Distinguishing a Firm Attribute from a Moment in Event-Prediction Studies: A Placebo-in-Time Design and a First Application

Maksim Aleksandrovich Andreev, Joseph Howden
Quark Quantitative Research

Abstract

Empirical studies that predict corporate-control events almost always compare firms that were acquired against control firms that were not, and report the discrimination a set of pre-event features achieves between the two groups. That comparison answers a composite question. A feature can separate the groups either because it marks the kind of company that eventually attracts a bidder, or because it marks the months during which a particular company is being approached. The two readings carry opposite operational content, and a case-control design cannot tell them apart, because the case firm and the control firm differ in identity and in moment simultaneously. We describe an arm that separates them. The same case firms are re-scored at moments deliberately chosen to contain no deal in either direction, screened by the control arm's own admissibility rules; firm identity is held fixed and only the observation moment moves. Under the hypothesis that a feature set measures a precursor, a case moment and the same firm's quiet moment are perfectly separable in principle and the placebo score collapses toward the control score. Under the hypothesis that the feature set measures a permanent firm attribute, the two are indistinguishable and the discrimination is 0.5 by construction. We report a first application on 2,853 acquisition episodes and 8,523 activity-matched control windows drawn from United States regulatory filings, validated under combinatorial purged cross-validation with a 180-day embargo. Scoring the case firms five, seven and nine years from their anchors recovers 56% of the case-versus-control gap: the mean score is 0.3034 at deal-free moments against 0.3779 at real case moments and 0.2087 at real control windows. Discrimination between a case moment and the same firm's quiet moment is 0.600, against 0.758 between a case and a different firm. Feature by feature, between 89% and 140% of the gap on total filing count, Form 4 count, 8-K count, filing-gap variance and item 5.02 is present with no deal within five to nine years in either direction. The only moment-linked columns are the strategic-review flag, which is the single-rung baseline the joint model was required to beat, and the soliciting-material count, which the unit's own leakage scrub removes. The unit settles measurement-limited on its calibration bar while its discrimination bar passes. Two properties of the frozen design generalise beyond this application. The benchmark rung named in the preregistration has discrimination 0.5011 on this panel, so the frozen success criterion was never testable as written. The base rate of 0.2508 is an artifact of one-to-three matching, so a calibrated model here would have been calibrated to a fictitious prevalence. All measurements reported are in-sample and none is forward-proven.

1. Introduction

A researcher who wants to predict which listed companies will be acquired has an obvious design available. Assemble the firms that were acquired, assemble a comparison group that was not, measure a vector of features on both at some fixed horizon before the announcement, and report how well the features separate the groups. The design is inexpensive and it runs on data that regulators publish, so it is the design most often reached for.

That number answers a question with two components bound together. Consider a feature that separates the two groups well, say a firm's rate of material-event disclosure in the year before the anchor date. Two readings are available. The feature may be marking the months during which a bidder is circling, in which case a high value carries information about timing and a desk can act on it. Alternatively the feature may be marking a durable property of the company, its size, its sector, its governance history or its disclosure culture, in which case a high value says that the firm belongs to the population from which targets are eventually drawn and says nothing about when. The

first reading would support a position taken on timing, whereas the second supports a watchlist and nothing beyond one.

A case-control comparison cannot distinguish these readings, and the reason is structural rather than statistical. The case observation and the control observation differ in two respects at once: they are different firms, and one of them sits adjacent to a deal. Any separation the features achieve is a sum of a firm-identity contribution and a moment contribution, and the design provides no second equation with which to solve for the split. No increase in sample size resolves it, because the confound is in the contrast rather than in the noise. Purged cross-validation, embargoes, permutation controls and leakage tripwires all address the question of whether the measured separation is real, and all of them are silent on what the separation is made of.

This paper describes an arm that supplies the missing equation. The construction is to score the same case firms at moments chosen so that no deal is present in either direction, and to compare the resulting scores with the scores those firms receive at their real case moments. Firm identity is held fixed by construction; only the moment moves. Everything durable about the firm, whether it is measured, mismeasured or unobserved, appears on both sides of the comparison and cancels. What remains is attributable to the moment.

> **Contribution.** A validity gate for event-prediction studies, with a null that is 0.5 by construction rather than by estimation, and a first application in which a majority of a case-control gap survives at deal-free moments. The application reports a null with an identified mechanism and claims no new edge, and it carries two further findings about the frozen design that generalise to any study of this shape.

