A Look-Ahead Convention in Intraday Price-Structure Backtests and the Size of the Edge It Manufactures

Maksim Aleksandrovich Andreev, Joseph Howden
Quark Quantitative Research

Abstract

An order block is defined as the last opposing candle preceding a displacement move, so no candle can be classified as one until the displacement has completed. The convention implied by a chart annotated after the fact timestamps the entry at the order-block candle rather than at the displacement bar that identifies it, and win rates in this family of material are ordinarily quoted under that convention. We measure the error directly. Holding bars, instrument, geometry, transaction costs, sizing and exit rules fixed, and varying only the bar at which the signal is timestamped, the retroactive convention adds a median +0.98 R per trade and a median 37.2 percentage points of win rate across thirteen cost-tier by bar-interval cells of United States equities and two cryptocurrency pairs over 2016–2022. On index and megacap names at fifteen-minute bars the retroactive version reports a 78.4% win rate at a two-to-one reward-to-risk target, a mean of +1.158 R net of measured costs and a day-clustered t of 108.4; the causal version of the same rule, on the same bars, wins 37.8%, earns +0.022 R and has a t of 1.87. Causal win rates across all thirteen cells lie between 33.7% and 38.4%, against the 33.3% a two-to-one target returns with no edge at all. Enforcement is two-way: each detector declares the largest bar index its definition reads, and a prefix replay independently re-derives every sampled signal from truncated bar history, consulting no bookkeeping the detector controls, so a detector that misreports its own timing is caught regardless. The manufactured effect is larger than any genuine effect measured elsewhere in the same study, which bounds the interpretation of published results in this family: a reported win rate near 80% at a two-to-one target is consistent with this error and with no edge.

1. Introduction

Technical trading rules have a long record of failing out-of-sample tests once data-snooping is priced in (Lo & MacKinlay, 1990; Sullivan, Timmermann & White, 1999; White, 2000; Park & Irwin, 2007), and the multiple-testing corrections now standard in the anomaly literature (Harvey, Liu & Zhu, 2016; Bailey & López de Prado, 2014) were developed for exactly that class of claim. The material examined here sits outside that literature. A body of intraday pattern instruction — order blocks, fair value gaps, liquidity sweeps, breaks of market structure, optimal trade entry retracements — circulates without peer-reviewed validation and, as usually taught, is not falsifiable: the entry is defined by what price did after it, the invalidation is discretionary, and the bar interval is selected after the outcome is known.

A subset of that material can be written as an exact predicate over open-high-low-close bars with an exact as-of time, and that subset is testable. In the course of testing it we measured something more portable than the verdict on the rules themselves. One of the detectors, the order block, admits two timestamps that differ only in which bar the signal is attributed to. One of them is causal and one reads the future. Everything else about the two runs is identical: the same bars, the same entry geometry, the same stop, the same target, the same maximum hold, the same per-instrument transaction costs, the same position sizing. The difference between them is therefore an unbiased estimate of what the timestamping error contributes, and it is large.

This paper reports that estimate and the procedure that isolates it. Section 2 states why the identification is retroactive. Section 3 describes the two independent causality checks, including the one that survives a detector lying about itself. Section 4 gives the design, Section 5 the results across thirteen cells, and Section 6 the mechanism by which the error produces an implausible win rate rather than merely an inflated

mean. Sections 7 and 8 give the implications for reading published results in this family and the limits of what was measured.

2. The order block and the timing of its identification

A displacement bar is one whose true range exceeds a multiple of the trailing average true range (Wilder, 1978); we use 1.5 times a fourteen-bar average computed strictly from the fourteen bars *before* the bar under test, and we exclude the overnight gap from true range at session boundaries, since including it makes the first bar of every session a displacement bar. An order block is the last candle of opposing colour within a short lookback preceding such a displacement.

The definition quantifies over a bar that lies in the future of the candle it names. Let p index the displacement bar and $j < p$ the last opposing candle before it. Candle j becomes an order block only at bar p, because until p completes there is no displacement for j to precede, and because a later opposing candle appearing between j and p would displace j from the role. The earliest bar at whose close the classification is available is therefore p, and the earliest legal entry is $p + 1$.

