



Do Options-Flow Signatures of Dealer Hedging Predict Price Direction Around Earnings?

Zhixing Han

10/15/26



Abstract

As investors trade options around a company's earnings announcement, the dealers who take the opposite side of the trade end up with exposure they didn't want, and normally offset it with an opposite position in the underlying shares. This mechanical hedging is not driven by investor sentiment, but may manifest itself as pressure on the short-term direction of a stock's price. This study combines four options-flow indicators (Volatility Smirk, Delta Difference, Open Interest Ratio, and Option Volume Ratio) representing four different types of dealer hedging pressure. Every indicator is referenced to a pre-earnings baseline specific to the ticker, and the four indicators are averaged into a composite indicator (Agg Score) that is weighted by each indicator's accuracy during the baseline. The model was tested on 30 earnings events from a stratified random sample of stocks across six sectors and a range of sizes, backtested with respect to the next-day price change direction on the 15 trading days around each earnings release. The composite achieved a hit rate of 48.16%, which was not significantly better than random (permutation test $p = 0.740$), and did not significantly outperform a naive momentum baseline (50.00%; McNemar $p = 0.619$) across 434 evaluations. None of the near- or far-from-earnings sub-periods, nor any single indicator, was significant, and accuracy was not focused around the time of the earnings release as predicted. These findings indicate that options-flow signals have little directional power around earnings, and that dealer hedging may impact volatility more than direction.

Keywords: options flow, dealer hedging, volatility smirk, open interest ratio, earnings announcements, composite z-score model, permutation test

1. Introduction

Options markets are not just indicative of expectations; they also establish a duty on the part of those who play the other side of the trade, the dealers, or market makers. A dealer who sells an option normally hedges it by buying or selling the underlying stock, which is referred to as delta hedging. This paper investigates whether delta hedging, visible through patterns in the flow of options, can be used to forecast a stock's near-term direction around an earnings announcement.

The question is important to a number of groups. The volume of options traded through retail accounts has recently surpassed 60% of the total options market volume (Bryzgalova, Pavlova, & Sikorskaya, 2023), making it important to many individual traders to be aware of whether options data has predictive power or if it is merely sentiment. Market microstructure scholars obtain insights into the relationship between dealer positioning and price dynamics, while risk desks and dealers are directly affected by the overall impact of dealers' hedges on the market.

In this paper, the author presents a composite measure (Agg Score) synthesizing four options-chain measures: Volatility Smirk measures the magnitude of dealer hedging pressure, Delta Difference measures its direction, Open Interest Ratio measures its concentration at nearby strikes, and Option Volume Ratio measures its immediacy. The model does not give equal weight to the measures, but rather assigns a weight based on the accuracy of each measure during a pre-earnings baseline period.

Four hypotheses are presented with the study:

H1: The Agg Score predicts the next day's price direction more accurately than chance.

H2: Accuracy is highest in the sessions closest to the earnings release, consistent with the theory that hedging pressure is greatest in those sessions.

H3: The Agg Score is more accurate than any of its four indicators alone.

H4: The Agg Score is more accurate than a naive momentum baseline assuming that each stock keeps moving in the same direction as the previous day.

All four hypotheses are tested on the 15 trading days surrounding 30 earnings events, using permutation tests against chance, a near- versus far-from-earnings comparison and a day-by-day accuracy comparison for H2, and direct method comparisons for H3 and H4. None of the four was supported. The prior research is reviewed in Section 2, linkages between each indicator and dealer hedging are made in Section 3, the data and methodology are described in Sections 4 and 5, significance tests and the benchmark comparison are described in Section 6, the findings are described in Section 7, limitations are discussed in Section 8, and conclusions are drawn in Section 9.

2. Literature Review

Studies have demonstrated that options markets can provide information earlier than stock prices, in part because informed traders may choose to trade options rather than stock. Three of these studies are directly related to individual indicators in this paper, while a further line of research documents the dealer-hedging mechanism behind all four.

As stated by Xing, Zhang and Zhao (2010), the volatility smirk can be defined as the difference in implied volatility between out-of-the-money puts and at-the-money calls. On a risk-adjusted basis, on a stock-by-stock basis from 1996 to 2005, stocks with the steepest smirks underperform those with the least pronounced smirks by about 10.9% per year, and steep-smirk firms have the largest negative earnings shocks the following quarter. The authors

attribute this to informed traders with negative news buying out-of-the-money puts and to the slow response of the equity market to that information. This paper's Volatility Smirk follows the same structure but applies it day to day around individual earnings announcements. Importantly, while Xing, Zhang, and Zhao document weeks-months of predictability, this paper tests next-day direction, and given that the smirk's information is incorporated into prices gradually, it may be absent at a 1-day horizon, a likely reason this indicator fails to show predictability here (Section 7c).

Using intraday data, Easley, O'Hara, and Srinivas (1998) model informed traders who decide whether to trade in options markets or stock markets and conclude that volume in the option markets can predict prices of the underlying stocks, especially when split into positive volume (buying calls or selling puts) and negative volume (buying puts or selling calls). This spurs the Option Volume Ratio. Since there is no indicator of trade direction in this paper's data, however, the ratio can only distinguish puts from calls, and in doing so it loses the information that Easley, O'Hara, and Srinivas demonstrate is vitally important: raw volume of puts combines put buying (bearish) with put selling (bullish). This may be the reason for the null result for this indicator (Section 7c).

Fodor, Krieger, and Doran (2011) find that changes in open interest predict returns: large increases in call open interest lead to stronger returns, whereas increases in put open interest lead to weaker returns (less robustly after controls), and the recent change in the call-to-put ratio is a predictor of returns over the next week. This reinforces the directionality of the Open Interest Ratio, albeit this paper is based on the ratio's level (not change). The level fits the dealer-hedging interpretation (Section 3c) because the hedging adjustments that dealers might make depend on the total number of positions close to the current price, not on how much that number changed since the previous day. Daily change is also unreliable here, due to mechanical jumps in open interest caused by contract rolls (Section 4c), and because new same-day activity is already reflected in the Option Volume Ratio. A change-based version is a natural extension.

