An Exploratory Study of Random-Day Placebos Under Persistent Conditioning: Calibration and Non-Overlapping Event Retention in Single-Stock Conditional Event Studies

Peilun Cai
ORCID: 0009-0007-5364-6883

October 30, 2026
Working paper v1

Abstract

Backtests are difficult to assess when specification searches and overlapping events are not disclosed. We examine a fixed single-stock rule that calls a direction at a random-day percentile of $\geq .95$ or $\leq .05$, nominally two-sided 10%. Across 124 stocks, circular shifts of persistent condition masks by at least 252 trading days yield a random-day rejection rate of 0.314 (95% CI 0.302–0.324), compared with 0.102 for a shift-based placebo. Thinning events to at least 20 trading days apart retains a median 0.366 of the nominal sample (95% CI 0.356–0.376). Most initially directional readings lose their direction after thinning. This does not establish that genuine return effects disappear: thinning removes information, and the shift placebo shares the mechanism used to construct the null. General false-positive control remains unverified. The stock pool is selected, thinned events may remain dependent, and stock-resampling intervals have unverified coverage under a shared calendar. This retrospective exploratory study was not prospectively preregistered. Evaluation requires the search history and event-overlap accounting; neither is replaced by a small reported p-value. A planned simulation comparison will test calibration and power against established methods.

Keywords: placebo calibration; conditional event studies; overlapping returns; circular shifts; exploratory research.
JEL classifications: C12, C15, G11, G17.

1 Introduction

A backtest that reports a successful rule without its search history leaves readers unable to assess whether the result was selected from many unsuccessful specifications. Overlapping events create a separate problem: the reported event count can overstate the amount of distinct information. These concerns motivate data-snooping and multiple-testing procedures (White, 2000; Harvey et al., 2016). We do not estimate how prevalent selective reporting is in public trading-signal research. We examine the overlap problem in a specific evaluation procedure, while retaining the limits imposed by its incomplete historical search record.

The procedure encodes today's stock condition as a fingerprint, finds historical days with the same fingerprint, and compares their average return over the next 20 trading days with averages from randomly drawn days. A fingerprint average in the top or bottom 5% of the random-day distribution is classified as directional.

This is the third layer of the evaluated system. Its fingerprint combines the GMMA trend side, price relative to the quarter-anchored VWAP, 60-day strength relative to SPY, and a volume z-score of at least 1. The placebo draws 400 random sets with the same number of days. The study snapshot contains reports using this procedure.

Fingerprint conditions can persist for weeks, so qualifying dates cluster. Adjacent 20-day forward windows share 19 of their 20 daily returns. Independently sampled random days preserve neither the observed runs of qualifying dates nor their exact overlap. This comparison may therefore assign a directional reading too often under a persistence-preserving reference procedure. Overlapping observations and clustered events are established inference concerns (Hansen and Hodrick, 1980; Newey and West, 1987; Brown and Warner, 1985).

The local analysis plan recorded two questions. Its timing cannot be independently established, so the study is retrospective and exploratory (Section 6.1):

- **Q1.** Under circular shifts that preserve the real fingerprint's run structure, how often does the random-day rule ($p \geq .95$ or $p \leq .05$) reject?
- **Q2.** What share of the nominal fingerprint sample remains when forward-return windows are required not to overlap?

Circular shifts provide the reference procedure for Q1; they are not proven to represent every economically relevant null (Section 6). The plan called the Q2 count "effective independent events" and described H2 as overlap at least halving the effective sample. We retain the quantity, the 0.5 threshold and the test, but call it a non-overlapping event count because non-overlap does not establish independence (Review Response R1).

The distinction between classifications and economic effects is central to the interpretation. The study measures rejection frequencies under a specified shift procedure and changes in directional classifications after thinning. It estimates no signal returns, and cannot identify which classifications correspond to genuine effects.

A local analysis plan exists, but its external timestamp was obtained only after the experiment.

The exploratory contributions are:

- A comparison of fixed-rule rejection frequencies under a persistent-mask shift procedure.
- An account of how event thinning changes sample retention and directional classifications.

These diagnostics do not establish a new method, predictive performance or a complete historical search denominator.

2 Related Literature

Overlapping return windows and clustered event dates motivate the dependence concern in this study (Hansen and Hodrick, 1980; Newey and West, 1987; Brown and Warner, 1985). Block and stationary bootstraps preserve aspects of serial dependence when constructing sampling distributions (Künsch, 1989; Politis and Romano, 1994). The circular-shift comparison used here instead moves the observed condition mask relative to the price path. Its calibration under its own shift mechanism does not establish validity under all economically relevant nulls.