An annotated chart does not draw it that way. The rectangle is placed on candle j, which is where the entry is subsequently described as occurring, and a backtest written from that description rests its order from $j + 1$. Between $j + 1$ and p lies the displacement move itself, so an order resting there is filled by the very event that defines the pattern. The entry is therefore taken on a classification that does not exist when the order is placed.

> **The two timestamps.** Both versions read bar p; that is what the definition requires. They differ in the bar at which the signal is claimed to be knowable. The causal version claims p. The retroactive version claims j while still reading p, which is the discrepancy the enforcement in Section 3 detects.

3. Enforcing causality: a declaration and an independent replay

Every detector in the harness returns two indices per signal. The signal index is the bar at whose close the detector claims the pattern became knowable. The evidence index is the largest bar index the pattern definition reads. The harness refuses to trade any signal set in which the evidence index exceeds the signal index anywhere, and the refusal is total rather than per-signal: on the retroactive order block applied to the first instrument, all 17,135 of 17,135 signals declared evidence from a later bar and the whole set was rejected. Enforcement can be disabled explicitly, which is how the measurement in Section 5 was taken.

A declaration is only as reliable as the detector making it, and a detector that reports its evidence index incorrectly — through carelessness or otherwise — would pass. The second check does not use the declaration. For a signal claimed at bar s, the prefix replay re-runs the detector on the truncated series ending at s and again on the series ending at $s - 1$, and requires the signal to appear in the first and not in the second. Two distinct failures are separated. A signal absent from the series ending at s is not derivable at its claimed time, which is a causality violation. A signal already present in the series ending at $s - 1$ is derivable earlier than claimed, which is conservative rather than wrong but means the stated lag is not the real one.

The replay reads nothing the detector controls, so the two checks fail independently. The self-test includes a detector constructed to lie: the retroactive order block with its evidence index overwritten to equal its signal index. That detector passes the declaration check and is returned as look-ahead by the replay anyway. On the retroactive order block the replay returned six of six sampled signals not derivable at their claimed time; on the causal version and on every other detector in the study it returned causal.

> **Why both are needed.** The declaration check is cheap and covers every signal, which is what makes a blanket refusal possible. The replay is expensive, so it is sampled, but it is the one that has teeth against an incorrect declaration. Neither alone is sufficient: the first trusts the detector, and the second cannot be run over hundreds of thousands of signals.

4. Experimental design

Twenty-eight instruments were grouped into four tiers by measured round-trip transaction cost, and each tier was run at the bar intervals for which its cost admits a two-to-one target. The index and megacap tier holds ten names at fifteen-minute, one-hour and four-hour intervals and three exchange-traded funds at five minutes, the large tier six names, the small and mid tier nine names, and the cryptocurrency tier two pairs. The window is 2016-01-04 to 2022-12-31, comprising 1,762 equity sessions and 728 cryptocurrency days.

Entry is a limit order resting at the near edge of the order block, valid for twelve bars from the first legal entry bar and filled only if a bar trades through it. The stop is the far edge of the block and the target is twice the stop distance. Maximum hold is twenty bars. When a single bar spans both the stop and the target the stop is taken, which is the conservative resolution and a material one: at a two-to-one target on intraday bars a meaningful fraction of bars contain both, and resolving them favourably is on its own sufficient to produce a positive expectancy. Position size is one percent of capital at risk in whole shares.

Transaction costs are charged per instrument from a measured round-trip table rather than as a flat figure, because a flat figure inverts the tier ordering. The measured round trip for the S&P 500 exchange-traded fund is 0.355 basis points, the one-cent tick floor; for the larger cryptocurrency pair it is 20.93 basis points, fifty-nine times as wide. Costs enter the R metric by division by the stop distance, so a fixed basis-point toll is a larger fraction of an R when the stop is tight. Statistics are means of net R per trade with a day-clustered Newey-West standard error (Newey & West, 1987), the clustering applied because signals on the same session across ten instruments are not independent.

The retroactive and causal runs differ in the signal index and in nothing else. Both read the displacement bar, both rest at the same price, both stop at the same price, both target the same multiple, both are charged the same table and sized by the same rule. Trade counts differ because the set of resting orders that fill differs, which is itself part of the effect and is taken up in Section 6.

5. Results

Table 1 gives both versions on all thirteen cells. The final two columns are the difference attributable to the timestamp: manufactured R per trade and manufactured win rate in percentage points.

