Methodology and Reproducibility for Participant-Level Order Flow Forecasting

A Pre-Registered Pipeline

Joerg Osterrieder

2026-05-09

Pre-registration tag: `pre-reg-v1` at commit `0f7741f`
Manifest SHA: `7be9eca…b6fe6d`
Repository: Digital-AI-Finance/daily-order-flow-based-excess-alpha

Outline

The replication crisis in empirical finance

Harvey (2017): most claimed cross-sectional return predictors do not survive out-of-sample testing.

Welch (2019): the workflow itself is the failure point. Researchers select specifications, splits, and hypothesis families AFTER seeing the data, then report the survivors.

Common root: enough degrees of freedom remain after observing data to amplify positive results without declaring the selection.

A pre-registered methodology stack removes those degrees of freedom by locking every substantive choice in the tag history before any reference run touches the outcome data.

What this paper delivers

A pre-registered, reproducibility-first methodology stack for daily participant-level order flow forecasting, validated on synthetic data.

The contribution is the methodology itself, ten primitives:

- closed-family hypothesis design [ADR-008]
- walk-forward CV with explicit purge and embargo [ADR-003, ADR-018]
- pre-committed estimator universe [ADR-004]
- deterministic seed graph
- fixed-cost portfolio harness [ADR-017]
- paired hypothesis tests with HAC and small-sample correction
- Benjamini-Hochberg FDR over a closed family
- serial-correlation-aware block bootstrap with locked parameters
- three-tier claim-to-evidence taxonomy
- reproducibility capsule for archival re-derivation

Synthetic-first development [ADR-014, ADR-015]

No code path may read real-data files before the `pre-reg-v1` tag is cut.

Four-layer source firewall [ADR-014]:

1. `data/real/` is gitignored. Real CSVs never enter the public commit history pre-tag.
2. Pipeline `Source` enum hard-codes `synthetic` or `real`; the `real` path requires a config flag plus the pre-reg tag.
3. `io.firewall` validates source-of-data at load time.
4. CI matrix never installs real data; only synthetic fixtures.

The constraint was binding throughout development. The empirical paper is a companion submission, gated on the Phase 0.5 history rewrite that scrubs licensed CSVs from the public commit history.

Synthetic generator

Per-column independent `numpy.random.Generator` streams via `SeedSequence.spawn()` [Phase 1f US-072].

Stochastic call sites derive seeds via `_seeds.derive(master, label)` with a registered `label` in `SEED_LABELS`. Adding a new label requires a code edit and is gated by a unit test.

Reference runs set `REFERENCE_RUN=1` which forces `n_jobs=1` on every sklearn estimator, removing thread-scheduling nondeterminism.

The Phase 1e single-stream generator artefact (Type-I rate at IR=0 above nominal 5%) was traced to adjacent rows sharing RNG state. Per-column independent generators restored proper Type-I calibration.

Walk-forward purge plus embargo [ADR-003, ADR-018]

Splitter: `ExpandingWindowPurgedEmbargo`.
Embargo enforced in trading-day index space:

$$\texttt{test\_max(k)} + \texttt{embargo} + \texttt{purge} \leq \texttt{test\_start(k+1)}$$

Horizon-parameterised purge per ADR-018: purge equals horizon plus two.

Inner CV for `TunedRandomForestWrapper`: `_InnerPurgedCV` prevents PR-AUC inflation from leakage into the hyperparameter search.

Why it matters: in time-series data with autocorrelated targets, naive k-fold CV leaks future information into the training set; PR-AUC and ROC-AUC inflate, and the eventual out-of-sample performance disappoints.

Pre-committed estimator universe [ADR-004]

Five estimators, locked at `pre-reg-v1`:

- `MajorityBaseline`: predicts the empirical majority class
- `PersistenceBaseline`: predicts $y_t = y_{t-1}$
- `MomentumBaseline` for $k \in \{5, 20, 60\}$: sign of the k-day return
- `LogisticBaseline`: logistic regression on the registered features
- `TunedRandomForestWrapper`: `RandomizedSearchCV` with `n_iter=50`

No additions without an ADR amendment.

This is the constraint that closes the model-zoo selection-bias gap. A reviewer cannot ask "did you cherry-pick the best baseline?" because the baseline set was committed to the tag history before any reference run.