The Reality Check of White (2000) addresses whether the best model found in a specification search outperforms a benchmark. The superior predictive ability test of Hansen (2005) modifies

that comparison using studentization and a sample-dependent null distribution to reduce sensitivity to poor alternatives. Romano and Wolf (2005) develop stepdown multiple-testing procedures with family-wise error control. Those questions concern a family of candidate models or hypotheses; the historical rejection rate reported here is for each fixed final fingerprint and excludes a search over features or thresholds.

Harvey et al. (2016) examine multiple testing in the cross-section of expected returns. The deflated Sharpe ratio of Bailey and López de Prado (2014) addresses selection bias and non-normal returns in performance evaluation. This paper estimates no strategy returns or Sharpe ratios. It does not run a common-design size-and-power comparison against the Reality Check, superior predictive ability, Romano–Wolf procedures, block bootstraps or the deflated Sharpe ratio, and it makes no superiority claim. These references locate the scope of the diagnostic; they do not supply missing validation.

3 Data

Sample. The snapshot holds 164 non-SPY symbols from the primary snapshot of commercial market data (not redistributed). For each we take the parquet file with the latest basis date and require a SPY file for the same date, at least 2,000 bars, and at least 30 retained Tier A days. The 32 files from a different commercial source are ignored. Data checks reject non-increasing indexes, missing OHLCV and non-positive closes; none failed. Soft flags (daily moves over 50%, zero-volume days) appear for 12 stocks; they are listed in `results.json` and exclude none.

Detailed input and output hashes, the reproduction boundary and code availability are recorded in Appendix B. The publication contains derived statistics only; no price series or downloadable datasets are provided.

4 Method

4.1 Definitions and notation

Let T identify a stock, n_T its number of daily observations, and $C_{T,t}$ its adjusted close at trading-day index t. The daily dividend-adjusted bars come from commercial market data (not redistributed); SPY bars with the same basis date serve as benchmark. The Tier A fingerprint mask is

$$M_T = (M_{T,t})_{t=0}^{n_T-1} \in \{0,1\}^{n_T}. \tag{1}$$

It is built by the engine's `features()` and the same logic as `compute()` lines 490–498. The direction of each leg is the state on the last bar. Missing values count as 0.

$$r_{T,t}^{(20)} = \frac{C_{T,t+20}}{C_{T,t}} - 1, \qquad t < n_T - 21. \tag{2}$$

For a mask M, let $\mathcal{I}_T(M)$ be its retained event indices. For a nonempty retained sample, the event count and mean forward return are

$$\mathcal{I}_T(M) = \{t : M_t = 1,\ 0 \le t < n_T - 21\}, \quad n(M) = |\mathcal{I}_T(M)|, \quad S_T(M) = \frac{1}{n(M)} \sum_{t \in \mathcal{I}_T(M)} r_{T,t}^{(20)}. \tag{3}$$

For each of 400 random-day draws, A_b is a set of $n(M)$ days sampled without replacement from $\mathcal{P}_T = \{60, \ldots, n_T - 22\}$. The engine's `placebo()` percentile is

$$p_{\mathrm{RD}}(M) = \frac{1}{400} \sum_{b=1}^{400} \mathbf{1}\left\{ \frac{1}{n(M)} \sum_{t \in A_b} r_{T,t}^{(20)} < S_T(M) \right\}. \tag{4}$$

The engine's `rotation_placebo()` uses 400 circular shifts of the mask. With independent draws v_b from the discrete uniform distribution on the stated integer set,

$$\begin{aligned} v_b &\sim \mathrm{Unif}\{252, \ldots, n_T - 253\}, \\ p_{\mathrm{SH}}(M) &= \frac{1}{400} \sum_{b=1}^{400} \mathbf{1}\{S_T(\mathrm{roll}(M, v_b)) < S_T(M)\}. \end{aligned} \tag{5}$$

Both percentiles use a strict less-than comparison. The engine's layer-3 rejection indicator is

$$D(p) = \mathbf{1}\{p \geq .95 \text{ or } p \leq .05\}, \tag{6}$$

a nominal two-sided 10% rule. The outer calibration procedure draws null masks

$$M_T^{(r)} = \mathrm{roll}(M_T, u_r), \qquad u_r \sim \mathrm{Unif}\{252, \ldots, n_T - 252\}. \tag{7}$$

Here roll is a circular shift with wrap-around at the ends. It keeps the density and run structure of the real mask and moves it at least 252 bars from its real dates. "Under the null" below is shorthand for "under this shift procedure". The endpoints in equations (5) and (7) intentionally differ, matching the frozen engine and analysis wrapper.

