to survive costs in liquid markets. The protocol this paper defends is *underpowered precisely where genuine, durable edges live*, and eight years of daily data cannot fix that; only more data, or priors, can. *Second*, the reason deflation retains any power at low SR is itself a caution: when all K variants share the true alpha, selecting the best of them adds an expected ≈ 1.87σ of luck that nearly cancels the deflation charge. This cancellation is an artifact of independent variants. With realistically correlated variants (common factor, ρ = 0.8), the effective number of trials shrinks, the luck subsidy disappears, and power at SR 0.5 falls from 28% to **8.6%** (null size falls too, to 0.16% — deflation over-charges correlated trials, a known conservatism). *Third*, LBEO's severity has a measured price: roughly two-thirds of deflation's power at SR 0.5–0.75. A program adopting LBEO is buying insurance against era-locked false positives with detection probability on modest true edges, and should decide — ex ante, in writing — whether that trade matches its cost of each error type.

We would rather publish these numbers than the reassuring subset. A methods paper that reports only its Type I performance is a survivorship exercise about itself.

5. Case files: the shapes of death

Individual proprietary results stay home. The *shapes*, however, generalize. Four recur in our records often enough to deserve names, and each carries a diagnostic signature computable from standard fold-level output — the taxonomy is intended as a toolkit, not a museum.

5.1 The protective device that protects against recovery. Exit overlays — stops of every flavor — are the most reliably seductive family we test, because their in-sample story is visually irresistible: each stop, examined on the chart, visibly avoids some historical pain. Under the folds, every variant we tested — on our universe, at our horizons — died the same death: the pain avoided and the recovery forgone were frequently the *same episodes*, and the overlay attended the loss while missing the rebound. The theoretical literature says this is not a universal law but a process property: Kaminski and Lo (2014) show analytically that stop-loss rules destroy value under

random-walk returns yet earn a positive stopping premium under momentum and regime-switching processes. Our burials are therefore a *measurement*, not a refutation: the margins we studied evidently lack the return-process property that makes stops pay — which is precisely the kind of statement a graveyard, and only a graveyard, licenses. *Signature:* the overlay's per-fold benefit correlates negatively with the underlying's per-fold recovery speed; episodes truncated and episodes forgone share timestamps.

5.2 The patience that exists only in sample. Extending holding periods — "just hold longer" — improves nearly any backtest containing a long upward-drifting regime, and the improvement evaporates under combinatorial evaluation because it was a property of the sample's shape, not the signal. This is the most common death among ideas contributed by intelligent newcomers, and Experiment 2 of Section 4 is its idealized form. *Signature:* fold performance correlates with fold *calendar location* — regress per-fold return on fold start date; era-locked candidates show significant loading where robust candidates show none. Section 4.3's leave-best-era-out criterion is this signature promoted to a pre-registered kill rule, with its size and power measured.

5.3 The signal below the noise floor. A large class of candidates — often the most theoretically appealing — produce effects real in sign and negligible in magnitude: detectable with decades more data, indistinguishable from a coin flip at any horizon the program can act on. The literature that inspired such candidates typically established plausibility, not actionability. *Signature:* the candidate's fold-level statistics sit inside the interquartile range of its own placebo family's distribution; separation, not significance, is the failing test.

5.4 The improvement that is a rearrangement. Candidates that redistribute performance — across time, names, or market states — while leaving the aggregate untouched; in-sample presentations read these as risk reduction, the folds reveal a conservation law. The recurring cause is a mechanism that cannot, on economic grounds, create information — only reshuffle exposure to it. *Signature:* an offsetting diagnostic pair — every improvement in one pre-registered diagnostic is matched, within noise, by deterioration in another; the vector of diagnostic changes has near-zero projection on the program's objective.

What the four shapes share: each produces a legible, attractive, in-sample artifact, and each is diagnosed only by machinery — folds, placebos, deflation, pre-registered diagnostics — that the artifact's presentation omits.

6. The graveyard: record-keeping as infrastructure

The protocol produces verdicts. The verdicts would be worth little without the records.