There is literature dedicated to the underlying mechanism, dealer delta hedging. As demonstrated by Ni, Pearson, and Poteshman (2005), the closing prices of optionable stocks tend to cluster around strike prices on expiration days, driven largely by delta-hedge rebalancing, which changes returns by at least 16.5 basis points per expiration day. Beyond expiration dates, Ni, Pearson, Poteshman, and White (2021) discover that there is a significant negative relationship between stock return volatility and the net purchased option positions of likely hedgers: hedging reduces volatility when hedgers are net long options, and increases volatility when hedgers are net short options. The results of these studies are that dealer hedging influences stock prices, primarily by size of move and not by direction of move. This paper's Delta Difference indicator (Section 3b) tests whether the same mechanism also carries directional information around earnings.

All this literature implies that the price of options, the volume of trading, and standing open interest each contribute information on future stock prices, and that the actions of dealers in hedging influence the price of the stock, especially its volatility. To the author's best knowledge, no study has merged these four signals together or has examined whether such a signal predicts direction in the window around earnings announcements. This paper fills in that gap.

3. Background / Theoretical Framework

If an options dealer sells options to satisfy the demand of the customer, the dealer will be exposed and will need to hedge the exposure by trading the underlying asset. The strongest evidence for market effects of this hedging comes from Ni, Pearson, Poteshman and White (2021), who discover a significant negative relationship between the net purchased option positions of likely hedgers and stock return volatility. In foreign-exchange options, Anderegg, Ulmann and Sornette (2022) report consistent evidence that dealers' net-short gamma positions significantly increased the volatility of spot rates. In this paper, four option-flow indicators are proposed as proxies for different aspects of this hedging pressure: Volatility Smirk (magnitude), Delta Difference (direction), Open Interest Ratio (concentration), and Option Volume Ratio (immediacy). The definitions of the underlying metrics are in Appendix A.

3a. Volatility Smirk

Dealers are usually net short the out-of-the-money (OTM) puts that investors buy for downside protection. A dealer with a short put will hedge with a short stock position that must increase as the stock price declines toward the strike. As put implied volatility rises relative to call implied volatility, dealers are demanding more compensation for this downside risk, signaling that they are more likely to be forced into further selling if a sell-off starts. A high Smirk is thus seen as bearish, while a low or negative Smirk is seen as bullish, although negative Smirks seldom occur and most literature concentrates on high smirks.

This interpretation differs from Xing, Zhang and Zhao (2010), who attribute the smirk's predictive power to informed traders who purchase puts when the news is negative. Both explanations treat a high Smirk as bearish, but for different reasons: in the former, it reveals negative information not yet reflected in prices; in the latter, it signals that dealers may be forced to sell. This paper does not make any distinction between them.

3b. Delta Difference

Delta is often used as the approximate probability that an option will finish in the money. Dealers typically short the most heavily traded contracts, the call one strike above and the put one strike below the current price, holding long stock against the short call and short stock against the short put. If the call side dominates, dealers must buy more stock as the price rises, supporting higher prices; when the put side dominates, they must sell stock as the price falls, supporting lower prices. The idea of Delta Difference (call delta minus absolute put delta) is to see which side dominates, with a high number reflecting a bullish outlook.

An important caveat is that it only employs deltas. Delta Difference is based on the price sensitivity of each option, not on the number of contracts held by the dealers, so Delta Difference does not indicate which side the dealers hold more of. The two deltas at the strikes just above and below the current price primarily indicate where the price sits between the two strikes: the call delta increases as the price moves closer to the upper strike, and the absolute put delta decreases, meaning that the indicator increases regardless of how dealers are positioned. Because Delta Difference is running on price location and not on dealer positioning, this could be the reason it is the least accurate indicator (39.63%) among all the indicators (Section 7c). This could be solved by weighting each delta by the open interest at its strike.

3c. Open Interest Ratio

Open interest is the total of positions that have been added over time. If put open interest below the current price is high compared to call open interest above it, more short-put hedges will require dealers to sell stock if the price falls toward the put strike. However, heavy open interest does not necessarily mean that the price has a definite direction, as it can keep a price

pinned at the strike or drive it through, depending on the net position of the dealers. This paper nonetheless follows Fodor, Krieger and Doran (2011) in treating a high Open Interest Ratio as bearish and a low one as bullish.

3d. Option Volume Ratio

Volume reflects same-day trading activity. Increased volume for at-the-money put options compared to call options may indicate that dealers are taking on more short puts and selling stock to hedge that exposure during the session. Volume does not show whether a trade opened or closed a position, or the dealer's side of the trade, so this interpretation is uncertain, but a high Option Volume Ratio is considered bearish and a low one is considered bullish. It is the most current of the four signals.

3e. The Dealer-Short Assumption

The interpretations assume that the dealers are net short the options that are traded by the customers. If dealers are net long, the direction would switch or not work. The assumption cannot be confirmed based on open interest, volume and delta information because these do not provide any insight into the dealer's side of trades (Section 8b).

3f. Direction Versus Magnitude

The indicators are used in a directional manner, but the best evidence regarding dealer hedging is on the size of price moves, rather than their direction. Net-short dealers must sell after declines and buy after gains, amplifying moves in either direction, while net-long dealers dampen them (Ni, Pearson, Poteshman, & White, 2021), and hedging tends to push prices toward strikes on expiration dates, rather than in a consistent direction (Ni, Pearson, & Poteshman, 2005). A dealer's hedge might thus be more helpful in predicting the size of moves than their direction. Since this paper only tests direction, this is likely a prime cause of the non-significant results (Section 7) and is a justification for testing volatility instead (Section 9).

In summary, high values of Volatility Smirk, Open Interest Ratio and Option Volume Ratio are bearish, and a high value of Delta Difference is bullish (Section 5b).

4. Data

4a. Source & Parsing

All options chain data was manually recorded from the options chain page on Thinkorswim's platform into a spreadsheet, one row for each ticker for each trading day. Some fields use trailing signs (e.g., "12.5-" for negative, "0.02+" for positive), a "(%)" sign, or a thousands separator, which are not interpreted by a standard parser; custom parsing functions (Appendix B) were used to identify these values, convert them to signed numbers, and remove the percentage signs and thousands separators. Entries recorded as "N/A" or left blank were considered missing, and one malformed entry (a letter typed instead of a decimal point) was corrected. Analysis was conducted in Python (pandas and NumPy) and the figures were created in Matplotlib.

