Bias Fingerprints: Diagnosing Data-Handling Errors from the Cross-Section of Reported Factor Performance

Raunak Sood
University of Southern California
raunakso@usc.edu

September 2026

Abstract

A table of factor t-statistics is usually all an outside reader gets. The allocator reading a manager's factor deck, the referee reading a submitted backtest, the desk pricing a vendor signal: none of them can run the pipeline that produced the numbers, and the literature on look-ahead and survivorship bias is written for someone who can.

We propose *bias fingerprinting*, a framework that turns a controlled research pipeline into a diagnostic instrument for uncontrolled ones. Each data-handling defect is represented as the vector of shifts it induces across a fixed factor cross-section; that vector is measured once on a pipeline the analyst owns, then used as a known quantity in a hypothesis test on a reported table. The framework has four parts: a generator that re-runs one study under alternative data conventions, a measured signature per defect, an inference model that separates signature from unknown true performance, and a pre-specified validation protocol.

We instantiate it on a twenty-two factor US equity study and calibrate two signatures. Joining fundamentals on fiscal period end rather than filing date inflates mean information coefficient by 59% among the eleven factors that read a filed figure and manufactures four t-statistics above 2.0, while leaving the eleven price-only factors unchanged to machine precision. Conditioning the universe on current index membership does not inflate at all: mean IC *falls* by 0.0043 while `amihud_illiquidity` flips from $t = -1.11$ to $+2.80$ and the low-volatility anomaly inverts. The two signatures correlate at -0.088 and disagree in sign on value factors, so they are separable directions in factor space.

We then run the protocol. Both signatures are stable across disjoint subsamples and survive a block bootstrap, but separability holds for only one of them. Against a τ resembling a real factor table, universe conditioning is recovered at 99% accuracy while period-end dating is recovered at 64% and labels 70% of clean tables defective. The reason is geometric and measurable in advance: after removing family means the dating signature is smaller than the within-family variation of true performance and correlates $+0.543$ with it, so a clean table already supplies most of the defect's shape. This yields a condition a signature must satisfy before it can be used diagnostically, checkable without collecting a single reported table.

1 Introduction

Factor research is consumed as a table. An allocator sees mean IC, an information ratio and a t-statistic for each of a manager's signals. A referee sees the same for a submitted paper. A desk evaluating a vendor sees less. In none of these cases is the pipeline available, and a Sharpe ratio carries no information about which construction choice produced it.

The literature on data-handling bias is prescriptive and self-directed. Banz and Breen [5] established that look-ahead inflates backtested returns; survivorship is textbook. Both are taught as things to avoid in one's own code: date the fundamentals correctly, build the universe point-in-time, shift a feature forward and check that the result degrades. All of it assumes the reader controls the pipeline. Nothing in it helps the reader who controls only the table.

This paper proposes a framework for that reader. The idea is that a data-handling defect is not a scalar amount of inflation but a *direction* in factor space, and that the direction is a property of the defect rather than of the study it damages. A defect that corrupts filed fundamentals cannot move a factor built from prices alone. A defect that corrupts the universe moves liquidity and volatility factors hardest, because it selects on size. Different defects therefore leave different patterns, and a pattern is something a reader can see in a table.

The framework. Bias fingerprinting has four components, set out in Section 3. A *generator* re-runs one factor study under alternative data conventions with everything else held fixed, so the difference between two arms is attributable to the convention. A *signature* δ_b is the vector of t-statistic shifts a defect b induces across the factor cross-section, measured on that generator. An *inference model* treats an observed table as true performance plus a signature plus noise, and recovers the signature using exclusion restrictions and within-family contrasts rather than a fitted classifier. A *validation protocol* states, in advance, the four measurements that have to succeed before the diagnostic may be applied to a real table, and the order in which they are made.

The framework is not specific to the two defects calibrated here, or to equities. It applies wherever a research pipeline can be re-run under a defective convention, which is to say wherever the analyst owns a correct pipeline. That is what makes the labelled data for validation free: biased panels are generated rather than collected, so the diagnostic can be tested at scale with no annotation and no external dataset.

Calibration. We instantiate the framework on a twenty-two factor cross-sectional US equity study and calibrate two signatures. Joining fundamentals on fiscal period end inflates mean IC by 59% of its true value among the eleven factors that read a filed figure, and turns four insignificant factors into t-statistics above 2.0. Its eleven price-only factors do not move at all, with maximum absolute drift 0.0: exact equality in floating point, which is the strongest form an exclusion restriction can take. Conditioning the universe on current index membership behaves differently in every respect. Mean IC falls by 0.00432 and only 8 of 22 factors improve, but `amihud_illiquidity` moves from $t = -1.11$ to $+2.80$, `turnover_1m` becomes significant with the wrong sign, and the low-volatility anomaly inverts. The channel is not the survival filter usually named: the biased panel wrongly *includes* 203 names a day that had not yet been admitted to the index, against 88 it wrongly excludes, and the included cohort is a third the size and a third the liquidity of a genuine constituent.