6.1 The burial record

Every rejected family receives a burial record: a short structured document fixing what the idea was, why it was plausible, what configuration it faced, how it died, and the shape of failure it exemplifies (template in Appendix A). The discipline resembles a post-mortem in site reliability engineering: the case matters less than the maintained corpus. Three economic properties follow.

A validated no has a long shelf life. Markets recycle ideas; the same intuition returns every few quarters wearing different clothes — and post-publication decay (McLean and Pontiff, 2016) means even the genuine article returns weaker than its citation. A graveyard with good record-keeping converts each burial into a standing rejection: the second time an idea arrives, the cost of dismissal is a lookup, not a study. Over years this is the largest identifiable saving of research time in our program.

Negative results discipline the prior. Every failed conditioning scheme is a small measurement of how efficient the studied margin actually is. After enough funerals one carries a calibrated posterior — not a slogan — about where edge is unlikely to live, and research attention is the scarcest input the program has.

The graveyard is what makes survivors credible. A candidate that outlived dozens of siblings under an unchanging constitution carries different epistemic weight than a candidate born validated. Survivorship means nothing if nothing was allowed to die.

6.2 The drawer

One category surprises even sympathetic readers: ideas that *passed* the protocol and were rejected anyway — capacity too small to matter, operational surface rising faster than expected contribution, fragility to one identifiable structural change, redundancy with machinery already in place. These go to the drawer, with records noting that the mathematics said yes and the program said no, and why. Methodologically the drawer enforces the distinction between statistical and decision-theoretic acceptance; culturally it is the strongest evidence that the protocol is not a rubber stamp in either direction.

7. Why the funerals are hidden

If mortality is the modal outcome of honest research, why does public discourse contain almost none of it? The incentives are overdetermined. Survivors are legible and screenshot-able; funerals are not. Survivors accrue audience, clients, and employment. An author with a genuine edge has no incentive to publish it; an author without one has no incentive to publish failures; the observable corpus is therefore composed of noise presented as signal, or signal too small to survive costs, presented without the trial count that would price it — and the reader cannot distinguish the cases from presentation alone.

Duke (2018) names the behavioral half of the diagnosis: *resulting*, grading a decision by its outcome rather than the process that produced it. A reader calibrated on survivor-only writing is trained to result — to treat a green equity curve as evidence of good process and a funeral as evidence of bad — when in a noisy domain the mapping runs through the protocol or does not run at all. The pedagogical damage is real: newcomers whose thirtieth idea fails conclude they are bad at this, when they are experiencing, for the first time, a base rate their reading diet concealed. A funeral under a pre-registered protocol was a good decision to test and a good decision to kill.

The practical import is a due-diligence heuristic that costs nothing to apply: *ask for the graveyard.* A researcher, a fund, a published factor, a vendor model — the honest artifact is not the survivor but the documented population from which it survived. Absence of a graveyard is not proof of dishonesty; it is proof that the number being shown cannot be priced.

8. Discussion, limitations, and how this argument fails

Four limitations deserve statement, and stating them is not modesty — each marks a condition under which the paper's advice would mislead.

Mortality is a joint statistic. The temptation is to read high mortality as high rigor. It is not: mortality confounds protocol severity with idea quality, and a program with excellent hypothesis generation under a severe protocol could legitimately show high survival. What is diagnostic is not the rate but its decomposition: high survival is uninformative *unless the protocol's severity is independently verifiable* — pre-registered criteria, demonstrated placebo separation, an honest trial count. The correct slogan is not "distrust survivors" but "price the protocol before pricing the survivor."

The graveyard is gameable. Anyone can bury fifty deliberately bad ideas and present the corpse count as discipline. Graveyard credibility therefore requires *ex-ante plausibility of the buried* — evidence that the ideas were worth testing when tested: origin fields citing literature or observation, pre-registration timestamps, and the presence in the graveyard of ideas the researcher demonstrably favored (advocated for in writing before the verdict). The heuristic of Section 7 does not eliminate judgment; it relocates judgment from an unverifiable number to a partially verifiable record, which is an improvement, not a solution. Adverse selection survives all record-keeping; it is merely charged a higher price.