For non-overlapping event retention, walk $\mathcal{I}_T(M)$ in time order, keeping the first day and then a day only if it is at least 20 trading days after the previously kept day. If g_j denotes the retained sequence,

$$\begin{aligned} g_1 &= \min \mathcal{I}_T(M), \qquad g_{j+1} = \min\{t \in \mathcal{I}_T(M) : t \geq g_j + 20\}, \\ n_{\mathrm{ind}}(M) &= |\{g_j\}|, \qquad \rho_T = \frac{n_{\mathrm{ind}}(M_T)}{n(M_T)}. \end{aligned} \tag{8}$$

The recursion stops when the next set is empty. Greedy selection gives the largest set of retained days whose 20-day windows do not overlap. Those days are not thereby independent, because returns more than 20 days apart can still be correlated.

Let $R = 200$ be the number of null masks per stock, $\mathcal{T}$ the included stock set and $K = |\mathcal{T}|$. For $a \in \{\mathrm{RD}, \mathrm{SH}\}$,

$$\pi_{a,T} = \frac{1}{R} \sum_{r=1}^{R} D\left(p_a(M_T^{(r)})\right), \qquad \pi_a = \frac{1}{K} \sum_{T \in \mathcal{T}} \pi_{a,T}, \qquad \widetilde{\rho} = \mathrm{median}_{T \in \mathcal{T}}\, \rho_T. \tag{9}$$

The symbols π_{RD} and π_{SH} denote equal-weighted means across stocks; $\widetilde{\rho}$ is the median non-overlapping share. Indicator $\mathbf{1}\{\cdot\}$ equals one when its condition is true and zero otherwise.

4.2 Calibration and inference

Null calibration. For each included stock we draw $R = 200$ shifted masks. For each one we compute p_{RD} and p_{SH} with the engine functions, unchanged, and record whether each rule rejects.

For the full sample, the seed key is `SeedSequence` with seed key $(20261022, i_T, r)$, where i_T is the stock index and r the replicate index; append 1 for the first half or 2 for the second. Its three spawned child streams generate the null shift, random-day draws and shift-placebo draws, respectively.

Non-overlapping events. n_{ind} and ρ_T come from the real mask. We use greedy thinning because it is transparent and has no tuning parameter. It counts events whose forward windows do not overlap and is not an effective sample size. Estimating an effective sample size would need a variance-based method such as HAC variance (Newey and West, 1987) or block or stationary bootstraps (Künsch, 1989; Politis and Romano, 1994); we did not compute one.

Inference. We resample stocks with replacement with $B = 2000$ and seed 20261022 and take percentile 95% intervals. The original analysis plan specified a family of two one-sided tests, H1a: $\pi_{\text{RD}} > 0.10$ and H2: $\widetilde{\rho} < 0.5$. Holm's procedure (Holm, 1979) controls family-wise error at 0.05. The one-sided p of PREREG §5 is the share of bootstrap replicates on the null side of the threshold. It is a Monte Carlo quantity with resolution $1/B$; Section 5 shows how a count of zero is handled. Resampling stocks treats them as independent clusters. They share one market calendar, so the intervals and p-values are conditional on that assumption; their actual coverage and size were not checked.

Pre-specified checks outside the family. (i) H1b: the 90% interval of π_{SH} must lie inside [0.07, 0.13]. The null is itself a circular shift, so this mostly confirms the wrapper code is correct, and failing it stops the analysis. (ii) Setting the shift to zero must reproduce the engine's stored `placebo_pct` for every facts file whose inputs are in the snapshot.

In-sample / out-of-sample. Each stock is split at $\lfloor n/2 \rfloor$. Features are computed on the full series and then cut. Forward windows do not cross the cut, and halves shorter than 1,000 bars are skipped. The original plan required a hypothesis to pass Holm on the full sample and have point estimates in both halves on the same side of the threshold. Leg directions come from the last bar and eligibility from the full series, both fixed before the split, so the halves are a consistency check across time specified in the local plan, not independent out-of-sample tests.

5 Results

Table 1: Sample. Counts of symbols and exclusions.

	Stocks
Non-SPY symbols in primary snapshot	164
Excluded: no SPY file for latest basis date	14
Excluded: fewer than 2,000 bars	20
Excluded: fewer than 30 retained Tier A days	6
Excluded: ETF (symbol inspection, deviation 2) / data checks	0 / 0
Included	**124**

Source: `out/results.json` (October 22, 2026).

Null replicates: 24800 (124 stocks × 200). Median retained Tier A days per stock: 172; median non-overlapping events: 59.

Table 2: Main estimates. Equal weight per stock, stock-clustered percentile bootstrap. The intervals treat stocks as independent; with a shared calendar their coverage is not established (Section 6). "Under null" means under the shift procedure of Section 4.