4b. Snapshot Timing & Earnings-Day Alignment

Each value was taken at 12:45 PM Pacific Time (3:45 PM Eastern), 15 minutes before the close, and the stock price was taken at the same time, so each return is not between closing prices, but between the stock price at this snapshot and the stock price at the same snapshot on the next trading day. Because open interest updates only once a day, the Open Interest Ratio reflects positions taken through the previous day, while volume shows all trading up to the time of the snapshot, so the other three indicators reflect the current day. Day counts refer to trading days.

Day E is the first trading session in which the price can reflect earnings: for Before-Market-Open (BMO) releases, Day E is the release day, and for After-Market-Close (AMC) releases, Day E is the trading day after the release. The windows are E-17 through E+7 with an additional price at E+8 for the last return.

4c. Contract Selection & Expiration-Day Exclusion

For each day's indicators, the monthly contract nearest to expiration on that day was used (excluding weeklies), since monthly contracts offer higher volume and open interest and greater stability in ratios. If this contract expires, the next monthly contract is applied, which means that a window can span two contracts. If the strike an indicator requires (e.g., the 95% put strike) had no data, it was replaced by the closest strike with data, with the same rule applied to all four indicators.

All trading days that fell on monthly expiration dates (March 20, April 17 and May 15, 2026) were excluded for mechanical reasons, as open interest collapses and implied volatility distorts on those days. This excluded 27 of 300 baseline days and 16 of 450 evaluation days. Only indicator values are omitted; the excluded day's price is still used for the previous day's return. Each ticker provides 9 or 10 usable baseline days (E-17 to E-8) and 14 or 15 evaluation days (E-7 to E+7).

4d. Sample Construction

Stratified random sampling was used to select tickers by both sector and company size. The universe was all U.S.-listed optionable stocks with earnings releases in April or May 2026. Sectors are defined in accordance with Yahoo Finance categories, and Consumer Cyclical and Consumer Defensive have been grouped together; Sunrun (RUN), which Yahoo Finance classifies under Technology (Solar), was moved to the Energy sector as its business is residential solar generation. Yahoo Finance market capitalization on April 20, 2026 was used to determine size. The largest companies vary widely by sector (no U.S. energy company is worth more than \$1 trillion), so within each of the six sectors, the eligible companies were split into 5 groups by size relative to the other companies in that sector, and one ticker was drawn at random from each of the resulting 30 groups, producing a mix of companies from under \$5 billion to more than \$1 trillion. This design reduces the likelihood that results are driven by a particular sector or size range.

The sample includes 30 tickers that each have one earnings event (sectors and size ranks in Appendix C): GS, FUBO, LLY, U, ABBV, GM, JPM, XOM, CVX, PFE, SOFI, DIS, RIVN, NVDA, OXY, BBAI, SNAP, META, AMZN, A, COST, NFLX, RUN, GRAL, KLAC, DVN, QCOM, UPST, V and SG. Earnings dates are between April 13 and May 29, 2026, and data collection, including baselines, took place from March 18 to June 10, 2026.

5. Methodology

5a. Indicator Construction & Baseline Statistics

The four indicators are calculated as follows:

- Volatility Smirk = put implied volatility (95% strike) – call implied volatility (at-the-money strike)
- Delta Difference = call delta (strike immediately above at-the-money) – |put delta (strike immediately below at-the-money)|
- Open Interest Ratio = (put open interest + 1) ÷ (call open interest + 1), using the strikes immediately below and above at-the-money
- Option Volume Ratio = (put volume + 1) ÷ (call volume + 1), at the at-the-money strike

The +1 terms (additive, or Laplace, smoothing) keep both ratios defined when the call-side value is zero. Each ticker's baseline window (E-17 to E-8, excluding expiration days) precedes all evaluation days, preventing look-ahead bias. Within it, the mean (m) and sample standard deviation (s , with Bessel's correction) of each indicator are computed per ticker:

$$s = \sqrt{[\sum (x_i - m)^2 \div (n - 1)]}$$

5b. Seven-Tier Scoring

Each indicator's values are divided into seven tiers, from +3 to -3, at ± 0.5 , ± 1.0 , and ± 1.5 standard deviations from its baseline mean:

Tier	Range	Label
+3	$x \geq m + 1.5s$	Extremely bullish
+2	$m + 1.0s \leq x < m + 1.5s$	Strongly bullish
+1	$m + 0.5s \leq x < m + 1.0s$	Bullish
0	$m - 0.5s \leq x < m + 0.5s$	Neutral
-1	$m - 1.0s \leq x < m - 0.5s$	Bearish
-2	$m - 1.5s \leq x < m - 1.0s$	Strongly bearish
-3	$x < m - 1.5s$	Extremely bearish

Each tier includes its lower and excludes its upper boundary. Because high values of Volatility Smirk, Open Interest Ratio, and Option Volume Ratio reflect downside hedging pressure (Section 3), their scores are inverted (e.g., tier +3 is scored -3); Delta Difference is not inverted.

For a single indicator, only whether a score is positive, negative, or zero matters (Section 5c), so the tiers reduce to bullish, neutral, and bearish. Their main purpose is in the composite, where bounded tiers cap each indicator's contribution at ± 3 so that no single extreme value dominates. This matters because standard deviations estimated from only 9 or 10 baseline days are noisy and occasionally very small, which can turn modest changes into extreme z-scores (Section 8d).

5c. Hit-Rate Scoring Rules

Each prediction compares the signal on day t with the return from t to $t+1$ (Section 4b):

- Directional Match: a positive score is a hit if the return is positive, and a negative score is a hit if the return is negative; otherwise, including a return of exactly 0%, it is a miss.
- Neutral Zero Rule: a score of exactly 0 is a hit if $|\text{return}| \leq 1\%$.

These rules are used to set the weights (Section 5d) and to score the individual indicators (Section 7c), the Agg Score (Section 7a), and the momentum baseline (Section 6b).