2. The identification problem stated

Write $s(i, t)$ for the score a fitted model assigns to firm i observed at date t. A case-control study evaluates the contrast between $s(i, t_i)$ at case firms observed shortly before their announcement anchors and $s(j, t_j)$ at control firms observed at matched dates. Decompose the score additively into a firm component and a moment component, setting any interaction between the two aside:

```
s(i, t)  =  a(i)  +  m(i, t)  +  noise

a(i)      durable firm attribute: sector, size,
complexity, disclosure
          culture, governance history, age
m(i, t)   moment component: whatever is
different about firm i during
          the months adjacent to a corporate-
control event
```

The case-control contrast estimates E[a(case) + m(case, anchor)] minus E[a(control) + m(control, matched date)]. Matching reduces the first difference and cannot eliminate it, since matching operates on the variables the designer thought to match on. The contrast is therefore a sum of two unknowns, and a study that reports it as evidence of a precursor has assumed the first term away rather than measured it.

The placebo-in-time contrast differs in the one respect that resolves the decomposition. Score the case firms at dates far from any deal and compare within firm:

$$\begin{aligned}\Delta_{\text{placebo}}(i) &= s(i, t_{\text{case}}) - s(i, t_{\text{quiet}}) \\ &= [\, a(i) + m(i, t_{\text{case}}) \,] - [\, a(i) + m(i, t_{\text{quiet}}) \,] \\ &= m(i, t_{\text{case}}) - m(i, t_{\text{quiet}})\end{aligned}$$

The firm term $a(i)$ enters both halves of the difference and cancels identically. It does so whether or not it is observed, whether or not it has been matched on, and whether or not the designer is aware that it exists, which is the property that separates this contrast from covariate adjustment. Matching removes the confounders an analyst has thought to name; differencing a firm against itself removes every time-invariant firm term at once, including the unnamed ones.

The null hypothesis attached to this contrast is unusually clean. If a feature set carries no moment information, then a case moment and the same firm's quiet moment are exchangeable and a classifier separating them achieves 0.5. That value is a property

of the construction rather than an estimate from a control sample, which is what distinguishes this arm from a permutation control. A permutation test establishes that some association between features and labels exists; it says nothing about whether that association is carried by identity or by timing, because shuffling labels destroys both components together.

> **Definition (placebo-in-time arm).** For each case firm i with anchor t_i, draw one or more placebo dates t displaced from t_i by an interval long enough that no deal process plausibly spans it, and admit each placebo observation only if it satisfies the control arm's own admissibility rules. Score those observations with the model fitted on the case-control panel. Report the share of the case-versus-control gap recovered at placebo moments, and the discrimination between a case moment and the same firm's placebo moment.

Two implementation requirements deserve emphasis, because they are the places where the construction can silently fail. The placebo observation must pass the control arm's admissibility screen, since otherwise the comparison reintroduces a composition difference through the back door. And the displacement must be checked against calendar drift, because a feature whose distribution shifts over the sample period will produce a placebo-versus-case gap that reflects the era rather than the moment.

3. The application and its data

The design was first run inside a registered research unit on pre-deal signatures, which asked whether the pre-deal signs taken together separate future acquisition targets from ordinary companies better than the best single measured sign does, and whether the resulting score is calibrated as well as ranked. Its protocol was frozen on 12 August 2026, before any joint statistic existed, and records a prior probability of 0.30 that the primary bar would pass.

Cases are deal episodes carrying an 8-K-anchored announcement date. The case builder produced 3,056 episodes across 1,737 filer identifiers with anchors spanning 16 March 1994 to 7 July 2026; 2,853 survive into the modelling panel, with 201 dropped for insufficient filing history and two for pre-1995 anchors. Feature vectors are cut 180 days before the anchor. A point-in-time tripwire in the case builder dropped 2,628,932 records dated after each row's cut across 48,055 gate calls.

Controls are activity-matched rather than randomly sampled, and the matching clause is the direct inheritance of an earlier failure in the same programme. An insider-trading precursor unit had drawn controls from the population of all registrants, which includes funds, trusts and shells; only 31% of its control windows carried any insider filing against 79% of its case windows, and every filing-count feature inherited that confound. The clause frozen here requires a control to have filed an annual report and at least one 8-K within the prior twelve months, and to carry no deal-family filing in the window running from 360 days before the anchor to 180 days after it. Each control is additionally matched to its case on calendar quarter and on filing-volume decile, and three controls are drawn per case.

The clause worked on the variable it names. Insider-filing presence runs 88.0% for cases against 81.7% for controls, in place of the 79% versus 31% split that ended the earlier unit. The standardised mean difference on total prior-year filings, which is the matching variable itself, is −0.037, so the confound that had compromised the predecessor is removed.

Two imbalances survive the match, and both are reported here because they bear directly on the identification problem this paper addresses. Firm age at the cut is 13.05 years for cases against 7.90 years for controls, a standardised mean difference of +0.710, larger than that of any feature in the model. Industry is unmatched entirely; among rows with a resolved industry classification, cases over-weight chemicals and pharmaceuticals by 11.7 points, depository institutions by 7.7, holding and investment offices by 7.5 and business services by 6.9. Both of these are firm attributes in the sense of Section 2, which is precisely the term the case-control contrast cannot isolate.