Quantity	Full sample	95% CI	IS (first half)	OOS (second half)
π_{RD}, random-day reject rate under null (nominal 0.10)	0.314	0.302 – 0.324	0.294	0.312
π_{SH}, shift-placebo reject rate under null (nominal 0.10)	0.102	0.098 – 0.105	0.093	0.093
$\widetilde{\rho}$, median n_{ind}/n (non-overlapping share)	0.366	0.356 – 0.376	0.369	0.367

Source: `out/results.json` (October 22, 2026).

IS and OOS used 124 and 124 stocks; no half was under 1,000 bars.

Table 3: Tests specified in the local analysis plan (exploratory).

Hypothesis	Bootstrap replicates on the null side	95% upper bound of the tail probability	Holm-adjusted bound	Full sample	IS / OOS same side	Original plan rule (exploratory; see Limitations)
H1a: $\pi_{\mathrm{RD}} > 0.10$	0 of 2000	0.0015	0.0030	pass	yes (0.294, 0.312)	meets original rule
H2: $\widetilde{\rho} < 0.5$	0 of 2000	0.0015	0.0030	pass	yes (0.369, 0.367)	meets original rule

Source: `out/table3_input_provenance.json` (prepared October 29, 2026); `out/results.json` (October 22, 2026) and `out/review_checks.json` (October 23, 2026).

PREREG §5 defines each p as the share of the 2000 bootstrap replicates that fall on the null side of the threshold, so it moves in steps of $1/B = 0.0005$. A count of zero shows that no replicate crossed; it does not show that the underlying tail probability is below $1/B$. Treating the count as binomial, the one-sided 95% upper bound for the tail probability is 0.0015 for H1a and 0.0015 for H2; after Holm's step the bounds are 0.0030 and 0.0030. All are below 0.05, so the Holm decisions are the same whether one uses the share or the bound. The bounds cover simulation error only. They assume that resampling stocks is a valid model of sampling variation, which Section 6 does not establish. These bounds were added after review and are not preregistered.

Check H1b. The 90% interval for π_{SH} is 0.099–0.105, inside [0.07, 0.13]. This full-sample output passed the numerical check. It does not establish compliance with the PREREG §9 stop rule in the earlier smoke/full-run sequence; see Deviations and limitations.

Reading. Under the specified mask-shift procedure, the random-day rule calls a direction 0.314 of the time against a nominal rate of 0.10. The excess is 0.214, about 3.1 times nominal, with similar rates in both temporal halves. Per-stock π_{RD} ranges from 0.105 to 0.440; 124 of 124 stocks have a point estimate above 0.10. This description does not depend on stock resampling, but each estimate has Monte Carlo error from 200 replicates, and the count is not a test. A separate-seed rederivation put 123 of 124 above 0.10, illustrating that sensitivity. Across the 24800 null p_{RD} values, 0.211 are below 0.1 and 0.213 are at or above 0.9; a uniform percentile would place 0.1 in each tail. The decile counts are 5243, 2082, 1855, 1637, 1562, 1638, 1616, 1775, 2107 and 5285.

The shift-placebo rejection rate is 0.102. Its proximity to nominal under this procedure does not establish general false-positive control: the placebo and the null share the same shift mechanism. The median stock retains 0.366 of its nominal Tier A dates after requiring non-overlapping windows, 0.134 below the 0.5 threshold. This is an event-retention ratio, not an effective sample size.

The real-mask thinning comparison in E2 answers a different question. Of 43 initially directional readings, 37 lose their direction; 1 flips sign and 5 retain it. Another 5 stocks gain a direction, leaving 11 directional readings after thinning. Thus most initially directional readings disappear after thinning in this sample. The results do not show that most genuine return effects disappear, or that combining shifting and thinning produces a validated test. Thinning discards information, and no returns or independently known truth are used to adjudicate the lost and retained readings. The evidence supports a narrower conclusion: the evaluated random-day classifications are sensitive to event dependence, and its nominal rejection rate is not reproduced under this particular persistent-mask reference procedure.

Source for the separate-seed rederivation: `out/rederive_separate_seed_comparison.json`; processing `code/rederive_independent.py`. The rederivation is same-family and is not a substitute for independent cross-family QA.

5.1 Additional exploratory analyses

- **E1. Published reports.** Of 20 non-superseded facts files in the snapshot, 2 had a Tier A random-day percentile at or beyond the .95/.05 cut. The local session record says the two threshold-crossing values were first inspected after the claimed lock; the record cannot independently establish that timing. This is a description only.
- **E2. Non-overlapping recomputation.** For each included stock we thinned the real mask to non-overlapping events and recomputed p_{RD} with the engine's default seed. The rule gave a direction for 43 stocks on the full mask and for 11 on the thinned mask. Of the 43, 37 lost the direction, 1 flipped sign, and 5 kept it; 5 stocks gained a direction after thinning. Thinning also discards information, so this does not show which reading is right.
- **E5. Clustering and miscalibration (not preregistered).** Across stocks, the Spearman correlation between $\pi_{\mathrm{RD},T}$ and ρ_T is -0.62. Stocks whose fingerprint days cluster more also reject more often under the null. This is a correlation, not a tested effect; a separate-seed rederivation gave -0.603, showing sensitivity to the Monte Carlo draws.
- **E3.** Not done in this draft.
- **E4. Realised outcomes of published reports.** Not analysed. The 20-day windows overlap almost entirely in calendar time and amount to roughly one draw of the market, so they cannot support inference.