Across the twenty-two factors the two signature vectors correlate at -0.088, and they disagree in sign on value: dating inflates `earnings_yield` and `roe` by $+0.98$ and $+1.07$ t-units while universe conditioning deflates them by -1.28 and -1.24. A table cannot be explained by both. Together with the exclusion restriction, that separation is the empirical content the diagnostic rests on.

Contribution. Three things are new. The framework itself, which reverses the direction of the standard treatment of data-handling bias, from prescription for the producer to diagnosis for the consumer. The two calibrated signatures, measured with identical prices, universe, tradability screen and forward returns, changing one convention at a time. And the characterisation of the universe channel as a forward-looking growth filter rather than a survival filter, which explains why it relocates measured performance instead of inflating it.

The framework is specified and calibrated, and gates one through three of Section 3.4 are run in Section 8. Gate four, calibrated application to reported tables, is not, and the ordering is part of the method: a diagnostic that reports a label without a calibrated reference distribution is worse than no diagnostic. Section 3 sets out the framework, Section 4 the generator it is instantiated on, Sections 5 and 6 the two signatures, Section 7 their separability, and Section 3.4 the validation protocol and the conditions under which the diagnostic transfers.

2 Related work

Three literatures bear on this, and none of them answers the question above.

The errors themselves are old and well characterised. Banz and Breen [5] established that look-ahead inflates measured returns in accounting-based tests. Survivorship bias in performance evaluation was quantified by Brown, Goetzmann, Ibbotson and Ross [6] and by Elton, Gruber and Blake [9], and Shumway [20] and Shumway and Warther [21] showed that even CRSP needs care, because missing delisting returns bias the small-cap end of the cross-section. Kothari, Shanken and Sloan [13] showed that Compustat backfill and survivorship alone can manufacture a book-to-market premium, which is the same mechanism this paper's universe arm isolates. None of that work is superseded here. What it does not do is characterise the *shape* of each distortion across a factor cross-section, which is the object this paper measures.

Work on false discovery in the factor zoo concerns multiplicity, not provenance. Lo and MacKinlay [16] formalised data-snooping bias. Harvey, Liu and Zhu [10] argue that the accumulated search over published factors demands a t-statistic threshold near 3.0. Bailey and López de Prado [3] deflate the Sharpe ratio for selection and non-normality, and Bailey et al. [4] show how fast backtest overfitting manufactures apparent skill. McLean and Pontiff [18] document post-publication decay consistent with a substantial share of published anomalies being artifacts. Hou, Xue and Zhang [14] replicate 452 anomalies under a common construction and find most do not survive it, and Harvey and Liu [15] give a bootstrap procedure for judging a factor against the luck of the search that found it. Both treat construction as something to standardise; this paper treats it as something to infer.

That literature asks whether a result survives the number of hypotheses tested. It is silent on whether a result is an artifact of how the panel was *built*, and the two leave different traces.

Multiplicity inflates broadly and without regard to how a factor is constructed; Section 6 exhibits a construction defect that *lowers* mean IC while manufacturing significance in one specific corner of the cross-section. No multiple-testing correction would flag it, because nothing about the average looks improved.

The machine-learning literature names the mechanism but assumes access. Kaufman, Rosset and Perlich [12] give the canonical formulation of leakage and propose detection through legitimacy checks on the training data. Kapoor and Narayanan [11] document leakage across a wide range of published ML-based science and argue it constitutes a reproducibility crisis rather than a set of isolated mistakes. Both address a researcher holding the pipeline. The device used in Section 5 to establish the exclusion restriction is metamorphic testing [8]: identify a transformation of the input under which the output must be invariant, apply it, and treat any change as a defect. The idea is standard in software testing and, as far as we have found, has not been applied to the construction of an asset-pricing panel.

Reproducibility infrastructure. Chen and Zimmermann [7] release open-source replications of a large set of published cross-sectional predictors, and Novy-Marx and Velikov [19] price anomalies net of realistic trading costs. Both establish that construction choices dominate reported performance. Neither treats the construction choice as a quantity to be *inferred* from a reported table.

Where this sits. The forward direction, what a given defect does to a given study, is a measurement in the tradition of the first group, made with the controlled one-arm-at-a-time design the fourth group argues for. The inverse direction is what is new, and Section 3.4 is explicit that it is proposed rather than established.