The resulting panel has 11,376 rows, 2,853 cases and 8,523 controls, with anchors from 19 August 1995 to 10 August 2026 and a base rate of 0.2508 fixed by the matching ratio. Match status is full for 2,834 cases,

partial for 13 and unmatched for 6. The control pool holds 1,993 registrant identifiers, of which 1,153 carry 8-K item codes and are used, with a median reuse of 8 and a maximum of 11.

> **A contamination caught before any statistic.** The strategic-review outcome ledger used to populate one feature carries two cohorts, and its 2,500 rows labelled as control are an earlier study's seeded null, a random registrant paired with a random date at which no review occurred. Of the 1,993 identifiers in this unit's control pool, 672 appear in that seeded cohort against 397 with genuine detections. Indexing the ledger without filtering on cohort would have handed a third of the control arm a fabricated review event and inflated control prevalence from the correct 5.42% to a contaminated 7.91%. The builder caught this before any statistic was computed and two selftests lock it.

4. Model, validation architecture and frozen bars

The primary model is a penalised logistic regression on eighteen columns, chosen for interpretable coefficients because the question concerns which signs carry the separation rather than the maximisation of discrimination. Gradient boosting is reported alongside as a secondary. Validation is combinatorial purged cross-validation with a 180-day embargo, in the sense of López de Prado [2]: the panel is split into six calendar groups, two groups are held out at a time giving fifteen paths, and any training row whose information window overlaps a test fold's window is purged. The purge discards 53,908 rows across the fifteen paths against a mean of 3,990 surviving training rows per path, which is the expected geometry for a 730-day information window on a six-group split.

Three of the features named in the preregistration could not be populated and are reported as not computable; none of the three is imputed to zero. The routine-filtered insider purchase requires document-level parsing of more than a hundred thousand filings and was not run; point-in-time fundamentals were unavailable; and peer repricing density is computable on only 35 rows, because the available price panel begins in July 2025 while the anchors begin in 1994.

Four bars were frozen with the design, and their disposition is the substance of Sections 5 through 9.

BAR	FROZEN THRESHOLD	COMPUTED	RESULT
1. Brier increment over the single-rung baseline	≥ 0.010	+0.01112 as delivered; +0.00411 on the clean stratum	passes as delivered; not robust
1. Discrimination increment over the same baseline	≥ 0.05	+0.1635 as delivered; +0.1186 minimum across sensitivities	pass
2. Calibration of the top decile	realised rate inside the predicted 95% interval	predicted 0.6658 [0.6380, 0.6933], realised 0.6028	fail
3. Permutation control	joint skill above the 95th percentile of 200 within-quarter shuffles	observed +0.1333 against a 95th percentile of −0.0018	pass
4. Leakage tripwire (blocking)	scrubbed model loses ≤ 0.020 Brier skill	drop of +0.0077	does not fire; statistics reportable

The blocking bar deserves a note, since it governs whether anything in this paper may be reported at all. The tripwire refits the model with three columns scrubbed, namely the soliciting-material count and the two elapsed-time columns attached to the measured rungs, and requires that the scrubbed model not collapse. It loses 0.0077 of Brier skill against a bar of 0.020, so no feature is encoding the answer deterministically and the statistics are reportable rather than sealed. An obvious structural-leak hypothesis was tested separately and refuted: a zero 8-K count is more common among controls at 7.19% than among cases at 3.26%, and that indicator alone discriminates at 0.4803.

5. Discrimination as delivered

The joint model achieves an out-of-fold Brier score of 0.16284 against a climatology of 0.18789, a Brier skill score of 0.13334, and discrimination of 0.73115. The gradient-boosting secondary reaches 0.15960, 0.15054 and 0.74174 on the same folds. The operative single-rung baseline, a one-feature model on the strategic-review flag, scores 0.17396, 0.07415 and 0.5676. The delivered increments are therefore +0.011121 in Brier against a frozen bar of 0.010, and +0.1635 in discrimination against a frozen bar of 0.05.

The permutation control is clean. Two hundred within-quarter label shuffles give a 95th-percentile joint skill of −0.00177 against the observed +0.13334, and a 95th-percentile increment of −0.000194 against the observed +0.011121. The arithmetic was reproduced digit for digit by an independent scorer using its own Brier and rank-statistic implementations, and a brute-force date check confirmed that no training row overlaps any test fold's information window. Two verification passes recomputed 125,136 feature values from the raw caches using code that imported neither builder and found zero mismatches.

On the strength of these numbers alone the study would report that a combined pre-deal signature separates future targets from matched controls, that the separation survives purged cross-validation and a permutation control, and that no leak is detectable. Sections 6 and 7 give the reasons such a report would mislead.