6 Deviations and Limitations

6.1 Analysis-plan timing

This study was not prospectively preregistered.

The original timestamp is retrospective and was obtained after the original results.

The proof establishes that the plan hash existed by the block time; it does not establish when the study was planned or that the results were unchanged.

- Plan: `PREREG.md`, SHA-256 `431789bdc78ee6952c331e9df8c669e22901d434b4f76b8d072dbb91378e647a`, locked October 22, 2026 at 06:10:14 UTC by the local clock. Three local records made at the time support the order lock, then code, then results. (a) The transcript of the session that wrote the plan contains the write at 06:10:06 UTC and one wording edit at 06:10:13 UTC. Replaying both reproduces the locked hash, and the session printed that hash at 06:10:14 UTC. (b) At 06:14:05 UTC the company job tracker, a separate script, wrote into its hash-chained event log the SHA-256 of an inventory document that quotes the PREREG hash. Replaying that document from the same transcript gives the logged hash, and the chain is unbroken through all later entries. The document on disk has been edited since, so only its replayed 06:14:05 UTC state matches the log. (c) The analysis code was first written at 11:03:25 UTC and first run at 11:03:32 UTC, in a different session. All three records sit on the author's own machine, and the same user could in principle rewrite them. A later OpenTimestamps proof binds the plan hash to Bitcoin block 973207 at October 26, 2026 at 07:31:02 UTC, after the run. Verification uses two cached public header sources, not a locally validated full node. It does not establish that the plan was externally registered before the experiment. A third party therefore cannot verify the order; the study is therefore reported as exploratory and retrospective (receipts: review appendix `verify_lock_receipts.py`, `lock_receipts.json`).
- Data snapshot manifest SHA-256 `7487c62691abff9697c624590d450438aecd2afa49894c9d51e9ae673623596b`. Engine file SHA-256 `8a22ea2bfd6c6289cd8ceea67d4037b979fbe16f9ce8bfcf478844e4015aeadb`, checked by the script before any computation.
- Before locking, the author had seen the Tier A percentiles printed in published reports, a summary of one stock's facts file, and a 5-report outcome ledger through September 14, 2026 (PREREG §0); this account has not been independently verified. No null simulation or non-overlap calculation had been run. Before the lock, the plan session's own tool calls were file reads, listings, copies into the snapshot and a text check of one published report. One of its three read-only search subagent logs survives and shows only reads of other studies' files; the other two logs are not on disk.
- The mechanical check (PREREG §7) ran on 10 of 23 facts files. The others lack the matching price or SPY file in the snapshot. All 10 reproduced Tier A n, bar count, leg directions and `placebo_pct` exactly.

Deviations from the PREREG (full text in `DEVIATIONS.md`; none changes a threshold or test, and the one deviation in how the sample rule was applied, item 2, excluded nothing):

1. The engine module cannot be imported because two of its dependencies are missing from the snapshot. The five needed functions are AST-extracted from the hash-checked file and run unchanged.
2. The snapshot has no company profiles, so before the run the ETF screen was done by symbol inspection rather than by the profile `is_etf` field that PREREG §3 names. 0 ETFs were excluded. A check added after review (item 12) found 0 `is_etf` flags among the symbols.
3. The PREREG did not specify how seeds split across the shift, random-day and shift-placebo draws, or what seeds the halves use. Both are now fixed in the code.
4. A null mask with no retained days would count as not rejected, per the engine's `layers()`. This occurred 0 times.
5. Hard data checks were added; they excluded 0 stocks.
6. One smoke run (2 stocks, $R = 20$) was made before the full run. Its output files were deleted. The run session's transcript keeps the printed tail of that output and shows no code edit between the smoke run and the full run.