36-name hypothesis family [ADR-008]

Three horizons $\times$ three asset classes $\times$ four baseline families = thirty-six hypotheses, fixed at `pre-reg-v1`.

Tests:

- DeLong (`delong.py`): paired ROC-AUC test
- Diebold-Mariano (`diebold_mariano.py`): predictive-accuracy differential, HAC standard errors with the Harvey-Leybourne-Newbold small-sample correction

Multiple-testing correction: Benjamini-Hochberg FDR at $q = 0.05$ across the full thirty-six-name family (`bh_fdr.py`).

The 36-name family was published as the hypothesis pre-registration artefact at the `pre-reg-v1` tag commit. The empirical paper reports per-hypothesis decisions against this family.

Cost model [ADR-017]

Primary: fixed five-basis-point half-spread on every traded notional.

Sensitivity: Corwin-Schultz high-low spread estimator (Corwin and Schultz, 2012) as a daily-frequency upper bound.

The cost model is pre-committed for the same reason as the hypothesis family: post-hoc selection of a cost assumption that shifts a marginal result into significance is an open degree of freedom that the literature has historically exploited. Pre-committing closes it.

The repo-side enforcement: a pre-commit hook blocks any reference to `ASK`, `BID`, or `SPREAD` columns in `.py` files. The cost model uses fixed bps, not live spread quotes.

Three-tier claim-to-evidence taxonomy

Every numerical claim in the paper traces to one of three tiers:

Tier	Definition
Tier A	derived from the deterministic block (manifest hash); in-line OK.
Tier A-derived	bootstrap CI computed against the same manifest; in-line OK.
Tier B	Docker-only figure rendering; **prohibited** from in-line text; figures only.

Two enforcement scripts:

- `verify_caption_sources.py`: scans `paper.qmd` captions; numeric tokens must appear in `docs/claim-to-evidence.md` Tier α rows.
- `verify_claims.py`: scans the claim-to-evidence crosswalk for tier consistency.

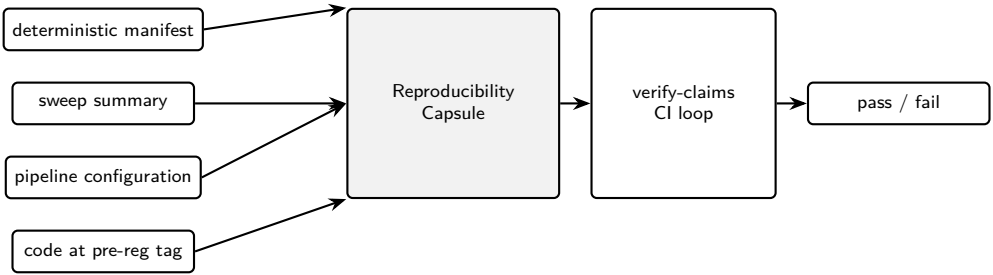
Reproducibility capsule, two tiers [ADR-016]

Tier	Artefact	Reproduces in
Tier A	`manifest.sha256` over the deterministic block	all CI matrix jobs, bit-identical
Tier B	per-figure SHA	Docker job only (font and float rendering sensitive)

Current Tier A manifest hash:

`7be9ecaadf707a0bd948b5b83574c89d729ceca4893f53a84f9d6b413ca6fe6d`

Reproducibility-capsule schematic



Same schematic as Figure 3 of `paper.qmd` (Tier A inline TikZ).

Synthetic-with-imposed-signal sweep (Section 4)