3 The framework

3.1 The generator

Fix a factor registry, an evaluation window, a forward-return definition and a walk-forward split. A generator is a pipeline that produces the resulting table of factor t-statistics under any of a set of data conventions, sharing every component that the convention does not touch. Prices, tradability screen, sector neutralisation, rank-normalisation and evaluation code are common across arms; the arm changes one construction decision.

The generator is what supplies the counterfactual. Without it, the effect of a defect can only be argued; with it, the effect is a difference between two runs of the same code. The requirement it imposes is real: the analyst must own a pipeline that is correct on the dimension being studied, because the correct arm is the baseline everything is measured against.

3.2 The signature

Let t^{pit} be the vector of factor t-statistics from the correct arm and t^b the vector from the arm containing defect b. The signature is the difference,

$$\delta_b = t^b - t^{\text{pit}}, \tag{1}$$

one component per factor, with $\delta_{\text{none}} = 0$ by construction. A signature is useful to the extent it has structure: components that are exactly zero, components that are large, and a direction that other defects do not share. Sections 5 and 6 measure δ_b for two defects and Section 7 reports the geometry of the pair.

3.3 The inference problem

An observed table from an unknown pipeline is a vector t of factor t-statistics. Model it as

$$t = \tau + \delta_b + \varepsilon, \tag{2}$$

where τ is the study's true factor performance, δ_b is the measured signature of whichever defect the study contains, and ε is sampling noise. The diagnostic problem is to infer b.

Since τ is unknown and twenty-two dimensional, b is not identified without further structure. The framework uses three sources of it, in increasing order of assumption:

1. **Exclusion.** Some defects cannot move some factors, for reasons of construction rather than magnitude. Dating bias cannot move a factor that never reads a filed figure. If a study's strongest results sit in the excluded block, that defect is ruled out whatever τ is. This requires no assumption on τ at all.

2. **Within-family contrast.** Assume τ is exchangeable within factor families: value factors share a true level, momentum factors share a different one. Family means then absorb τ, and within-family deviations are attributable to δ_b. This is the workhorse assumption, and it is testable against the factor-zoo literature.

3. **Projection.** Score a table by its projection onto each candidate signature after removing family means. Under the null that τ is exchangeable within families, that projection has a computable reference distribution, so the output is a p-value rather than a label.

Fitting a classifier from tables to defect labels is the obvious alternative and it does not survive the sample size. A twelve-year out-of-sample window gives six non-overlapping two-year observations; three conditions give eighteen labelled examples for a twenty-two dimensional feature vector, and cross-validation over overlapping windows would conceal the overfitting rather than reveal it. The signature is measured once from a pipeline we control and then held fixed. There is nothing to learn, only something to estimate and something to test.

3.4 The validation protocol

The framework commits in advance to four measurements, in order. Each is a gate: failing one ends the method rather than downgrading its claims.

1. **Signature stability.** δ_b must be a property of the defect rather than of the sample it was measured on. Re-measure on disjoint subperiods and random half-universes. If `amihud_illiquidity` does not move sharply positive under universe conditioning in every subsample, there is no signature.

2. **Signature uncertainty.** A point estimate cannot support a test. Bootstrap over dates for a sampling covariance, and report each signature with error bars.

3. **Separability under realistic noise.** Simulate tables from priors over τ, apply each defect, add noise at the level a twelve-year study actually carries, and report the confusion matrix. Whether two defects remain separable at that noise level is an empirical question with a measurable answer.

4. **Calibrated application.** Only then run the diagnostic on reported tables, and report a p-value or a posterior over {none, dating, universe, both}, never a bare label.

Gate three is where labelled data would normally be the binding constraint, and here it is free. Biased panels are produced by the generator itself, so the confusion matrix can be estimated over as many synthetic studies as compute allows, with no annotation and no external dataset. That property is what makes the protocol executable by one researcher, and it is a reason to prefer this formulation over a supervised one independent of sample size.

4 The generator, instantiated

Universe. S&P 500 point-in-time membership, reconstructed as dated constituent spells from the revision history of a public reference page. This is a free approximation to a point-in-time constituent file and is not CRSP; Section 9 treats the consequences.

Prices. Daily adjusted closes, 2010–2026, with a tradability screen on price, volume and index membership. Prices come from a free public source that does not carry delisted tickers, so 263 of the 990 names that held a membership spell from 2004 on have no price history in the current cache and are dropped. The point-in-time baseline is therefore itself survivorship-filtered: the universe arm measures the effect of conditioning on current membership *relative to* a universe that already lacks most of the names that left by failing, and a description of the universe as "including companies that went bust" holds only for names that left by acquisition or demotion. Delisting returns are absent for the same reason. Section 9 records the consequence.