An alternative would treat a score of 0 as no prediction. This paper scores neutral predictions because a neutral tier is a substantive prediction that hedging pressure is near baseline and no large move should follow; excluding them would ignore much of the model's output and reward indicators for predicting neutral more often. The cost is that methods

producing many neutral scores have lower chance-level accuracy (Section 7c), so each method is compared with its own permutation null distribution (Section 6a) rather than a fixed 50%.

5d. Accuracy-Based Weighting & the Agg Score

Rather than weighting indicators equally, the model weights each by its accuracy (A_i) on the ticker's baseline days, scored with the rules in Section 5c. Because the tiers are built from the same days, this accuracy is in-sample. Weights are normalized per ticker to sum to 1:

$$W_i = A_i \div (A_Smirk + A_Delta + A_OI + A_Option)$$

Every indicator receives a positive weight proportional to its accuracy, even at or below chance; if all four accuracies are zero, weights default to 25% each. On each evaluation day, the composite signal is:

$$\text{Agg Score} = (S_Smirk \times W_Smirk) + (S_Delta \times W_Delta) + (S_OI \times W_OI) + (S_Option \times W_Option)$$

where S_i is each indicator's tier score. If an indicator is missing, the Agg Score is the weighted sum of the available indicators divided by the sum of their weights, so missing values do not pull the score toward zero. The Agg Score ranges from -3 to $+3$, is rounded to four decimals, and is scored like a tier score: positive predicts an increase, negative a decrease, and exactly 0 is scored with the Neutral Zero Rule.

6. Validation

6a. Significance Testing

The accuracy of the composite was tested against chance using a permutation test. All (Agg Score, next-day return) pairs were combined into 434 observations ($30 \text{ tickers} \times 15 \text{ days} = 450$, minus 16 expiration days; Section 4c), and the real hit rate was calculated using the rules in Section 5c. A null distribution was created by shuffling the next-day returns 1,000 times across the entire pooled sample of returns, thus allowing a return to be matched with any ticker's Agg Score, and calculating the hit rate each time. This eliminates any relationship between signal and outcome while retaining identical signals and returns. The p -value is the proportion of shuffled hit rates that are equal to or better than the real hit rate, and the test is one-sided (better than chance); significance is set at $p < 0.05$. A normal approximation to the binomial distribution is used for 95% confidence intervals of hit rates.

Shuffles did not use any fixed random seed, so repeated tests yield slightly different results. Each test was therefore repeated 20 times; the p -values varied by no more than 0.08 (e.g., 0.718–0.767 for the pooled Agg Score test), and the lowest p -value from any test in any repetition was 0.426, which is still well above 0.05 (Appendix D). A single run is reported in Section 7.

Three additional analyses were performed:

1. *Proximity to earnings*. As a pre-specified secondary test of H2, the data were divided into two groups, far from earnings (E-7 to E-4 and E+4 to E+7; 226 observations) and near earnings (E-3 to E+3; 208 observations), each tested using the same permutation test. No test was applied to the difference between the groups; it was compared descriptively.
2. *Individual indicators*. The test was replicated for each indicator's tier scores (434 observations per indicator).
3. *Day-by-day accuracy*. Hit rates were calculated separately for each of the 15 day-positions, pooling all tickers. Each day-position has at most 30 observations, so these results can only be descriptive (Section 8d).

6b. Benchmark Comparison

The model was compared to a naive momentum baseline, a simple price rule, to see if the model provides added value over it. Define $r(t-1, t)$ as the return from the snapshot price on day $t-1$ to day t . If $r(t-1, t) > 0$, the baseline predicts a positive direction; if it is < 0 , it predicts a negative direction; and if it is exactly 0, it predicts neutral. Returns use only each ticker's own prices, and E-7's previous return is based on the E-8 price, thus providing a prediction for every evaluation day. Predictions are evaluated using the same rules as the Agg Score (Section 5c) on the same 434 evaluations, so differences in accuracy cannot be attributed to differences in sample. Like the Agg Score, the baseline rarely predicts neutral, so both have a chance-level hit rate near 50%, allowing their raw hit rates to be directly compared.

The baseline is tested with the same permutation test (Section 6a), where its predictions are kept fixed and the returns are shuffled instead. The two methods are also compared directly with an exact McNemar test, which only considers the days on which exactly one method is correct and asks whether either method is correct on significantly more of those days than the other (two-sided, $p < 0.05$).

7. Results

7a. Composite Model Accuracy

Table 1. *Composite model hit rate, null distribution, and significance results for the overall pooled test and the far-from-earnings and near-earnings sub-periods.*

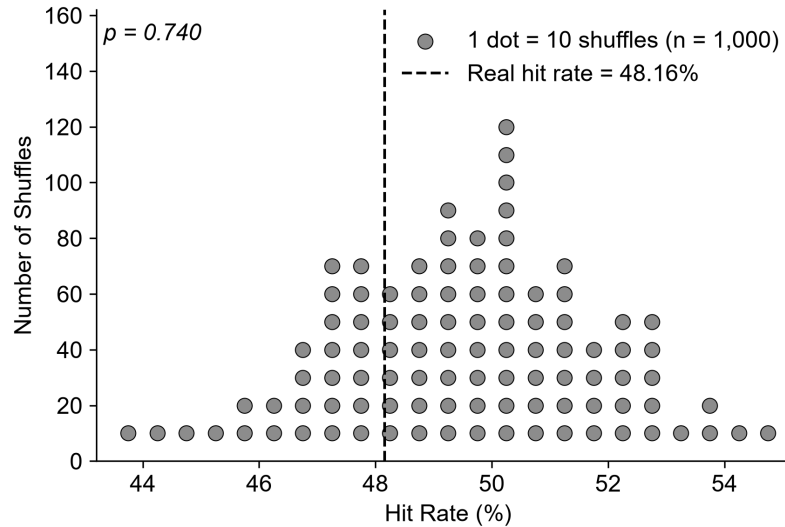
Analysis	Window	n	Hit rate	Null mean	Null SD	p	Significant
Overall pooled	E-7 to E+7	434	48.16% (209/434)	49.52%	2.27%	0.7400	No
Far from earnings	E-7 to E-4, E+4 to E+7	226	50.88% (115/226)	50.30%	3.20%	0.4360	No
Near earnings	E-3 to E+3	208	45.19% (94/208)	48.90%	3.42%	0.8710	No

Note. Hit rate = hits ÷ valid evaluations. Null mean and SD are from 1,000 permutations of next-day returns across all companies; p is the proportion of permutations with a hit rate at least as high as the observed rate. Monthly options-expiration dates were excluded.