7. The original outputs are in `results/` per PREREG §10; this review package carries copies in `out/`.
8. The full run was repeated once after CSV columns were added. The code change touches only the CSV writer, and summary statistics were identical.
9. Exploratory item E3 and both figures are not in this draft.
10. In the halves, the random-day pool starts at bar 60 of the half.
11. (added after review, October 23, 2026; not preregistered) PREREG Q2 and H2 call n_{ind} the "effective independent events" and H2 "overlap at least halves the effective sample". That label is withdrawn. $\widetilde{\rho}$, the 0.5 threshold and the test are unchanged.
12. (added after review, October 23, 2026; not preregistered) The evaluation pipeline's existing company profiles, which sit outside the snapshot, were read to check the ETF rule. 164 of 164 symbols have a profile, 0 carry `is_etf` true, 0 have a recorded file modification time after the local lock, and 0 included symbols lack one. File modification times are not independent evidence of when the profiles were last changed; their state at the lock is unknown. The sample is unchanged.
13. (added after review, October 23, 2026; not preregistered) The bootstrap p of PREREG §5 is reported as a count of replicates with its Monte Carlo resolution and upper bound, and Holm's step is shown on the bound. The test decisions are unchanged.

6.2 Stopping rule and incomplete trial record

The review ledger catalogues 28 documented groups of analysis and development attempts, including failed and missing-output attempts (internal review records: `out/documented_attempts.json` and `appendix/TRIAL_LEDGER_SNAPSHOT.md`). Some groups contain several attempts; this is not a count of independent experiments or an independently established exhaustive trial denominator. Full smoke-run outputs and some earlier development artifacts remain unavailable.

The archived smoke procedure was replayed on the original frozen snapshot using the archived code and numerical-library versions recorded for the full run. All surviving printed summary fields matched. The replay used two stocks with twenty null replicates per stock. Its H1b calibration check failed: the 90% interval was [0.10, 0.15], outside the required [0.07, 0.13]. The original stopping rule contains no smoke-run exemption and requires stopping without writing results when this check fails. The replay printed the warning but still wrote results and returned success, contrary to that rule. The draft was subsequently continued, and this deviation is retained here. The original smoke files and runtime metadata are unavailable, so the replay does not establish their byte identity or what the author knew at the time. The violation does not by itself overturn the full-sample numerical estimates, which remain unchanged; the later full-sample pass also cannot establish that the earlier sequence complied with the stopping rule. This paper remains exploratory and makes no methodological-breakthrough claim.

The missing original smoke-run files are identified as T01 in the review ledger. The replay is not a replacement for those missing historical artifacts.

6.3 Scope of inference

- The stock pool consists of names the evaluation pipeline downloaded on its basis dates, screened by triggers and turnover. Across seven dates represented in the snapshot, retained primary-source price-file counts range from 3 to 42, with a median of 40; the snapshot does not establish complete daily download totals. The pool is not a random sample, and the

results apply to this pool.

- Selection of the stock pool may introduce survivorship bias; its extent is unknown. No trading returns are estimated here.
- All stocks share market-wide shocks over the same calendar. Resampling stocks treats them as independent clusters, so the 95% intervals and the bootstrap p-values are conditional on that assumption. Their coverage and size under shared-calendar dependence were not checked and are unknown. The per-stock description in Section 5 does not rely on this assumption, but it is not a test.
- Leg directions come from each stock's state on its basis date, so each stock tests a different condition. The object of study is the procedure, not a common signal.
- Only the 20-day horizon is calibrated. The 1-, 5- and 10-day means and win rates shown in the reports have no placebo, and this paper does not address them.
- The null moves the real mask by 252 to $n - 252$ bars with wrap-around, and 0.314 is the random-day rule's rejection rate under this procedure. We have not shown that shifted masks are exchangeable with the real mask under every null in which the fingerprint carries no information about future returns. Dependence in volatility or drift that lasts longer than 252 bars, or the seam where the mask wraps around, could make them differ. Whether the rule's error rate under other such nulls equals 0.314 is unknown. The reported rejection rate evaluates each final fixed fingerprint. The null replicates did not repeat a search over features or thresholds, so the rate does not estimate the chance that any searched rule crosses the decision threshold.
- Greedy thinning counts non-overlapping events. It is not an effective sample size, because events 20 or more trading days apart can still be correlated. HAC- or block-based methods would be needed to estimate an effective sample size and were not run.
- The halves split each stock's history after leg directions (last bar) and eligibility (full series) were fixed, so IS and OOS are a consistency check across time, not independent out-of-sample validation.
- The analysis uses the primary snapshot of commercial market data (not redistributed) only. The snapshot also contains 32 files from a different commercial source, which were excluded.
- The claimed pre-experiment PREREG lock rests on local records (Section 6.1); the external proof postdates the experiment. The author works at the organisation that built the evaluation system.
- The complete trial record has not been independently established. The first full run's JSON and per-stock CSV have been reconstructed byte-for-byte to their previously recorded local hashes; the complete smoke-run output remains unavailable. This recovery neither authenticates the local chronology nor establishes that all attempts were recorded.
- The publication contains only derived statistics, with no individual price series or downloadable datasets. External reproduction requires the same original snapshot, which is not distributed. Publication follows a founder risk-acceptance decision; this does not assert contractual permission from a data provider.