Fundamentals. SEC XBRL company facts, keyed on the date each figure was filed. Restatements resolve to the earliest filing. Cumulative cash-flow spans are differenced to true quarters. Rows whose filing date is an XBRL-adoption artifact, meaning a period ending long before the filing that first reports it, are flagged rather than silently trusted.

Factors. Twenty-two published cross-sectional factors, each signed to its published direction, sector-neutralised and rank-normalised. Eleven read a filed fundamental. Eleven use only price and volume.

Evaluation. Information coefficient against 21-day forward returns, out-of-sample from 2014-01-01, with a purged and embargoed walk-forward following López de Prado [17]. The evaluation code is shared across every arm below, so the standard cannot drift between them.

5 Signature 1: fundamentals dated on fiscal period end

A quarter ending 31 March is not public on 31 March. It becomes public when the 10-Q is filed, a median of 34 days later (from scripts/filing_lag_study.py; the run output is not committed).

Joining on period end hands the backtest a month of information nobody had.

The generator produces three arms, holding everything else fixed:

- **PIT**: visible the day the filing landed, and correct.
- **LAG45**: every company files 45 days after quarter end, which is common practice.
- **NAIVE**: visible the instant the quarter closed, which is the defect.

5.1 The controls hold exactly

Eleven factors read no filed figure. Across all three arms their IC is identical, with maximum absolute drift 0.0. That is exact equality in floating point, not a small number.

This is what licenses the rest of the section. It establishes that nothing in the generator responds to the filing calendar except the factors that must, and it is the empirical basis for the exclusion restriction of Section 3.

One factor required care. `turnover_1m` is built from price and volume but divides by shares outstanding, so it inherits the filing calendar through its denominator. It is the only price-family factor that does. We determine dating sensitivity by measurement and check it against a declared set, so a factor that acquires a filed input is surfaced rather than hidden among the controls.

5.2 What the defect is worth

Convention	Mean IC inflation	Look better	Cross $t = 2$ spuriously
Fixed 45-day lag	+9% of true IC	8 of 11	**0 of 11**
Period-end join	**+59% of true IC**	10 of 11	**4 of 11**

Table 1: Effect of each dating convention, summarised over the eleven factors that read a filed figure. Averaging in the eleven controls, which cannot move by construction, would halve the reported inflation for free.

The four manufactured significances are `earnings_yield`, `cash_flow_yield`, `roe` and `accruals`. Each is insignificant under correct dating and above $t = 2.0$ under the period-end join.

The comparison between the two rows of Table 1 is the practical result. A fixed 45-day lag is wrong, being right on average and wrong for every filing that arrives later than that, but it manufactures no false discoveries in this sample. The period-end join manufactures four. The distinction between "approximately right" and "wrong" is the distinction between a bias and a bug, and only the second leaves a signature worth testing for.

6 Signature 2: a universe conditioned on current membership

Download today's constituents, pull their price history, backtest. The resulting panel conditions on index membership as of the end of the sample.

6.1 The dominant channel is not the one usually named

Survivorship is normally described as dropping losers: the acquired, the bankrupt, the demoted. In index research that is the smaller half.

Cohort	Names/day	Median $ volume	Median market cap
Genuine members	386	$95M	$10.1B
Wrongly **included**, not yet admitted	**203**	**$28M**	**$3.5B**
Wrongly **excluded**, later deleted	88	$54M	$5.3B

Table 2: (Figures from an uncommitted run; no script in the repository reproduces this table, so it is reported as recorded and unverified.) Daily averages, 2010–2015. The biased panel wrongly includes more than twice as many names as it wrongly excludes, and the added names are roughly a third the size and a third the liquidity of a genuine constituent.

These are companies that were small early in the sample and grew enough to be admitted later. Conditioning on current membership is therefore not principally a survival filter. It is a forward-looking growth filter aimed at the small and illiquid cohort, and that mechanism predicts the shape of the signature before the signature is measured.

6.2 The distortion relocates rather than inflates

Mean IC change across all twenty-two factors is -0.00432, and only 8 of 22 improve. Two factors cross $|t| = 2$.

Factor	PIT IC	Biased IC	PIT t	Biased t
`amihud_illiquidity` [1]	-0.0105	$+\mathbf{0.0224}$	-1.11	$+\mathbf{2.80}$
`turnover_1m`	-0.0047	-0.0232	-0.48	$-\mathbf{2.53}$
`idio_vol_60d` [2]	$+0.0059$	-0.0179	0.49	-1.53
`vol_60d`	-0.0013	-0.0234	-0.08	-1.49

Table 3: The four factors most affected by conditioning on current membership. Illiquidity flips sign and becomes significant; turnover becomes significant with the wrong sign; the low-volatility anomaly inverts.