Across the full evaluation window (E-7 to E+7), the composite Agg Score achieved a hit rate of 48.16% (209 of 434; 95% CI: 43.5% to 52.9%). By proximity to earnings, accuracy was 50.88% (115 of 226; 95% CI: 44.4% to 57.4%) in the far-from-earnings window (E-7 to E-4, E+4 to E+7) and 45.19% (94 of 208; 95% CI: 38.4% to 52.0%) in the near-earnings window (E-3 to E+3). All three intervals include the model's chance-level accuracy, and therefore the results fail to support H1.

7b. Significance Testing Results

Figure 1. Quantile dot plot of the overall pooled test's null distribution, showing the composite model's real hit rate (dashed line) relative to 1,000 randomly shuffled results.



Note. Each dot represents 10 of the 1,000 shuffled hit rates, placed at the median of its group after sorting. The dashed line marks the observed hit rate (48.16%). The p -value ($p = 0.740$) is calculated from all 1,000 shuffled results, not the grouped dots.

None of the three analyses in Table 1 reached significance ($p < 0.05$) under the permutation test (Section 6a). The real hit rate of this model is in the middle of the null distribution as shown in Figure 1, not in its upper tail.

The lowest p -value ($p = 0.436$) was for the far-from-earnings period, while the highest ($p = 0.871$) was for the near-earnings period. This does not support H2: accuracy was lower near the release (45.19%) than away from it (50.88%), although this difference was not statistically tested. The most likely explanation is the earnings announcement itself. The near-earnings window contains the E-1 to E return, reflecting the market's reaction to the news, and options can price the expectation of a big move, but not the direction of the surprise, so weak directional accuracy around the news is expected (Section 3f). Two other reasons could be involved: tiers calibrated on a calmer baseline period may not perform well in a more volatile one, and weights derived from a small per-ticker sample may not be representative of the days near earnings (Section 8d).

7c. Individual Indicator Accuracy

Table 2. *Hit rate, null distribution, excess over null, and significance results for the composite Agg Score and each of the four individual indicators, pooled across the evaluation window.*

Indicator	Direction	<i>n</i>	Hit rate	Null mean	Excess over null	Null SD	<i>p</i>	Significant
Agg score (composite)	–	434	48.16% (209/434)	49.52%	–1.36	2.27%	0.7400	No
Volatility smirk	Inverted	434	43.55% (189/434)	44.80%	–1.25	1.81%	0.7710	No
Delta difference	Direct	434	39.63% (172/434)	44.92%	–5.29	2.17%	0.9960	No
Open interest ratio	Inverted	434	44.24% (192/434)	44.71%	–0.47	2.27%	0.6040	No
Option volume ratio	Inverted	434	43.32% (188/434)	43.11%	+0.21	2.10%	0.4710	No

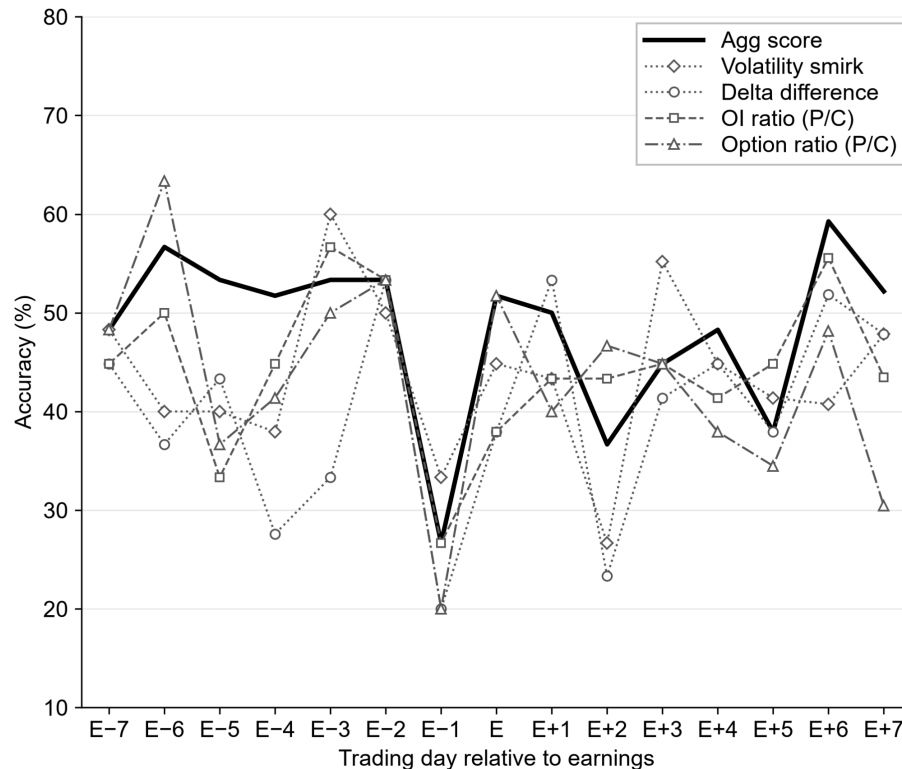
Note. Excess over null is the hit rate minus the null mean, in percentage points. Null distributions and *p*-values are as in Table 1; hits follow the rules in Section 5c. "Inverted" indicators are read as bearish when high.

None of the indicators was significant. Volatility Smirk, Open Interest Ratio, and Option Volume Ratio each performed close to their null means (43.55% vs. 44.80%, 44.24% vs. 44.71%, and 43.32% vs. 43.11%). Delta Difference was the weakest, falling well below its null mean (39.63% vs. 44.92%; $p = 0.996$). This aligns with the construction challenge in Section 3b: since it is based on deltas without open interest, Delta Difference mostly tracks where the price sits between strikes rather than how dealers are positioned, and may not reflect hedging pressure at all.