TIER	BARS	RETRO N	RETRO R	RETRO HIT	CAUSAL N	CAUSAL R	CAUSAL HIT	MANUF. R	MANUF. PP
Index / megacap	5 min	35,297	+1.332	83.7%	32,869	−0.023	35.7%	+1.355	+48.0
Index / megacap	15 min	53,684	+1.158	78.4%	41,673	+0.021	37.8%	+1.136	+40.6
Index / megacap	1 hour	14,119	+1.158	76.0%	10,966	+0.049	36.7%	+1.109	+39.3
Index / megacap	4 hour	3,829	+1.078	71.0%	3,559	+0.067	36.7%	+1.011	+34.3
Large	15 min	28,967	+0.896	74.5%	23,753	−0.035	38.4%	+0.930	+36.1
Large	1 hour	7,754	+0.996	74.0%	6,240	+0.015	36.8%	+0.981	+37.2
Large	4 hour	2,225	+1.118	74.3%	2,045	+0.037	36.4%	+1.081	+37.9
Small / mid	15 min	44,846	+0.439	68.0%	39,466	−0.189	36.9%	+0.628	+31.0
Small / mid	1 hour	12,111	+0.736	71.8%	10,165	−0.067	36.6%	+0.803	+35.2
Small / mid	4 hour	3,500	+1.028	74.5%	3,269	+0.001	36.5%	+1.027	+38.0
Crypto	15 min	17,386	−1.079	52.8%	13,010	−0.571	33.7%	−0.508	+19.0
Crypto	1 hour	5,259	+0.170	72.3%	3,677	−0.224	36.2%	+0.393	+36.1
Crypto	4 hour	1,612	+0.747	77.6%	1,061	−0.148	34.3%	+0.895	+43.3

Table 1. Retroactive and causal order block on identical bars, 2016–2022. R is mean net of measured per-instrument round-trip costs; hit is the share of trades reaching a two-to-one target before the stop. Median manufactured R is +0.98 and median manufactured win rate is +37.2 points.

Win rate is inflated in thirteen cells of thirteen, by between 19.0 and 48.0 percentage points, median 37.2. Mean net R is inflated in twelve of thirteen, by between +0.39 and +1.36, median +0.98. The causal win rates occupy a narrow band from 33.7% to 38.4% around the 33.3% that a two-to-one target returns under no edge, which is the shape a rule with no information content takes when its geometry is fixed. The retroactive win rates run from 68% to 84% in the equity tiers, a range no two-to-one rule can occupy for structural reasons.

The *t* statistics separate as sharply as the point estimates. Day-clustered Newey-West *t* on the index and megacap tier is 144.1 at five minutes, 108.4 at fifteen minutes, 65.9 at one hour and 33.0 at four hours under the retroactive timestamp; the corresponding causal values are −1.23, 1.87, 0.69 and 2.46. Across all thirteen cells no causal *t* reaches 3, and ten of thirteen retroactive *t* statistics exceed 26. A test statistic in the hundreds on a sample of fifty thousand intraday trades functions as a diagnostic that the trades are not what they claim to be, rather than as evidence of a large effect.

> **Magnitude relative to the study's real effects.** The largest genuinely causal effect measured anywhere in the parent study was +0.13 R per trade on one detector, one tier and one bar interval, and it survived a single sealed holdout. The timestamping error contributes roughly an order of magnitude more than that, in every equity cell, with no selection and no tuning. Any procedure capable of producing this error dominates whatever the rule itself does.

6. Mechanism

The inflated mean follows from the inflated win rate, and the inflated win rate follows from which resting orders fill. Under the causal timestamp the limit rests from the bar after the displacement, so it fills only if price returns to the block edge, which requires a

retracement against the move. Under the retroactive timestamp the limit rests from the bar after the block candle, so it is filled during the displacement itself. The fill and the event that defines the pattern are the same event.

That selection changes the R denominator, which is the block's own range. Re-deriving both versions on fifteen-minute bars of the S&P 500 exchange-traded fund over the same window, the median stop distance among filled trades is 0.96 average true range under the causal timestamp and 0.33 under the retroactive one. The retroactive fills are concentrated on narrow blocks, because a narrow block is exactly the one price runs through on its way out rather than returning to. A displacement bar has true range above 1.5 average true range by construction, so a target set two block-ranges away, where the block is a third of an average true range, sits inside the move already in progress. The trade is therefore resolved favourably by the same bar that identified it.