In a panel conditioned on current membership, being illiquid predicts returns, because the illiquid names in today's index are precisely those that were small years ago and then grew into it. Such a panel says: buy the illiquid, high-turnover, high-volatility names. That is not a risk premium. It is a description of a company on its way into the index, read backwards.

A defect that lowers mean IC is worth dwelling on, because it is invisible to the summary statistic an outside reader is most likely to be shown. Nothing about this study's average performance looks improved. Two of its factors are nonetheless significant for reasons that have nothing to do with returns.

7 Separability of the two signatures

The two signatures do not overlap. Figure 1 shows both shift vectors on a common axis, and Table 4 summarises them by factor family.

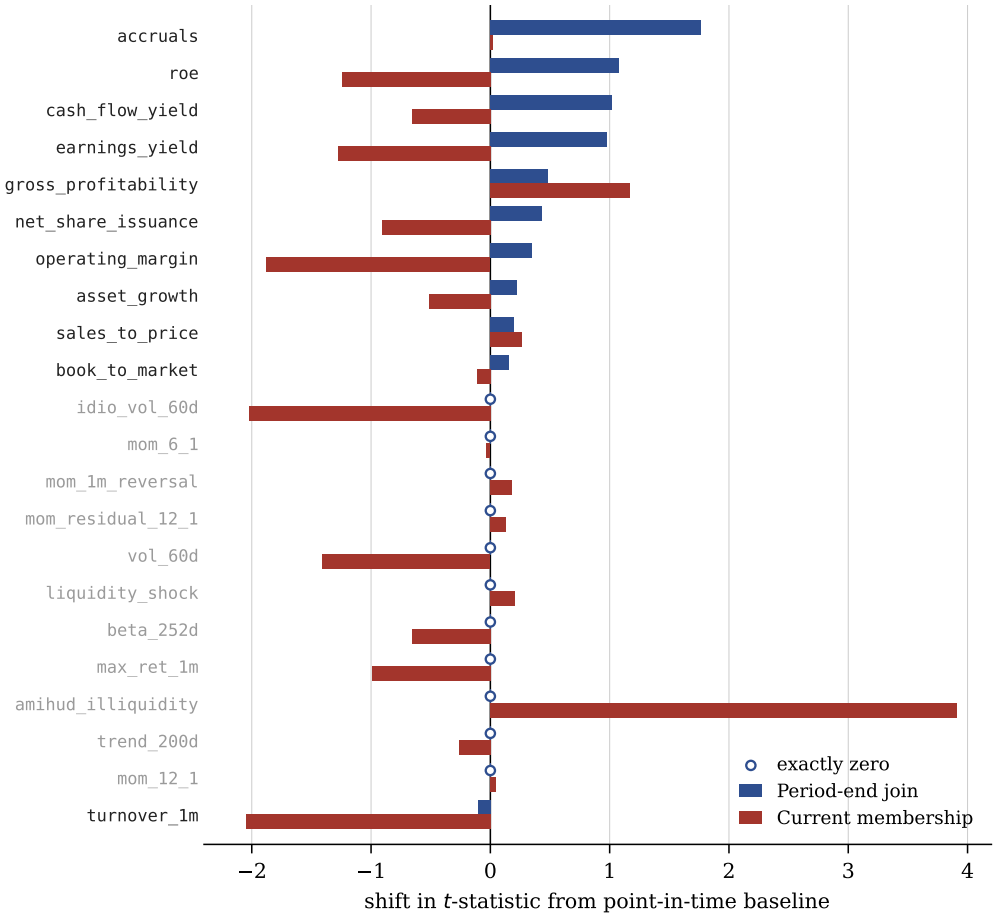


Figure 1: Shift in t-statistic from the point-in-time baseline, all twenty-two factors, ordered by dating shift. Factor names in dark type read a filed fundamental; those in grey use only price and volume. The eleven price-only factors carry an open marker rather than a bar on the period-end series: their shift is exactly zero, to machine precision. The current-membership series has no zeros anywhere, and its largest movement, `amihud_illiquidity` at +3.91, belongs to a factor that dating bias cannot touch.

Quantitatively, the two vectors correlate at −0.088 across the twenty-two factors. They are not two versions of "results look better"; they are close to independent directions in factor space. They also disagree in sign on value: dating inflates `earnings_yield` and `roe` by +0.98 and +1.07 t-units, universe conditioning deflates them by −1.28 and −1.24.