Methods should be compared using their excess over null, not their raw hit rate, since the Neutral Zero Rule results in a lower chance level for methods that are likely to predict neutral (Section 5c). The individual indicators often score 0, so their chance levels are about 43–45%, while the Agg Score is rarely 0 (18 of 434 evaluations; chance level 49.52%). Its raw hit rate is, indeed, higher than each indicator's, but its excess over null is no better than theirs (–1.36 versus –5.29 to +0.21). The results do not support H3, possibly because only about 10 baseline days are used for the weights, so noise may outweigh signal (Section 8d).

7d. Day-by-Day Accuracy by Day-Position

Figure 2. Descriptive accuracy (%) by day-position ($E-7$ to $E+7$) for the composite model (Agg score) and each of the four individual indicators, pooled across 30 companies.



Note. Each point is the share of companies whose day- t signal correctly predicted the return from t to $t+1$, so the $E-1$ point reflects the move into the earnings day (E). Sample size ranges from 23 to 30 companies per day because of expiration-date exclusions. Chance accuracy is about 45% for the individual indicators and about 50% for the composite. Values are descriptive; no per-day tests were performed.

There was a substantial amount of variation in accuracy over the 15 days but no clear pattern (Figure 2). The Agg Score ranged from 26.7% at $E-1$ to 59.3% at $E+6$. The worst day for nearly all five series was $E-1$: it was the lowest day for the composite, Delta Difference, Open Interest Ratio, and Option Volume Ratio (20.0–26.7%) and the second lowest for Volatility Smirk (33.3%), before accuracy rebounded at E and $E+1$. Because the $E-1$ point records the move into the earnings day, this dip reflects how poorly the signals anticipated the market's reaction to the announcement. Since one company shifts a day's accuracy by about 3–4 percentage points, this result should be treated with caution, but it suggests that $E-1$ accounts for a significant part of the gap between the near- and far-from-earnings windows: with $E-1$ removed, near-earnings accuracy increases from 45.19% to 48.31% (86 of 178), roughly halving the gap to the far-from-earnings window (50.88%).

No indicator performed better than the composite on every day, and none performed worse on every day. Most of the composite line is above the majority of the indicator lines, but this is mostly due to the composite's higher chance level (Section 7c). No sustained, day-specific concentration of predictive power emerges across the window, further evidence against H2.

7e. Benchmark Comparison Results

Table 3. *Hit rate, null distribution, and significance results for the composite model and the naive momentum baseline.*

Analysis	Window	n	Hit rate	Null mean	Null SD	p	Significant
Agg score (composite)	E-7 to E+7	434	48.16% (209/434)	49.52%	2.27%	0.7400	No
Naive momentum	E-7 to E+7	434	50.00% (217/434)	49.89%	2.45%	0.5030	No

Note. Both methods are evaluated on the same 434 predictions with the same scoring rules (Section 5c). The naive momentum baseline predicts that each stock's next-day return will have the same direction as its previous-day return. Null distributions and p -values are as in Table 1.

Neither method beats chance. The momentum baseline achieved 50.00% (217 of 434), almost exactly its null mean of 49.89% ($p = 0.503$), while the Agg Score achieved 48.16%, slightly below its null mean of 49.52% ($p = 0.740$). Both rarely forecast neutral, so their chance levels are very close and their raw hit rates can be compared directly: the composite was 1.84 percentage points less accurate than simply following the previous day's direction.

This gap is not significant. The methods disagreed on 198 days, with the composite correct on 95 of them and the baseline correct on 103 (exact McNemar test, $p = 0.619$), so these results do not support H4. In an exploratory breakdown not pre-specified in the study design, the baseline was somewhat more accurate than the composite in the near-earnings window (52.40% vs. 45.19%) and less accurate far from earnings (47.79% vs. 50.88%), but neither of these differences was significant (McNemar $p = 0.159$ and 0.547). The four options-flow indicators, when used together as they have been combined here, did not add any predictive power beyond a simple rule based on the latest price change of the stock.

8. Limitations

The limitations that follow are arranged in the order that they are expected to affect the conclusions.

8a. Direction Versus Magnitude

The key restriction is a mismatch between the hedging mechanism's prediction and the actual test conducted in this study. Dealer hedging evidence is more robust on the size of the price move than on the direction of the price move (Section 3f; Ni, Pearson, Poteshman, & White, 2021), and around earnings, options can price the probability of a big price change but not the direction of the earnings surprise. Since only next-day direction is assessed, the indicators may carry information about the size of upcoming moves that a directional hit rate cannot detect.

8b. Dealer-Short Assumption and Positioning Data

The model assumes dealers are net short, which is probably the case for retail-heavy names more than institutionally heavy ones. The assumption cannot be checked because there is no dealer-side data, and for the same reason the Option Volume Ratio only approximates the directional separation of volume described by Easley, O'Hara, and Srinivas (1998), while Delta Difference cannot measure positioning at all (Section 3b). If the assumption is incorrect, the

directional interpretation would be invalid, and the overall picture could be misleading when there are different dynamics across individual tickers.

8c. Monthly-Only Contracts and Contract Rolls

A significant portion of options trading activity, especially retail, is concentrated in short-term weekly contracts around earnings (Bryzgalova, Pavlova, & Sikorskaya, 2023), so indicators calculated from monthly contracts may miss much of the hedging flow the model aims to measure. Additionally, a ticker's window may cover a contract roll, where open interest and volume reset and time to expiration jumps by around a month, shifting the indicators independently of hedging pressure.

8d. Other Limitations

Statistical power. The study can only detect a moderately large effect, but the results are still informative: the 95% confidence interval for the composite's pooled hit rate (43.5% to 52.9%) rules out a directional edge of more than about 3 to 4 percentage points above chance. Day-positions have at most 30 observations, so day-by-day results are descriptive, and the nine related tests were not adjusted for multiple comparisons, which does not affect the conclusions since none of them was significant.

Small baseline and in-sample weights. The mean, standard deviation, and weight for each indicator are based on only 9 or 10 baseline days per ticker, so the standard deviations are noisy and the weights may contain noise; the weights are also calculated on the same days used to build the tiers.

Regime mismatch. Tiers calibrated within a more stable baseline (E-17 to E-8) are applied directly to the more volatile evaluation window, so tier assignments may reflect rising volatility rather than hedging pressure. All 30 earnings events are also in a single market environment (April 13 to May 29, 2026).