5.1 Comparison against the best single measured sign

The preregistration's operating question was whether the joint signature beats the best single measured sign, and at the selection size a desk operates at it does not. Comparing the fitted score against the strategic-review flag on the full panel, the rule selects 997 names at a precision of 0.6409 with a Wilson interval of [0.6107, 0.6701]. The model's top 997 achieve 0.6349 [0.6046, 0.6642], which is 0.60 percentage points worse, and its top decile of 1,138 achieves 0.6028 [0.5741, 0.6308], which is 3.81 points worse. Restricting to anchors from 2010 the model recovers, reaching 0.6520 [0.6219, 0.6809] at matched selection size and 0.6667 [0.6357, 0.6963] in the top decile, gains of 1.10 and 2.57 points with intervals that overlap the rule's almost entirely.

Where the rule and the model disagree the model is genuinely better, hitting 0.364 on the 77 names it selects alone against 0.221 on the 77 the rule selects alone. The +0.164 discrimination gain is real and it lives in the middle of the distribution, ranking the roughly ninety percent of names on which the single rungs are silent. That is a capability the rungs do not have, and it differs from a higher-precision top of book, which is the region of the distribution where sizing decisions are taken.

A diagnostic worth recording sits alongside this. Centring every feature on its matched-set mean, which asks whether the separation is a within-set contrast or a between-set composition effect, raises discrimination to 0.78249 and skill to 0.21453. The figure is diagnostic only and cannot be operated, because no matched set exists at the moment a desk scores a live name. It indicates that a meaningful part of the achieved separation is a comparison against the firm's own matched peers rather than an absolute property of the score.

6. The placebo result

The same case firms were scored at deal-free moments displaced five, seven and nine years from their anchors, admitted only where they satisfied the control arm's admissibility rules. The mean score at those placebo moments is 0.3034, against 0.3779 at the real case moments and 0.2087 at the real control windows.

```
share of the gap attributable to firm identity

  = ( placebo mean  -  control mean ) / ( case
mean  -  control mean )
  = ( 0.3034       -  0.2087       ) / ( 0.3779
-  0.2087       )
  = 56%

remainder attributable to the moment: 44%
```

Discrimination between a real case moment and the same firm's deal-free moment is 0.600. Discrimination between a case and a different firm, computed on the same placebo arm as its cross-firm comparator, is 0.758. Both figures are computed over the case firms with admissible placebo moments; the settlement record does not state the pair count over which they were evaluated, a gap the successor unit closes. The two numbers answer different questions, and their difference is the quantity the case-control design cannot recover on its own. Year-matched strata agree with the pooled figure, which rules out the calendar-drift failure mode identified in Section 2.

> **Reading the two discrimination figures.** The 0.758 cross-firm value is the placebo arm's own comparator and is not the same quantity as the 0.73115 out-of-fold discrimination reported in Section 5, which is computed across all fifteen purged cross-validation paths on the full panel. They are reported separately here for that reason. The comparison that carries the finding is 0.600 against 0.758 within the placebo arm, where both figures share a construction.

A model that discriminates a case from a different firm at 0.758 and a case moment from the same firm's quiet moment at 0.600 is measuring firm identity for the majority of its performance. The interval between 0.600 and 0.5 is where any genuine precursor content lives, and it is narrow.

7. Which features carry the moment, and which carry the firm

The decomposition can be run feature by feature, comparing each feature's mean at real case moments, at the same firms' placebo moments, and at real control windows. The share present without a deal is the placebo-minus-control difference over the case-minus-control difference.

FEATURE	CASE, REAL	CASE, PLACEBO	CONTROL, REAL	SHARE PRESENT WITH NO DEAL
Filing-gap variance	197.39	185.36	227.32	140%
8-K item 2.06 (impairment)	0.031	0.035	0.014	126%
8-K item 5.02 (officer or director change)	0.822	0.834	0.719	112%
Total filing count	72.951	72.637	77.023	108%
Insider filing count	37.404	37.811	31.607	107%
8-K count	13.066	12.706	9.665	89%
8-K item 4.02 (non-reliance)	0.030	0.033	0.037	48%
Strategic-review flag	0.290	0.100	0.054	19%
Soliciting-material count	0.857	0.578	0.517	18%

Shares above 100% arise where the placebo mean sits slightly beyond the case mean in the direction of separation, which is within sampling variation at these counts and carries the same interpretation as a share of 100%: the entire gap is present with no deal in either direction.