- The random-day percentile uses a strict less-than comparison. If all observed and placebo returns tie, the percentile is zero and the rule calls a direction even in a constant-price synthetic fixture. An author-side exact replay found no such ties in the original full-sample random-day null arm; it does not cover the split samples, other trials, or near ties (out/tie_audit_summary.json).

7 Conclusion

The total number of specifications tried and the overlap among events are necessary inputs to evaluating a reported backtest. Without them, a reader cannot assess the scope of selection or the amount of distinct event information; the reported significance alone is insufficient. These disclosures are not themselves proof that a result is valid. In this study, the complete historical trial denominator remains unestablished.

The existing results identify sensitivity in a particular single-stock evaluation procedure. Its random-day rule rejects more often than nominal under persistent-mask shifts, while the shift placebo stays near nominal under a null built from that same mechanism. Most initially directional real-mask readings lose their direction after non-overlapping-event thinning. Neither comparison shows that genuine return effects have disappeared, and neither establishes the shift method's general calibration or power. The results remain exploratory and do not estimate an investable return.

On October 29, 2026 we drafted a separate plan for a prospective confirmatory extension. The final pooled-panel protocol has now passed the assigned independent protocol review and has been locked with its software, cohort, source semantics and resource budget. OpenTimestamps calendar receipts have been obtained for the exact protocol and registration manifest. Bitcoin confirmation remains pending, so retained-data collection cannot begin. The locked design measures an algorithmic rejection probability conditional on a future panel; it is not a direct numerical replication of the historical estimate or evidence of market-wide calibration. It evaluates the full declared rule search on each null replicate and reports sensitivity to multiple minimum shift lengths. The present results remain exploratory.

The v2 methodological extension is planned to compare mask-shift testing with White's Reality Check and Romano–Wolf procedures using simulated data whose null and alternative generating processes are known. The comparison will measure false-positive rates and power under common settings; it will test whether the shift method is calibrated and whether its power is higher, rather than assume either result. This simulation comparison is additional to the locked pooled-panel protocol and must be specified before the comparison is run. No such comparative result is reported here.

Declaration of Generative AI and AI-Assisted Technologies in the Writing Process

During the preparation of this work, the author used Claude (Anthropic) and Codex (OpenAI) for code development, drafting assistance and language editing. The author designed the study, determined the principal arguments and interpretations, reviewed and revised all content, verified all data and citations, and takes full responsibility for the content of this article.

References

David H. Bailey and Marcos López de Prado. The deflated Sharpe ratio: Correcting for selection bias, backtest overfitting, and non-normality. *The Journal of Portfolio Management*, 40(5):94–107, 2014. doi: 10.3905/jpm.2014.40.5.094.

Stephen J. Brown and Jerold B. Warner. Using daily stock returns: The case of event studies. *Journal of Financial Economics*, 14(1):3–31, 1985. doi: 10.1016/0304-405x(85)90042-x.

Lars Peter Hansen and Robert J. Hodrick. Forward exchange rates as optimal predictors of future spot rates: An econometric analysis. *Journal of Political Economy*, 88(5):829–853, 1980. doi: 10.1086/260910.

Peter Reinhard Hansen. A test for superior predictive ability. *Journal of Business & Economic Statistics*, 23(4):365–380, 2005. doi: 10.1198/073500105000000063.

Campbell R. Harvey, Yan Liu, and Heqing Zhu. … and the cross-section of expected returns. *The Review of Financial Studies*, 29(1):5–68, 2016. doi: 10.1093/rfs/hhv059.

Sture Holm. A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, 6(2):65–70, 1979. URL `https://www.jstor.org/stable/4615733`.

Hans R. Künsch. The jackknife and the bootstrap for general stationary observations. *The Annals of Statistics*, 17(3):1217–1241, 1989. doi: 10.1214/aos/1176347265.

Whitney K. Newey and Kenneth D. West. A simple, positive semi-definite, heteroskedasticity and autocorrelation consistent covariance matrix. *Econometrica*, 55(3):703–708, 1987. doi: 10.2307/1913610. Final page awaits primary-source verification; retained from the frozen manuscript.

Dimitris N. Politis and Joseph P. Romano. The stationary bootstrap. *Journal of the American Statistical Association*, 89(428):1303–1313, 1994. doi: 10.1080/01621459.1994.10476870.

Joseph P. Romano and Michael Wolf. Exact and approximate stepdown methods for multiple hypothesis testing. *Journal of the American Statistical Association*, 100(469):94–108, 2005. doi: 10.1198/016214504000000539.

Halbert White. A reality check for data snooping. *Econometrica*, 68(5):1097–1126, 2000. doi: 10.1111/1468-0262.00152.