One property is a hard constraint rather than a tendency. Under dating bias the eleven price-only factors are unchanged to machine precision, because they never read a filed figure. An anomalously strong illiquidity result therefore cannot be produced by dating bias, whatever else is true of the study. This is the exclusion restriction of Section 3 in its strongest available form, and it does most of the discriminating work without any assumption on τ.

Read forward, the two signatures give the diagnostic its content. A study with a suspiciously strong illiquidity premium and a dead low-volatility anomaly has a universe problem. A study whose value and quality factors clear $t = 2$ while its price factors look ordinary has a dating problem. Section 3.4 states what has to be measured before either reading may be given a

Factor family	Period-end join	Current membership
Price and volume	**Untouched, exactly**	Heavily distorted
Value and quality	Inflated; 4 false positives	Mixed, mostly weakened
Liquidity and volatility	Untouched	Sign flips, false positives
Direction of effect	Uniform inflation	Relocation, not inflation

Table 4: The two signatures by factor family.

confidence.

8 Running the protocol

Section 3.4 commits to four gates. Gates one to three are run here. Each is computed from a cached per-date IC series for the three panels, which makes subperiod splits slices of that series and the bootstrap arithmetic rather than a sequence of panel rebuilds. Every t-statistic in this section is in the units of the factor tables above: the IC series is daily and the forward return is 21 days, so the 3,142 dates are deflated to about 150 independent observations before any t is formed. An earlier run of these gates used an undeflated t and reported shifts about $\sqrt{21} \approx 4.6$ times larger than Figure 1; the verdicts below are unchanged by the correction, the magnitudes are not.

8.1 Gate 1: the signatures are stable

Both signatures were re-measured on four disjoint subsamples: the first and second halves of the window, and odd- and even-numbered months. Every subsample signature correlates with the full-sample vector between +0.928 and +0.995. The stated criterion was that `amihud_illiquidity` must move sharply positive under universe conditioning in every subsample; it moves +3.55, +2.05, +2.98 and +2.59 t-units respectively.

The exclusion restriction was re-checked per subsample rather than only in aggregate. The maximum absolute shift on the eleven price-only factors under the dating defect is 0.0 in each of the four, so exact invariance is a property of the construction and not of the full-sample average. Gate 1 passes on the time splits. The protocol also specified random half-universe splits, which were not run; the stability claim rests on time splits alone.

8.2 Gate 2: the signatures carry error bars

Two thousand moving-block bootstrap draws over dates, with blocks of one 21-day horizon so that the serial dependence of overlapping returns is kept, give a sampling covariance for each signature. The largest shifts are estimated precisely: under the dating defect `accruals` shifts $+1.77 \pm 0.31$ and `roe` $+1.07 \pm 0.15$; under universe conditioning `amihud_illiquidity` shifts $+3.91 \pm 0.44$ and `turnover_1m` -2.05 ± 0.25. The point estimates agree with Figure 1 by construction. The -0.088 correlation between the two signatures has a bootstrap 95% interval of $[-0.225, +0.036]$: the interval covers zero and excludes ± 1 by a wide margin, which is the property the diagnostic needs. Gate 2 passes.

8.3 Gate 3: one signature is identified and one is not

As specified in Section 3.4, Gate 3 was a joint gate whose failure ends the method. It failed: dating and survivorship are not jointly separable. Scoring each defect separately, as below, is an amendment made after that result, and the per-defect verdicts should be read as post hoc.

Gate three simulates studies, applies a defect, adds sampling noise at the level this twelve-year window actually carries, and classifies by nearest signature after removing family means. Labels are free because the generator produces them, so the confusion matrix is estimated over four thousand simulated studies per condition. The classifier does not model τ: whatever within-family structure true performance has is treated as noise around the template. A classifier that estimated τ jointly with the label could do better, so the failures below are statements about this estimator under this τ, not lower bounds on what any estimator could recover.

The result depends on one assumption, which is the point of running it. Under τ exchangeable within families up to idiosyncratic noise of 0.5 t-units – the assumption the inference model of Section 3 is built on – the universe diagnostic is recovered at 99.7% and the dating diagnostic at 76.7%, both above the criterion. Under a τ resembling the measured factor table, it splits:

Diagnostic	Accuracy	Worst class
None vs. universe conditioning	99.2%	98.5%
None vs. period-end dating	63.9%	29.5%
Joint, four conditions	64.3%	29.2%

Table 5: Gate 3 under a realistic τ, 4,000 simulated studies per row. Accuracy is over equiprobable classes; the worst-class column is the criterion, because an average hides the failure that matters: calling a clean table defective.