The five features at the top of the table are the substance of the fitted model. Its three largest standardised coefficients are total filing count at −1.477, an odds ratio of 0.228 per standard deviation, insider filing count at +1.030 or 2.802 per standard deviation, and 8-K count at +1.024 or 2.785 per standard deviation, each with sign stability of 1.00

across all fifteen paths. Read at face value the model says that a future target files about as much in total as its matched peer but shifts the composition of that volume toward material-event and insider filings. The placebo decomposition refutes the precursor reading of that sentence: 108%, 107% and 89% of those three gaps respectively are present at moments with no deal within five to nine years in either direction. The composition shift describes the kind of company that is eventually acquired rather than the months in which it is approached.

The two columns that behave as genuine moment markers are the strategic-review flag at 19% and the soliciting-material count at 18%, and both are unusable for the purpose the study was built to serve. The strategic-review flag is the single-rung baseline the joint model was required to beat; recovering it is the preregistration's own stated kill condition, which reads that a model merely recovering the announcement of a review has added nothing. The soliciting-material count is one of the three columns the unit's leakage scrub removes. The model's fourth and fifth largest coefficients are exactly these two, at +0.372 and +0.248.

The insider-filing result also settles a question the predecessor unit left open. The gap in insider filing counts between arms had been a candidate signal, and 107% of it survives with no deal in either direction, so it reflects activity level and firm type rather than anticipation. The second-largest coefficient in the model sits on a raw count rather than on the routine-filtered purchase measure the preregistration names, and that measure has never been computed on this panel; dropping the raw count takes the Brier increment to +0.00797, below the frozen bar.

8. Calibration and the verdict

The frozen preregistration states that a discriminating but miscalibrated model settles measurement-limited, on the reasoning that the desk sizes positions on probability floors and an uncalibrated score is unusable however well it ranks. The calibration bar asks whether the realised rate in the top decile of predicted probability falls inside that decile's predicted 95% interval.

The realised rate falls outside that interval. The top decile holds 1,138 rows, predicts a mean of 0.6658 with an interval of [0.6380, 0.6933], and realises 0.6028 with a Wilson interval of [0.5741, 0.6308], 686 events. The reverse reading fails as well, since the predicted mean does not fall inside the realised interval. Four of ten deciles fall inside their predicted intervals under the delivered binomial construction, four under an exact Poisson-binomial construction, and five under matched-set clustering.

The mechanism is era mixing rather than a top-decile artifact or a general inability to state probabilities. Out-of-fold skill by era runs as follows.

```
era          n      climatology   joint
skill
1995-2000    173    0.18965       0.23391
-0.2334
2000-2005    863    0.19219       0.22123
-0.1511
2005-2010   1072    0.18750       0.19306
-0.0297
2010-2015   2692    0.18750       0.16337
+0.1287
2015-2020   3180    0.18750       0.15141
+0.1925
2020-2026   3396    0.18750       0.14513
+0.2260
```

Skill is negative on every era before 2010 and positive on every era after. Restricting to anchors on or after 2010 moves the top decile to a predicted 0.6954 against a realised 0.6663, which sits on the interval boundary. The statistics verifier read this as inside at [0.6656, 0.7249], while an independent recomputation at a decile size of 926 put the lower bound at 0.6663 against a realised 0.6663. A recovery that two implementations place on opposite sides of the boundary names a revival condition and does not establish calibration. The bar is frozen on the pooled panel in any case, and the failure stands there.

The binding bar here is calibration, for a reason specific to how the score would have been used. Discrimination licenses a ranking, and a ranking is

enough to order a research queue. Sizing requires a probability, and the probability this model states is wrong by six points at the only part of the distribution a desk would act on. The frozen consequence map has three branches, covering both bars passing, the primary failing while calibration passes, and the tripwire firing. None of the three covers the outcome that occurred, in which the primary passes as delivered while calibration fails, and the ranking-aid clause is conditioned on calibration passing. The unit therefore defaulted to the most conservative available reading and wired nothing: no desk gate, no probability, no ranking aid, no threshold change and no re-based card.

> **Generalisable instrument lesson.** A frozen consequence map must cover the full cross-product of its bars rather than the outcomes its author anticipated. This map was written by someone who expected discrimination to be the binding constraint, and the outcome that occurred had no entry.

9. Two properties of the frozen design that generalise

Independently of the placebo result, two features of this design would have compromised its conclusions had the bars gone the other way, and both are properties that any study of this shape can check before it runs.

9.1 The benchmark rung was not testable on the panel

The preregistration names the benchmark explicitly: the joint signature must beat the best single measured sign, which is a disclosed-proposal rung measured at 65.9% conversion within twenty-four months on 123 scored episodes, Wilson interval [57.1%, 73.6%], against 6.5% on 1,347 seeded control windows [3]. On the panel built for this unit that rung has discrimination of 0.5011 and a Brier skill of −0.0006, with a prevalence of 0.93% among cases and 0.26% among controls. A one-feature model on a column that is present in fewer than one case in a hundred cannot separate anything, so the frozen success criterion could not be evaluated as written.