A Preregistration and Timestamping

The prospective protocol was locked on October 30, 2026 (UTC lock receipt: `2026-10-30T07:35:40.506898+00:00`). Its SHA-256 is `29ba4f43ee2f0f473c0d7bff5b51988e20762fc05d72126a05ef010e543ec13d`. The registration manifest binds 120 files and has SHA-256 `9a2ca0a7bbd2e691268ad0a50affa18b3bf13722ca7cf1a8e8b284cc04d53e01`.

The initial OpenTimestamps protocol receipt has SHA-256 `8d8fb0db6af833cf7917064218b342ac2ce6eec3240cf501ab95caaef17d3ab6`; the manifest receipt has SHA-256 `fbf16d44fc65472f1132f96ff71ee0ca816e9e4ff6f830511e62e30d4f81a26f`. Each contains four calendar attestations. The receipt verification at `2026-10-30T07:43:11Z` recorded pending Bitcoin confirmation; calendar receipts alone do not establish Bitcoin confirmation.

The public registration is available at `https://axiocairn.com/research/prereg/paper1/`. The original retrospective timestamp and its limitations are described in Section 6.1.

B Data and Code Availability

The analysis is implemented in `code/run_prereg.py`, SHA-256 `87848c7f667134cf9d17603ccc1a8d2031b5dcec990a44c2b84eeb191fcc0cb8`, using Python 3.11.14, NumPy 2.3.5 and pandas 2.3.3. The evaluated engine is identified by SHA-256 `8a22ea2bfd6c6289cd8ceea67d4037b979fbe16f9ce8bfcf478844e4015aeadb`. Code and output identifiers are recorded here to support verification against the archived analysis.

The main output, `results/results.json`, has SHA-256 `6d5b94bfe318ec77097173e5d8ae3537db406566f4c76f4c824eba370e74a32f`; the per-stock output, `results/per_stock.csv`, has SHA-256 `e948f2392e290239c5691df7b1a0a49e0ceb756526fd49e701ffc79656974323`. The additional checks described in deviations 11–13 are implemented in `code/review_checks.py`; their output has SHA-256 `e8653d23b682959572d71657f3189bdeae103358f92ed5ad709247c49bd28169`.

The underlying data are commercial market data (not redistributed). Licensing restrictions preclude redistribution of the source data; only derived statistics are reported. No individual price series or downloadable datasets are provided. Reproduction of the numerical analysis requires access to the same original data snapshot, which is not distributed with this article.

Table replication identifiers

Table 1. Source: `out/results.json` (October 22, 2026, SHA-256 `6d5b94bfe318ec77097173e5d8ae3537db406566f4c76f4c824eba370e74a32f`); `code/render_tables.py` SHA-256 `48b48835140120bc341a964adbbab3aa2c134b4b7c96bfd5a53822924049f78e`; reproduced table SHA-256 `76fa716da6df0b19c72841903939a0d1497804fb9bdaca599da250adbbf4ba97`.

Table 2. Source: `out/results.json` (October 22, 2026, SHA-256 `6d5b94bfe318ec77097173e5d8ae3537db406566f4c76f4c824eba370e74a32f`); `code/render_tables.py` SHA-256 `48b48835140120bc341a964adbbab3aa2c134b4b7c96bfd5a53822924049f78e`; reproduced table SHA-256 `654adbb6f2977bb3537bad1001965f1e93eb6ea28b17d469151aeac543073f2e`.

Table 3. Source: `out/table3_input_provenance.json` (prepared October 29, 2026); `out/results.json` (October 22, 2026, SHA-256 `6d5b94bfe318ec77097173e5d8ae3537db406566f4c76f4c824eba370e74a32f`) and `out/review_checks.json` (October 23, 2026, SHA-256 `e8653d23b682959572d71657f3189bdeae103358f92ed5ad709247c49bd28169`); `code/render_tables.py` SHA-256 `48b48835140120bc341a964adbbab3aa2c134b4b7c96bfd5a53822924049f78e`; reproduced table SHA-256 `88464b5ce8a3d48d1c145aab9acfc7aca7e0ff5357dfd6200632f1772c717a3c`.

Source for the separate-seed rederivation: `out/rederive_separate_seed_comparison.json`, SHA-256 `679d6c9bcc4ba36e7caaa3a0524d6eed9e9e451a4285ec3f900803b9a52d1f7f`; processing `code/rederive_independent.py`, SHA-256 `3706bcb12b6c1e26e22a0e1d908b65776278195af7c57d8d54c1730f89f2f040`.

C Declarations

Competing Interests

The author is affiliated with Axiocairn, which develops the evaluated system.

This paper is for research purposes only and is not investment advice.