The universe diagnostic recovers both classes. The dating diagnostic labels 70% of clean tables as defective (clean recall 29.5%, dating recall 97.7%) and is unusable. In the four-condition problem the dating failure propagates: a universe-conditioned table is called "both" 68% of the time, because the classifier cannot tell whether a dating defect is also present, so the joint diagnostic can say "universe problem" but not "universe problem only".

The cause is measurable rather than inferred. Write $\tilde{x}$ for a vector with family means removed. Then $\|\tilde{\tau}\| = 2.19$, $\|\tilde{\delta}_{\text{dating}}\| = 1.40$ and $\|\tilde{\delta}_{\text{universe}}\| = 5.17$. The dating signature is *smaller* than the within-family variation of true performance it must be separated from, and worse, it points the same way: $\text{corr}(\tilde{\tau}, \tilde{\delta}_{\text{dating}}) = +0.543$, and the projection of $\tilde{\tau}$ onto the dating signature is $+0.85$ of it. A clean table already supplies six-sevenths of the defect's shape for free. The universe signature has the opposite geometry: correlation -0.268 with $\tilde{\tau}$ at more than twice its magnitude.

A second, structural obstacle compounds it. The exclusion restriction was supposed to do the discriminating work without assumptions on τ, but the family partition is nested inside the sensitivity partition. Value, quality and investment factors are entirely filing-sensitive; momentum, reversal and volatility are entirely price-only. Only the liquidity family spans both blocks. Within-family contrasts therefore give the excluded block almost no leverage on τ in the block where it is needed, and the strongest identification argument in the framework is unavailable to the test built on it.

8.4 What gate 3 establishes

The framework survives; one of its two instantiations does not. Read as a general condition, a signature is usable as a diagnostic when two things hold, and both are measurable before any table is diagnosed:

1. $\tilde{\delta}_b$ is large relative to $\|\tilde{\tau}\|$, and not positively aligned with it. Collinearity between a defect's shape and the true cross-section of factor performance is not a noise problem and does not improve with a longer sample.
2. The excluded block spans factor families, so that within-family contrasts transmit information about τ from factors the defect cannot touch to factors it can.

Universe conditioning satisfies both. Period-end dating fails both. That is a sharper statement than the framework had before the gates were run, and it is checkable for any future defect without collecting a single reported table.

9 Scope conditions

The framework transfers to any setting where a correct pipeline can be re-run under a defective convention. The two signatures calibrated here transfer less far, and the following conditions bound them.

- One index, one country, 2010–2026. The signatures are measured on the S&P 500 and should not be assumed to hold for small caps or non-US markets. Re-measurement is a run of the generator, not a new study.
- Membership spells are reconstructed from a public wiki revision history, not CRSP. The universe signature inherits whatever that approximation gets wrong.
- Fundamental coverage begins in 2010 because of the XBRL mandate phase-in, so the dating signature says nothing about the pre-XBRL era.
- The two arms are not additively decomposable. The universe arm rebuilds forward returns per universe, because a study with that defect never observes the returns of the names it excludes. Real studies contain several defects at once, and whether signatures compose is a question for gate three.
- The diagnostic needs a wide enough table to work with. A study reporting five factors gives a five-dimensional observation, and the exclusion restriction does not apply at all if none of the five are price-only.
- Gates one through three are run in Section 8. Gate four, calibrated application to reported tables, is not, and applies only to the universe signature: the dating signature does not survive gate three.
- The gate 3 simulation draws τ around this study's own measured factor table. That is the most realistic τ available, and it is one draw from one market and one window. A different factor cross-section could have a different alignment with the dating signature, which is an argument for measuring the alignment per application rather than assuming this one transfers.

- The gate 3 classifier is nearest-template with no model of τ. Part of what reads as "dating is undiagnosable" is a property of that estimator, and a likelihood that treats τ as a nuisance parameter is the natural next step.
- The price source drops delisted tickers, so the correct baseline is itself survivorship-filtered and the universe signature is measured against a universe that has already lost most names that failed. On CRSP with delisting returns the universe signature would plausibly be larger, not smaller, and the dating signature should be unchanged because it does not touch the universe.

10 Conclusion

Data-handling defects have shapes. A period-end join inflates fundamental factors uniformly by 59% of true IC and manufactures four false discoveries while leaving price factors provably untouched. Conditioning a universe on current membership relocates rather than inflates, flipping illiquidity to significance and inverting the low-volatility anomaly, through a channel better described as a forward-looking growth filter than as a survival filter. Neither shape is visible in a Sharpe ratio, and both are visible in the cross-section of what a study reports.