The substitution actually made was to the broader strategic-review flag, measured elsewhere at 31.7% conversion within eighteen months on 6,273 episodes against a 3.4% control base [3], which has a prevalence of 0.2903 among cases and 0.0542 among controls on this panel and therefore supports a fitted baseline. The substitution was declared before any statistic, and it lowers the bar the joint model must clear. Its consequence for the frozen criterion was not declared, and the criterion as frozen was never tested.

> **The check this implies.** A benchmark named in a preregistration must be evaluated for computability on the panel the study will actually build, before the bars are frozen. A benchmark whose prevalence is below roughly one percent in the case arm cannot serve as a fitted comparator whatever its measured conversion rate elsewhere, because the two quantities are computed on different populations.

9.2 The base rate is a construction artifact

The panel's base rate of 0.2508 follows from drawing three controls per case and has no relationship to the prevalence of acquisition in any real population. Every probability the model states is therefore expressed on a fictitious prevalence scale. The top decile's realised 0.6028 represents a lift of 2.40 times the design base, which is a meaningful quantity; the 0.6028 itself is not commensurable with the 31.7% of the review rung or the 65.9% of the proposal rung, both of which are conversion rates measured on populations.

The consequence for the frozen bars is sharper than a caveat about interpretation. The frozen consequence clause for a passing calibration bar requires the model's probability and its floor to enter a candidate class alongside the measured rungs. That operation is not executable without a prevalence correction, which this unit does not estimate. A perfectly calibrated model on this panel would have been perfectly calibrated to a prevalence that does not exist, so bar 2 could not have licensed its own frozen consequence even had it passed, and the bar was therefore incapable of discharging its stated role under either outcome.

10. A control-admissibility defect, and where the delivered increment lived

The delivered Brier increment of +0.011121 clears its frozen bar of 0.010 by eleven percent. An adversarial re-reading of the panel located the source of that margin in an asymmetry between how cases and controls were admitted.

The case builder drops any case whose first-ever filing postdates the start of the feature window, so every case is measured over a full 365-day window. Control admissibility imposes no equivalent rule and refuses only a filer that had not yet filed anything at all by the anchor date. The consequence, independently reproduced during settlement, is that 1,095 of 8,523 control rows, 12.8%, are measured over a truncated window. Their mean 8-K count is 4.70 against 10.40 for full-window controls and 13.07 for cases, so a truncated control is an artificially quiet control and the model separates it easily for a reason that has nothing to do with acquisition. Of the 2,853 matched sets, 854 contain at least one such control.

Splitting the panel on that stratum locates the entire increment inside it.

```
stratum                                  n
baseline   joint      increment
matched sets WITH a truncated control  3416
0.16704    0.13957    +0.02747
matched sets WITHOUT one               7960
0.17693    0.17283    +0.00411

identity check:
  (3416 x 0.02747 + 7960 x 0.00411) / 11376  =
0.01113
  delivered increment                      =
0.01112
```

Refitting on the clean stratum rather than restricting to it gives +0.00367. Either reading fails the bar of 0.010. The verifier's t-statistic on the clean stratum against the bar is −5.78, and against zero is +4.03, so the increment there is real and small rather than absent.

Three further readings point the same way. Under the unit's own leakage scrub the increment is +0.00967, below the bar, so bars 1 and 4 are jointly unsatisfiable as frozen and the run never computed this. Exactly one specification of the eighteen columns clears the bar, and four single columns each individually sink it when dropped: 8-K count leaves +0.00129, total filing count +0.00619, insider filing count +0.00797, soliciting-material count +0.00952. Dropping all rung columns leaves +0.00225, and a strict scrub leaves +0.00515. Under every clustering the increment is not distinguishable from the bar it must clear, with a bootstrap 95% interval over 2,853 matched sets of [0.00947, 0.01283] that straddles 0.010.

```
clustering             clusters   SE        t
vs zero   t vs bar
matched set            2853       0.000868
12.82       +1.29
calendar quarter        122       0.001626
6.84       +0.69
calendar year            32       0.002548
4.37       +0.44
calendar group            6       0.006083
1.83       +0.18
```

The increment is unambiguously non-zero, and its magnitude is not distinguishable from 0.010. The six-cluster read is a degenerate estimate and over-states the pessimism; the defensible middle is the 32-cluster calendar-year figure at 4.37 against zero.

The register records the outcome of the primary bar as a pass, because the programme honours frozen bars mechanically and the delivered arithmetic clears the threshold. The registered Brier contribution against a frozen probability of 0.30 is therefore 0.49. The defect-corrected reading, under which the primary fails, would score 0.09; taking it would let a discovery made after the outcome flatter the forecast ledger, which is the behaviour preregistration exists to prevent. Both readings are recorded, and the disclosure forms part of the settlement itself.