The single negative row in Table 1 is explained by the same mechanism rather than being an exception to it. The cryptocurrency fifteen-minute cell shows the usual win-rate inflation of 19.0 points, so the look-ahead is present, but the tight retroactive stop makes a fixed basis-point toll an enormous fraction of an R: the causal version of that cell already pays 0.589 R per trade in costs, the largest in the grid, and the retroactive version pays more because its denominator is smaller. The look-ahead is uniform across the grid; the toll is not, and on the widest instrument in the sample it buries the manufactured gross edge. Reading that row as an absence of the error would invert its meaning.

7. Implications for the published record

The measurement bounds what can be inferred from a reported result in this family without access to its code. A two-to-one target admits a maximum win rate consistent with an edge of plausible size; the observed causal band is 33.7% to 38.4%, and a rule genuinely earning +0.1 R per trade would sit near 37%. A quoted win rate of 75% to 85% at that reward-to-risk ratio corresponds instead to the timestamp error, at the magnitudes Table 1 records: 78.4% at fifteen minutes, 83.7% at five minutes, 76.0% at one hour.

Three properties of the error make it durable in published work. It requires no data snooping, so it survives out-of-sample splits, walk-forward procedures and holdout seals unchanged — the fake edge is present in every window because it is a property of the detector rather than of the sample. It produces test statistics far above any multiple-testing bar, so deflation procedures (Bailey & López de Prado, 2014; Harvey *et al.*, 2016) pass it without difficulty. And it grows rather than shrinks with sample size, since it is a bias in the estimand and not variance in the estimate. A backtesting protocol that checks only for overfitting (Arnott, Harvey & Markowitz, 2019; López de Prado, 2018) will not catch it.

The check that does catch it is available to any reader with the code: re-derive each signal from the bar history truncated at the bar the signal claims, and confirm it is present there and absent one bar earlier. Where the code is not available, the reported win rate at a stated reward-to-risk ratio is already diagnostic, because the two quantities are jointly constrained.

8. Limits

The measurement covers one asset class and one detector family. All thirteen cells are intraday United States equities and two cryptocurrency pairs over 2016–2022; nothing here establishes the magnitude on other markets, other periods, or daily and higher intervals, though the mechanism in Section 6 is not specific to any of them. The order block is the detector for which the retroactive convention is cleanly instantiated, because its definition quantifies over a strictly later bar; other members of the family are retroactive to varying and less separable degrees, and we did not attempt to quantify them.

The instrument sample is a live-name sample selected in 2026 and is survivorship-selected by construction. That bias applies to both arms identically and cancels in the difference, which is the quantity reported, but it does apply to the levels in either arm taken alone. The

five-minute index cell rests on three exchange-traded funds and the cryptocurrency tier on two pairs, so those rows are thinner in cross-section than their trade counts suggest. The prefix replay is sampled rather than exhaustive; the declaration check is exhaustive.

The causal arm is not a finding. Its +0.022 R per trade at fifteen minutes carries a t of 1.87 against a multiple-testing bar above 3, and its counterparts in the other twelve cells range from −0.571 to +0.067 with no t reaching 3. The causal numbers are reported as the reference against which the error is measured and should not be read as an estimate of a small positive edge in the rule. Nothing in this paper was traded, no allocation or risk limit was changed on the basis of it, and no position is taken on whether the rule family has merit under any other specification.

References

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

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

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

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

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

López de Prado, M. (2018). *Advances in Financial Machine Learning*. Hoboken, NJ: Wiley.

Newey, W. K. & West, K. D. (1987). A simple, positive semi-definite, heteroskedasticity and autocorrelation consistent covariance matrix. *Econometrica*, 55(3), 703–708.

Park, C.-H. & Irwin, S. H. (2007). What do we know about the profitability of technical analysis? *Journal of Economic Surveys*, 21(4), 786–826.

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

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

Wilder, J. W. (1978). *New Concepts in Technical Trading Systems*. Greensboro, NC: Trend Research.

Andreev, M. A. & Howden, J. (2026e). Ontological limits of historical validation for wavefunction-calibrated financial systems. *Quark Quantitative Research Working Paper*.