Falling mortality is ambiguous. Section 6.1 claims the graveyard disciplines the prior — implying idea quality, and therefore survival, should rise over time. A falling mortality rate is thus either learning or gaming, and the aggregate rate cannot distinguish them. The discriminator is *where* the survival improves: learning shows up as fewer submissions in death shapes already catalogued (the graveyard pricing repeat temptations to zero), while survival rising *within* previously fatal shapes, or under criteria revised after results, indicates the protocol is being negotiated with. Our records are structured (Appendix A's shape field) to make exactly this decomposition computable.

Scope. The protocol's severity is calibrated to our setting — a small program, long horizons of intended use, high cost of false adoption relative to false rejection. A program with cheaper reversibility might rationally accept more Type I risk; the constitution argument requires only that criteria be fixed ex ante, not that every program adopt ours. Severity also has a measured price: Section 4.4 shows the protocol detects a stable annualized Sharpe of 0.5 only about a quarter of the time on eight years of daily data — and materially less when trial variants are correlated — so a program adopting these hurdles is explicitly trading recall on modest durable edges for protection against false ones, and should make that trade in writing. And everything reported here is backtest

and simulation research; we make no claims about live implementation, and the framework neither requires nor implies any.

9. Conclusion

We keep the graveyard for the same reason a serious laboratory keeps its failed experiments: it is the record of contact with reality. The ideas were plausible; many were elegant; the evidence said no, and the evidence was *allowed* to say no — that last clause being the entire game.

The machinery is public and has been for years: combinatorial folds, purging, deflation, placebos, pre-registered criteria — and Section 4 quantifies both its value and the failure of any of its layers alone. What is scarce is not machinery but the willingness to live under it: to precommit to accepting rejections, to count the embarrassing trials, to write the funeral down where one's future self can find it, and to treat the resulting corpus, rather than any single survivor, as the product. A research program should be embarrassed by an empty graveyard, and suspicious of a full trophy case.

The referee is hostile. That is what we pay him for.

Appendix A. The burial-record template

Reproduced in full because its value is its brevity: a funeral that takes an hour to document will be documented; one that takes a day will not.

```
BURIAL RECORD #NNN
Family:           [one-line description of the idea]
Origin:           [literature / observation / recycled — link prior record]
Economic story:   [why this could have worked, two sentences]
Ex-ante case:     [written BEFORE testing — the §8 gameability defense]
Protocol config:  [criteria version — reference, not values]
Trial count:      [per §3.3 rule; never revised downward]
Verdict:          REJECTED — [statistical | placebo | deflation |
                   decision-theoretic (→ drawer)]
Death shape:      [protective-device | in-sample-patience | noise-floor |
                   rearrangement | other]
What died:        [the specific claim the evidence killed]
What survives:    [residual fact worth keeping — often "none"]
Resurrection ban: [what a future proposal must show to reopen this]
Date · analyst · time spent
```

Two fields do most of the compounding. *Death shape* turns cases into the maintained taxonomy that Section 5 reports. *Resurrection ban* converts "we tried that once" — the weakest sentence in research — into a falsifiable reopening condition, closing the question in every other state of the

world. *Ex-ante case* and *Trial count* exist to keep the record auditable against Section 8's failure modes.

Appendix B. Monte Carlo design

Common structure. Families of K = 20 variants; each variant a series of T = 2,016 i.i.d. $N(0, \sigma^2)$ daily returns, σ = 1%; the candidate is the variant with the best full-sample t-statistic. CPCV uses six temporal groups with test sets of two groups: C(6,2) = 15 splits. Replications: 20,000 families for the pure null, 10,000 for the era-locked, power, and correlated experiments; standard errors are binomial (≤ 0.4pp at the reported rates).

Hurdles. The deflation benchmark uses the Bailey–López de Prado closed form for the expected maximum of K standard normal deviates, $t_0 = (1-\gamma)\Phi^{-1}(1-1/K) + \gamma\Phi^{-1}(1-1/(Ke)) = 1.9007$ for K = 20 (γ the Euler–Mascheroni constant); the acceptance hurdle is $t_0 + 1.645$, a DSR > 0.95 analogue. Direct simulation of E[max] gives 1.867 (200,000 replications); the closed form is slightly stricter and is what we use. LBEO applies the same hurdle scaled by $\sqrt{(5/6)}$ to the t-statistic computed after deleting the candidate's best temporal group.