Bias fingerprinting takes that observation and builds a method on it: measure the shape once on a pipeline you own, then test for it in tables you do not. The pieces the method needs are in place: a generator that isolates one convention at a time, two signatures that are near-orthogonal and carry an exact exclusion restriction, an inference model that absorbs unknown true performance through family contrasts, and a protocol whose hardest step has free labelled data because the labels are generated rather than collected.

Running the protocol splits that claim in two. Both signatures are stable and both carry tight error bars, but only universe conditioning is separable from clean performance at the noise level of a twelve-year study. Period-end dating is not, and the reason generalises: its shape is nearly parallel to the true cross-section of factor performance, so a correct table already looks like a defective one. That is a property of the defect, not a shortage of data, and no longer window fixes it.

What the framework gains from failing that gate is a condition rather than a caveat. A signature is diagnostically usable when it is large relative to within-family variation in true performance, not positively aligned with it, and its excluded block spans factor families. All three are measurable on a pipeline the analyst already owns, before any reported table is touched. The rest of the catalogue (delisting-return assumptions, restatement handling, index-reconstitution timing) can therefore be screened for usability at the point of calibration, and the ones that fail can be reported as measurements of what a defect costs without being promoted to diagnostics.

Reproduction. All results reproduce with `make bias`. The dating arm is `scripts/pit_vs_naive.py`; the universe arm is `scripts/survivorship.py`. Gates one to three are `scripts/gate_ic_series.py` followed by `scripts/gates.py`, with the full output in `project/reports/gates_1_to_3.txt`, which is committed. Per-factor outputs are written to `reports/pit_vs_naive.csv` and `reports/survivorship.csv`. Prices are fetched per user rather than redistributed. The point-in-time fundamentals layer and the membership spells are released under CC BY 4.0.

References

[1] Y. Amihud. Illiquidity and stock returns: cross-section and time-series effects. *Journal of Financial Markets*, 5(1):31–56, 2002.

[2] A. Ang, R. J. Hodrick, Y. Xing, and X. Zhang. The cross-section of volatility and expected returns. *Journal of Finance*, 61(1):259–299, 2006.

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

[4] D. H. Bailey, J. M. Borwein, M. López de Prado, and Q. J. Zhu. Pseudo-mathematics and financial charlatanism: the effects of backtest overfitting on out-of-sample performance. *Notices of the American Mathematical Society*, 61(5):458–471, 2014.

[5] R. Banz and W. Breen. Sample-dependent results using accounting and market data: some evidence. *Journal of Finance*, 41(4):779–793, 1986.

[6] S. J. Brown, W. Goetzmann, R. G. Ibbotson, and S. A. Ross. Survivorship bias in performance studies. *Review of Financial Studies*, 5(4):553–580, 1992.

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

[8] T. Y. Chen, S. C. Cheung, and S. M. Yiu. Metamorphic testing: a new approach for generating next test cases. Technical Report HKUST-CS98-01, Hong Kong University of Science and Technology, 1998.

[9] E. J. Elton, M. J. Gruber, and C. R. Blake. Survivorship bias and mutual fund performance. *Review of Financial Studies*, 9(4):1097–1120, 1996.

[10] C. R. Harvey, Y. Liu, and H. Zhu. ... and the cross-section of expected returns. *Review of Financial Studies*, 29(1):5–68, 2016.

[11] S. Kapoor and A. Narayanan. Leakage and the reproducibility crisis in machine-learning-based science. *Patterns*, 4(9):100804, 2023.

[12] S. Kaufman, S. Rosset, and C. Perlich. Leakage in data mining: formulation, detection, and avoidance. *ACM Transactions on Knowledge Discovery from Data*, 6(4):1–21, 2012.

[13] S. P. Kothari, J. Shanken, and R. G. Sloan. Another look at the cross-section of expected stock returns. *Journal of Finance*, 50(1):185–224, 1995.

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

[15] C. R. Harvey and Y. Liu. Lucky factors. *Journal of Financial Economics*, 141(2):413–435, 2021.

[16] A. W. Lo and A. C. MacKinlay. Data-snooping biases in tests of financial asset pricing models. *Review of Financial Studies*, 3(3):431–467, 1990.

[17] M. López de Prado. *Advances in Financial Machine Learning.* Wiley, 2018.

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

[19] R. Novy-Marx and M. Velikov. A taxonomy of anomalies and their trading costs. *Review of Financial Studies*, 29(1):104–147, 2016.

[20] T. Shumway. The delisting bias in CRSP data. *Journal of Finance*, 52(1):327–340, 1997.

[21] T. Shumway and V. A. Warther. The delisting bias in CRSP's Nasdaq data and its implications for the size effect. *Journal of Finance*, 54(6):2361–2379, 1999.