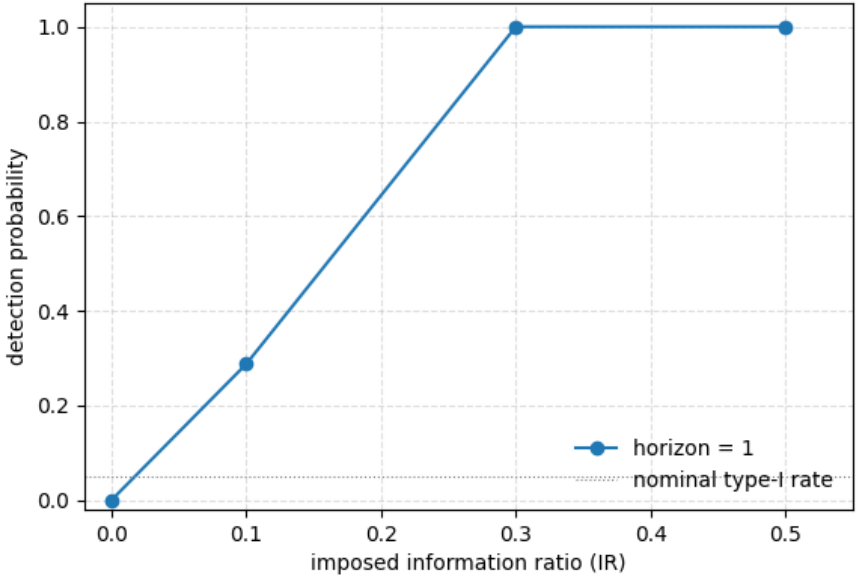
Imposed information ratio levels: $\{0.0, 0.1, 0.3, 0.5, 1.0\}$
Forecast horizon in the canonical sweep `sweep-v2.json`: $h = 1$.

For each IR level:

- Generate the synthetic panel via per-column independent generators.
- Impose a known information ratio.
- Run the full pipeline (CV, baselines, RF, DM test, BH-FDR).
- Record detection probability across seeds.

Type-I check: at $\mathrm{IR} = 0$, detection probability does not exceed nominal 5% under per-column independent generators (the Phase 1e single-stream artefact was the regression test that motivated Phase 1f).

Detection probability curve



SPY 2023 OHLCV demonstration (Section 4.5)

To show the pipeline runs on publicly available real data, a simplified four-estimator version (Majority, Persistence, Logistic, RandomForest) was applied to SPY daily OHLCV for calendar 2023.

Scope: 259 trading days; sign(log return) at t forecasts sign(log return) at $t+1$; five OOS folds; sixty-day minimum training window.

Honest caveat: sign(log return) is a crude proxy for participant- level signed order flow. The equality of PR-AUC across the four estimators is expected (each has access to one binary feature). The result confirms mechanics, not edge.

Real edge requires the participant-decomposed Xetra panel, which is the subject of the companion empirical paper.

What `pre-reg-v1` delivers to a reviewer

The 36-name hypothesis family, the cross-validation splitter, the estimator universe, the cost model, the test statistics, the FDR procedure, and the claim-to-evidence taxonomy were fixed before outcome data were seen.

A reviewer asking whether the authors could have chosen a different AUC threshold after seeing results inspects the tag history and confirms they could not.

A reviewer asking whether the result is robust to the embargo length re- derives the answer from the deterministic seed graph and the locked `pipeline.yaml`.

The review conversation collapses from "prove you did not p-hack" to "are the pre-committed choices appropriate for the research question?"

Replication recipe

The Zenodo capsule (DOI to be issued at submission) bundles the full codebase at `pre-reg-v1`, the synthetic reference run, and the manifest verification script.

A replicator reproduces every Tier A number in the paper:

- the power-analysis sweep
- the cross-validation metrics on synthetic data
- the bit-identical manifest hash

One verification command:

```
python scripts/check_signoffs.py && uv run pytest -q
```

Exit zero means the methodology is intact.

Y-then-X two-paper arc [ADR-021]

This paper (Y): methodology plus reproducibility, validated on synthetic data. Null-agnostic; ships independent of empirical outcome.

Companion paper (X): empirical application of this stack to the complete Xetra participant-category daily order-flow panel. Requires the Phase 0.5 history rewrite plus the data-availability questions to be resolved.

No methodology choice in the empirical paper differs from the choices locked at `pre-reg-v1`.

The two papers are designed to be read as a pair: this one as the pre-committed primitive layer, the empirical one as the outcome layer.

Methodology and Reproducibility for Participant-Level Order Flow Forecasting

Pre-registration tag:	`pre-reg-v1` at `0f7741f`
Manifest hash:	`7be9eca…b6fe6d`
Repository:	Digital-AI-Finance
SSRN:	v2 staged for upload
Companion paper:	forthcoming