Fixed neutral threshold. The 1% Neutral Zero Rule favors low-volatility stocks when the model predicts neutral, and it reduces the chance level of methods that frequently predict neutral (Section 7c).

Sample structure. Each ticker has one earnings event, and its up to 15 consecutive days are not likely to be fully independent. The pooled permutation test also treats returns as interchangeable across companies with different volatility levels; it was used rather than a within-company shuffle because a within-company shuffle would reduce each null distribution to only 14 or 15 observations.

Data collection. Data were manually recorded from a single platform at one snapshot a day, 15 minutes before the close; one data recording error was corrected (Section 4a), and others may remain.

9. Conclusion & Future Work

The present study examined the ability of options-flow indicators to forecast short-term stock price movements, with these indicators serving as proxies for dealer hedging pressure, around earnings announcements. Each indicator was standardized against the ticker's pre-earnings baseline, converted into a seven-level score, and combined into a composite weighted by each indicator's accuracy, the Agg Score, which was evaluated over 434 predictions and 30 earnings events.

The composite did not beat chance, its accuracy was not concentrated around the release, it did not outperform its individual measures, nor did it outperform a naive momentum baseline. The findings are nevertheless informative: the confidence interval excludes a

directional edge of more than approximately 3–4 percentage points over chance, meaning that any directional information in options flow around earnings is small (Section 8d).

The results have practical implications for the groups mentioned in the Introduction. The options-flow measures, as constructed here, which retail traders can observe in public, are not suited to use as tools to trade the direction of a stock around its earnings; they don't fare better than a coin flip or a rule which follows the previous day's trading direction. For risk desks and dealers, the results indicate that the hedging activity implied by these measures does not push prices predictably in one direction, but rather tends to be more evident in the magnitude of price movements, which is the area of interest of risk management. For researchers, the results point to a mismatch between what the dealer-hedging mechanism predicts, the size of moves, and what directional tests can detect (Section 3f).

There are a few extensions that could test this more directly:

- *Predicting the size of moves.* The most direct next step is to determine whether the same indicators forecast the size of the next day's move, either the absolute return or the realized volatility. This is what the hedging mechanism predicts (Section 3f), and it can be tested using the data already collected.
- *Increased size and scope of samples.* Using the same tickers across multiple earnings cycles and market regimes would increase statistical power, enable tests that take repeated observations of a ticker into account, and show whether the results hold beyond April–May 2026.
- *Weekly contracts and fixed contract windows.* A more complete portrayal of the relevant hedging flow, and the elimination of contract-roll artifacts, would be achieved by adding weekly contracts, fixing the contract for each window, or standardizing the indicators by time to expiration.
- *Improved construction and calibration.* The main issues identified in Sections 3 and 8 could be addressed by weighting Delta Difference by open interest (Section 3b), testing a change-based Open Interest Ratio (Fodor, Krieger, & Doran, 2011), estimating weights from a separate or pooled calibration sample, recalibrating tiers for earnings-period volatility, and using a ticker-relative neutral threshold.
- *Directly measured dealer positioning.* Signed volume data indicating the direction of trades, whether buys or sells and whether opening or closing, like the Cboe Open-Close data used by Pan and Poteshman (2006), would allow the dealer-short assumption to be tested directly, including whether accuracy varies based on whether a ticker is more or less likely to be dealer-short, and would allow put buying to be separated from put selling, which Easley, O'Hara, and Srinivas (1998) demonstrate is necessary for option volume to be informative.
- *Independent replication.* Applying the unchanged model to an independent sample of earnings events would determine whether the null result is specific to this sample and period or reflects a genuine lack of signal.

Acknowledgments

The author thanks Ben B. for mentorship and guidance throughout this project.

References

- Anderegg, B., Ulmann, F., & Sornette, D. (2022). The impact of option hedging on the spot market volatility. *Journal of International Money and Finance*, 124, Article 102627. <https://doi.org/10.1016/j.jimonfin.2022.102627>
- Bryzgalova, S., Pavlova, A., & Sikorskaya, T. (2023). Retail trading in options and the rise of the big three wholesalers. *The Journal of Finance*, 78(6), 3465–3514. <https://doi.org/10.1111/jofi.13285>
- Charles Schwab & Co. (n.d.). *Studies library*. Thinkorswim Learning Center. Retrieved September 30, 2026, from <https://toslc.thinkorswim.com/center/reference/Tech-Indicators/studies-library>
- Easley, D., O'Hara, M., & Srinivas, P. S. (1998). Option volume and stock prices: Evidence on where informed traders trade. *The Journal of Finance*, 53(2), 431–465. <https://doi.org/10.1111/0022-1082.194060>
- Fodor, A., Krieger, K., & Doran, J. S. (2011). Do option open-interest changes foreshadow future equity returns? *Financial Markets and Portfolio Management*, 25(3), 265–280. <https://doi.org/10.1007/s11408-011-0164-z>
- Ni, S. X., Pearson, N. D., & Poteshman, A. M. (2005). Stock price clustering on option expiration dates. *Journal of Financial Economics*, 78(1), 49–87. <https://doi.org/10.1016/j.jfineco.2004.08.005>
- Ni, S. X., Pearson, N. D., Poteshman, A. M., & White, J. (2021). Does option trading have a pervasive impact on underlying stock prices? *The Review of Financial Studies*, 34(4), 1952–1986. <https://doi.org/10.1093/rfs/hhaa082>
- Pan, J., & Poteshman, A. M. (2006). The information in option volume for future stock prices. *The Review of Financial Studies*, 19(3), 871–908. <https://doi.org/10.1093/rfs/hhj024>
- Xing, Y., Zhang, X., & Zhao, R. (2010). What does the individual option volatility smirk tell us about future equity returns? *Journal of Financial and Quantitative Analysis*, 45(3), 641–662. <https://doi.org/10.1017/S0022109010000220>

Appendix A: Variable Definitions

The four underlying metrics are defined below, following the Thinkorswim platform's documentation (Charles Schwab & Co., n.d.).