> **Generalisable instrument lesson.** A matched-control panel must apply the case builder's admissibility rules to controls. Any rule that governs which cases are measurable becomes a composition difference the moment it is not mirrored, and a composition difference in a matched design is indistinguishable from signal at the level of the fitted score.

11. Scope and limitations

Everything reported here is an in-sample measurement on a panel assembled in August 2026, and no quantity in this paper is forward-proven. The programme that produced it operates a forward ledger for exactly this reason. That ledger stands at 136 resolutions on record against a quorum of thirty, which is therefore met. A provenance caveat attaches to the count: the archive writer had discarded entry spreads, the entry records were recovered by a backfill from the desk's snapshot history, and 68 earlier resolutions predate that history and cannot be recovered. The measured base rates remain unadopted pending settlement of the registered hypothesis that governs them. A measured in-sample contrast and a forward-validated edge are different objects, and the distinction is load-bearing throughout the series [5].

Two further caveats bind the benchmark quantities this paper compares against, and both are stated here rather than deferred. The conversion rates quoted for the measured rungs are established against a delisting proxy for deal completion, so completion figures such as the disclosed-proposal rung's 61.0% are lower bounds: the proxy is tight for cash deals, where the target's listing genuinely ends, and loose for stock-for-stock transactions whose surviving entity keeps a listing [3]. Separately, the programme's convention for panels assembled from current index membership is that contrasts are citable and levels are not, because membership today is not membership at the time of the observation. The same reading discipline applies to this panel for a different reason: its levels are expressed on a base rate of 0.2508 that one-to-three matching created, so the contrasts reported here, the 56% placebo share, the 0.600 against 0.758 discrimination comparison and the per-feature shares, are the findings, and the levels are not population quantities.

Within those limits the placebo arm licenses a specific and useful conclusion. The feature set assembled for this unit measures a firm attribute for the majority of its performance, and the study would have reported a precursor without the arm. That conclusion does not generalise to all feature sets or all event classes; it is a measurement on this panel with these columns. What generalises is the design and the threshold it establishes for this class of study.

> **The standing gate.** A candidate signature must discriminate a case moment from the *same firm's* deal-free moment. On this panel the reference figure for that comparison is the same-firm value of 0.600; a study that reports only the cross-firm figure, here 0.758, has reported a quantity that a permanent firm attribute can produce on its own.

Several claims that bear on the result remain unresolved, and each is named here. Whether the primary bar clears under a symmetric window rule is unknown, since the clean-stratum figures were computed by restriction and by refit on a reduced panel rather than by rebuilding the panel under the corrected rule. Whether calibration recovers after 2010 is unknown, for the knife-edge reason given in Section 8. Whether the insider channel contributes anything was unknown at the time of writing, because the routine-filtered purchase measure had not been computed here. A successor unit has since computed it under the same-firm design this paper advocates, and the channel discriminates a case moment from the same firm's quiet moment at an area under the curve of 0.6141, above the 0.600 same-firm reference this paper sets. The direction is inverted from the accumulation premise: every buying feature runs backwards, and the discriminating quantity is the absence of a non-routine purchase over the trailing year rather than any property of purchase timing. The feature is right-censored at that same one-year lookback, so it carries no information about a firm's own purchase rhythm, and the finding should not be paraphrased as one about cadence. Whether the signature survives matching on firm age and industry is unknown, and firm age carries the largest imbalance in the design at a standardised mean difference of +0.710. What the score means as a probability is unknown without a prevalence correction this unit does not estimate.

The revival conditions attached to the unit follow from those gaps directly: apply the case builder's first-filing rule to control admissibility and rebuild; carry the placebo arm as a standing validity gate in any future

signature unit; and match on firm age or restrict controls to an age band. Three further conditions concern the bars themselves: restate the primary bar against a defined model, since as frozen it is read on a model containing a column the unit's own leak definition removes; estimate a prevalence correction before any score is stated as a probability; and either compute the routine-filtered insider measure or drop the insider channel, in place of substituting a raw count for it.

11.1 Relation to the wider literature and to the rest of the series

The multiple-testing discipline this programme applies follows Harvey, Liu and Zhu [1], whose finding that conventional significance thresholds produce large numbers of false discoveries in the cross-section motivates the practice of freezing bars before any statistic is computed. The validation architecture follows López de Prado [2] on purged and combinatorial cross-validation, which addresses the leakage that arises when overlapping information windows place related observations on both sides of a split.

Neither corrective addresses the identification problem described in Section 2, and this is the point the paper is built around. A study can freeze its bars, purge its folds, embargo its windows, run a permutation control and pass a leakage tripwire, and still report a firm attribute as a precursor, because every one of those instruments asks whether the measured separation is real. The placebo arm asks a different question, which is what the separation is made of, and it is cheap to add: the same fitted model, the same feature code, the same admissibility screen, evaluated at dates the study already knows how to construct.