Alternatives. Era-locked: each variant receives a drift confined to one uniformly chosen group, scaled so a typical candidate's full-sample $t \approx 2.5$. Power: all variants share a constant drift equal to annualized SR ∈ {0.5, 0.75, 1.0} (full-sample t of 1.41, 2.12, 2.83). Correlated variants: returns share a common Gaussian factor with loading $\sqrt{\rho}$, ρ = 0.8.

Dependence caveats. The 15 CPCV splits are heavily dependent — each group appears in five test sets — so "n of 15 splits" has far fewer effective degrees of freedom than 15; we use the split counts only as a rule under test, never as independent evidence. We likewise do not perform path reconstruction (López de Prado, 2018, ch. 12, assembles the splits into φ = 5 complete backtest paths, which is the machinery's full form); our simplified rules are deliberately the *practitioner's* readings of CPCV, and Section 4's negative results about them should not be read as results about path-level CPCV inference. Purging is omitted because i.i.d. returns have no label overlap by construction; in applied settings it is not optional.

Analytic checks. $1 - \Phi(2)^K = 36.89\%$ (naive and reused-holdout null); $1 - \Phi(2) = 2.28\%$ (clean holdout); $1 - \Phi(3.5457)^K = 0.39\%$ (deflated null). Simulated values match within Monte Carlo error throughout.

Code. Single-file Python/NumPy, no dependencies beyond NumPy, fixed seed: https://alexvidovich.com/paper/monte-carlo-v2.py — reproduces every number in Section 4 in a few minutes on a laptop.

References

Arnott, R., Harvey, C.R., Markowitz, H., 2019. A backtesting protocol in the era of machine learning. *Journal of Financial Data Science* 1 (1), 64–74.

Bailey, D.H., Borwein, J.M., López de Prado, M., Zhu, Q.J., 2014. Pseudo-mathematics and financial charlatanism: the effects of backtest overfitting on out-of-sample performance. *Notices of the American Mathematical Society* 61, 458–471.

Bailey, D.H., Borwein, J.M., López de Prado, M., Zhu, Q.J., 2017. The probability of backtest overfitting. *Journal of Computational Finance* 20 (4), 39–69.

Bailey, D.H., López de Prado, M., 2014. The deflated Sharpe ratio: correcting for selection bias, backtest overfitting, and non-normality. *Journal of Portfolio Management* 40 (5), 94–107.

Chen, A.Y., Zimmermann, T., 2022. Open source cross-sectional asset pricing. *Critical Finance Review* 11 (2), 207–264.

Duke, A., 2018. *Thinking in Bets: Making Smarter Decisions When You Don't Have All the Facts*. Portfolio/Penguin, New York.

Harvey, C.R., Liu, Y., 2015. Backtesting. *Journal of Portfolio Management* 42 (1), 13–28.

Harvey, C.R., Liu, Y., Zhu, H., 2016. …and the cross-section of expected returns. *Review of Financial Studies* 29, 5–68.

Hou, K., Xue, C., Zhang, L., 2020. Replicating anomalies. *Review of Financial Studies* 33 (5), 2019–2133.

Jensen, T.I., Kelly, B., Pedersen, L.H., 2023. Is there a replication crisis in finance? *Journal of Finance* 78 (5), 2465–2518.

Kaminski, K.M., Lo, A.W., 2014. When do stop-loss rules stop losses? *Journal of Financial Markets* 18, 234–254.

López de Prado, M., 2018. *Advances in Financial Machine Learning*. Wiley, Hoboken.

McLean, R.D., Pontiff, J., 2016. Does academic research destroy stock return predictability? *Journal of Finance* 71 (1), 5–32.

Nosek, B.A., Ebersole, C.R., DeHaven, A.C., Mellor, D.T., 2018. The preregistration revolution. *Proceedings of the National Academy of Sciences* 115 (11), 2600–2606.

Sullivan, R., Timmermann, A., White, H., 1999. Data-snooping, technical trading rule performance, and the bootstrap. *Journal of Finance* 54 (5), 1647–1691.

White, H., 2000. A reality check for data snooping. *Econometrica* 68 (5), 1097–1126.