- Delta: the expected change in an option's price for a \$1 change in the price of the underlying stock. Call deltas range from 0 to 1 and put deltas from -1 to 0; delta is also commonly used as an approximation of the probability that an option will expire in the money.
- Implied volatility: the volatility of the underlying stock implied by an option's market price, expressed as an annualized percentage. It reflects the market's expectation of future price movement.
- Open interest: the total number of option contracts at a given strike and expiration that are outstanding, meaning they have been opened and not yet closed, exercised, or expired. It is updated once per day.
- Volume: the number of option contracts at a given strike and expiration traded during the current session.

Appendix B: Data Parsing Code

The following Python functions convert raw values recorded from Thinkorswim into numbers. `parse_signed_metric` handles fields that may use trailing-sign notation (Volatility Smirk and Delta Difference); `parse_plain_number` handles all other numeric fields, removing thousands separators and percentage symbols. Values that cannot be read are returned as missing (NaN).

```
import re
import numpy as np
def parse_signed_metric(raw):
    if raw is None:
        return np.nan
    s = str(raw).strip()
    if s == "" or s.lower() in ("nan", "none", "null", "-"):
        return np.nan
    s = s.replace("%", "").replace("$", "").strip()
    m = re.match(r"^[+-]?[\d.]+\s*([+-]?)$", s)
    if not m:
        try:
            return float(s)
        except ValueError:
            return np.nan
    number_str, sign = m.groups()
    try:
        val = float(number_str)
    except ValueError:
        return np.nan
    if sign == "-":
        return -abs(val)
    elif sign == "+":
        return abs(val)
    return val
def parse_plain_number(raw):
    if raw is None or str(raw).strip() in ("", "nan", "None", "-"):
        return np.nan
    s = str(raw).replace(",", "").replace("%", "").strip()
    try:
        return float(s)
    except ValueError:
        return np.nan
```

Appendix C: Sample Composition

Table C1. Sample Companies by Sector and Size, with Fiscal Quarter and Earnings Timing

Sector	Ticker	Size Rank	Fiscal Quarter	Release Date	Timing	Day E
Financials	JPM	5	Q1 FY2026	Apr 14, 2026	BMO	Apr 14, 2026
	V	4	Q2 FY2026	Apr 28, 2026	AMC	Apr 29, 2026
	GS	3	Q1 FY2026	Apr 13, 2026	BMO	Apr 13, 2026
	SOFI	2	Q1 FY2026	Apr 29, 2026	BMO	Apr 29, 2026
	UPST	1	Q1 FY2026	May 5, 2026	AMC	May 6, 2026
Technology	NVDA	5	Q1 FY2027	May 20, 2026	AMC	May 21, 2026
	QCOM	4	Q2 FY2026	Apr 29, 2026	AMC	Apr 30, 2026
	KLAC	3	Q3 FY2026	Apr 29, 2026	AMC	Apr 30, 2026
	U	2	Q1 FY2026	May 7, 2026	BMO	May 7, 2026
	BBAI	1	Q1 FY2026	May 5, 2026	AMC	May 6, 2026
Consumer	AMZN	5	Q1 FY2026	Apr 29, 2026	AMC	Apr 30, 2026
	COST	4	Q3 FY2026	May 28, 2026	AMC	May 29, 2026
	GM	3	Q1 FY2026	Apr 28, 2026	BMO	Apr 28, 2026
	RIVN	2	Q1 FY2026	Apr 30, 2026	AMC	May 1, 2026
	SG	1	Q1 FY2026	May 7, 2026	AMC	May 8, 2026
Energy	XOM	5	Q1 FY2026	May 1, 2026	BMO	May 1, 2026
	CVX	4	Q1 FY2026	May 1, 2026	BMO	May 1, 2026
	OXY	3	Q1 FY2026	May 5, 2026	AMC	May 6, 2026
	DVN	2	Q1 FY2026	May 5, 2026	AMC	May 6, 2026
	RUN	1	Q1 FY2026	May 6, 2026	AMC	May 7, 2026
Healthcare	LLY	5	Q1 FY2026	Apr 30, 2026	BMO	Apr 30, 2026
	ABBV	4	Q1 FY2026	Apr 29, 2026	BMO	Apr 29, 2026
	PFE	3	Q1 FY2026	May 5, 2026	BMO	May 5, 2026
	A	2	Q2 FY2026	May 27, 2026	AMC	May 28, 2026
	GRAL	1	Q1 FY2026	May 5, 2026	AMC	May 6, 2026
Communications	META	5	Q1 FY2026	Apr 29, 2026	AMC	Apr 30, 2026
	NFLX	4	Q1 FY2026	Apr 16, 2026	AMC	Apr 17, 2026
	DIS	3	Q2 FY2026	May 6, 2026	BMO	May 6, 2026
	SNAP	2	Q1 FY2026	May 6, 2026	AMC	May 7, 2026
	FUBO	1	Q2 FY2026	May 6, 2026	BMO	May 6, 2026

Note. Size rank orders the five companies within each sector by market capitalization (Yahoo Finance, April 20, 2026), from 5 (largest) to 1 (smallest). Sectors follow Yahoo Finance classifications, with Consumer Cyclical and Consumer Defensive combined, and Sunrun (RUN) placed in Energy (Section 4d). Fiscal quarters follow each company's own fiscal calendar; NVIDIA's fiscal year 2027 began in February 2026. BMO = released before market open; AMC = released after market close. Day E is the first full trading session in which earnings information could be reflected in price: the release date for BMO releases and the following trading day for AMC releases (Section 4b). NFLX's Day E coincided with a monthly options-expiration date (April 17, 2026), so its Day E indicator values were excluded (Section 4c).



Appendix D: Permutation-Test Reproducibility

Table D1. *Range of Permutation-Test p -Values Across 20 Repetitions*

Test	Reported p	Lowest p	Highest p
Agg score, pooled	0.740	0.718	0.767
Agg score, far from earnings	0.436	0.426	0.494
Agg score, near earnings	0.871	0.858	0.895
Volatility smirk	0.771	0.769	0.817
Delta difference	0.996	0.989	0.998
Open interest ratio	0.604	0.594	0.631
Option volume ratio	0.471	0.433	0.507
Naive momentum	0.503	0.471	0.522

Note. Each permutation test (1,000 shuffles of next-day returns across all companies) was repeated 20 times without a fixed random seed. Reported p is the value given in Section 7. No test produced a p -value below 0.05 in any repetition.