Within the series, the measured rungs used as benchmarks here are established in the companion measurement paper on pre-announcement and in-process deal probabilities [3], and the data limits that shaped which features were computable are catalogued in the companion paper on the reachable boundary of free filing data [4]. The distinction between an in-sample measurement and a forward-validated claim is treated at length in [5].

12. Reproducibility

The pipeline runs in three stages, each guarded by a selftest that must pass before the stage writes anything. The first stage builds the case panel. It reads the cached index of regulatory filings, identifies the deal episodes that carry an 8-K-anchored announcement date, and emits one feature row per episode cut 180 days before its anchor. Features for both arms are computed by a single function that is blind to the arm a row belongs to, and the point-in-time gate that discards any record dated after a row's cut operates inside this stage. Fourteen checks guard it, and it produced the 3,056 case rows reported in Section 3.

The second stage assembles the matched panel. It draws three activity-matched control windows for each case under the admissibility clause set out earlier, filters the strategic-review outcome ledger on cohort so that seeded null windows cannot be read as genuine detections, and joins the two arms into the 11,376-row panel the model consumes. Alongside the panel it emits a pre-modelling report carrying the prevalence table, the recorded defects and the amendments, and a reconciliation that checks the cached filing indexes against live retrieval. Thirty-two checks guard this stage, two of them locking the cohort filter, and the control-admissibility asymmetry examined in Section 10 originates in the two clauses at lines 334 and 391 to 392.

The third stage fits and evaluates. It consumes the matched panel, fits the penalised logistic primary and the gradient-boosting secondary under combinatorial purged cross-validation with a 180-day embargo, and computes the four frozen bars, the calibration table, the ablations and the permutation control. It writes those to a results record, together with a second record holding design-matrix coverage and the per-column standardised mean differences. Twenty-seven checks guard it. The random seed recorded with the delivered run is 20260812, and this stage completed in 282 seconds.

The reported figures divide cleanly across those outputs and two further documents. The frozen bars, the power block, the prior probability of 0.30 and the

consequence map discussed in Section 8 are read from the preregistration, which was frozen before any joint statistic was computed. The bars, coefficients, calibration table and ablations of Sections 4, 5, 8 and 10 are transcribed from the results record. The placebo tables of Sections 6 and 7, the stratum decomposition of Section 10 and the defect list are transcribed from the settlement record. The panel counts given in Section 3 appear in both. The benchmark rung quoted in Section 9.1 is taken from the companion ladder measurement's own results [3], and the registered verdict, including the recorded probability score of 0.49 and the disclosed alternative of 0.09, is held in this unit's register entry.

Exact file names for each of the three stages, for each record they write and for the two documents are recorded in the repository and in that register entry.

13. Conclusion

A case-control study of corporate-control events estimates the sum of a firm-identity effect and a moment effect and reports it as the second. Holding the firm fixed and moving only the observation moment separates the two, and it does so against a null of 0.5 that follows from the construction rather than from an estimate. In the first application of the design, 56% of the case-versus-control gap survives at deal-free moments five to nine years from the anchors, and discrimination against the same firm's quiet moment is 0.600 against 0.758 across firms. Between 89% and 140% of the gap on the model's principal features is present with no deal in either direction. The only columns that behave as moment markers are the baseline the model was required to beat and a column the study's own leakage definition removes.

The unit settled measurement-limited, wired nothing to the desk, and recorded the two structural findings that would have compromised its conclusions had the bars fallen the other way: a benchmark that could not be evaluated on the panel that was built, and a base rate manufactured by the matching ratio. The result is a null with an identified mechanism and no claim of an edge. Its transferable content is the arm itself, which costs one additional scoring pass over dates the study already knows how to construct, and which any study of this shape can run before it reports a precursor.

References

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

[2] M. López de Prado. *Machine Learning for Asset Managers*. Cambridge University Press, 2020.

[3] M.A. Andreev & J. Howden. A measured ladder of pre-announcement and in-process deal probabilities, reconstructed entirely from free EDGAR filings. *Quark Quantitative Research Working Paper*, 2026.

[4] M.A. Andreev & J. Howden. The reachable boundary of free filing data for event-driven research: a catalogue of measurement limits with named defect classes. *Quark Quantitative Research Working Paper*, 2026.

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

Quark Quantitative Research Platform — Working Paper
Generated 2026-08-14 22:39 UTC

This paper proposes a placebo-in-time arm as a standing validity gate for event-prediction studies and reports a first application in which 56% of a case–control gap survives at deal-free moments. The result is a null with an identified mechanism; no predictive edge is claimed, and no measurement reported here is forward-proven. Correspondence: M.A. Andreev, Quark Quantitative Research.