The Reachable Boundary of Free Filing Data for Event-Driven Research: A Catalogue of Measurement Limits with Named Defect Classes

Maksim Aleksandrovich Andreev, Joseph Howden
Quark Quantitative Research

Abstract

Event-driven research on United States corporate filings is widely described as accessible, on the grounds that the filings themselves are free and machine-readable. We report the boundary at which that accessibility stops. Over a single preregistered programme built on the Electronic Data Gathering, Analysis and Retrieval system and on free daily price series, eight units settled at the verdict class `settled as measurement-limited`, meaning that a limit of the measurement prevented the frozen question from being answered as posed. Whether a unit's statistics were additionally sealed is a consequence fixed per unit by its preregistration, and several of the units below settled with reportable statistics. We catalogue all eight, together with four adjacent units that settled null or failed their own validity bars, and we group the failures into six named defect classes: extraction of transaction terms from filing text, which limited or killed four separate units; delisting and price coverage, where a provably-real delisted target is exactly the name a free price source cannot serve; control-pool composition, where a seeded panel of registrants is populated with funds and shells so that any filing-count feature inherits a confound; entity-level versus episode-level scoping, which killed two units; keying on filer identity for documents the index carries under both bidder and target; and selftests that cannot fail, of which we give two instances and the positive-control discipline that replaced them. For each wall we state what breaching it would require, separating what a paid subscription buys from what only forward accrual buys. Every figure below is in-sample and none is forward-proven; the programme's forward ledger for deal completion holds 136 resolutions, recovered by a backfill from the desk's snapshot history, so its thirty-resolution quorum is met while the measured base rates remain unadopted pending settlement of the registered spread hypothesis. Completion rates rest on a delisting proxy and are lower bounds, tight for cash consideration and loose for stock-for-stock survivors. The quality panel is screened on current index membership, so its contrasts are citable and its levels are not.

1. Introduction

The published record of quantitative research is selected on success. A method that produced a statistic gets a paper; a method that ran into a data limitation before producing one usually gets nothing, or a single sentence in a limitations paragraph. The consequence for a replicator is that the accessible literature describes the reachable region of a data source without describing its boundary, and the boundary has to be rediscovered by each new group at full cost. For filing-based event research the cost is substantial, because the discovery typically happens after the harvest, the parser and the join have all been built.

This paper reports the boundary for one programme. The programme measures pre-announcement and in-process corporate-event probabilities from the Electronic Data Gathering, Analysis and Retrieval system (hereafter the filing system), supplemented by free daily price series. Its units are preregistered: a design document is frozen before any statistic exists, naming the bars the unit must clear and the validity gates that fire before those bars are read. A unit settles at the verdict class `settled as measurement-limited` when a limit of the measurement prevents the frozen question from being answered as posed, whether that limit is detected by a validity gate ahead of the bars or established at the bars themselves. Sealing is a separate consequence applied per unit under the preregistration's own language: a sealed statistic is computed, archived and explicitly not read as a finding, and while some units sealed every outcome statistic, others settled with statistics that remained reportable. Eight units have settled at this class.

We treat those eight settlements, plus four adjacent units that settled null or failed their own validity bars, as the primary data of this paper. Our object is the

defect that fired the gate rather than the hypothesis the unit was testing, and we group defects into classes because two of the classes killed more than one unit before they were named.

> **Contribution.** We publish a catalogue of measurement limits for free filing data, organised into six named defect classes, with the gate that detected each limit, the numeric threshold it failed, and the archived results file in which the figure is recorded. For each class we separate the remedy that a paid subscription supplies from the remedy that only forward accrual supplies. The catalogue is intended to be used before a harvest rather than after one.

2. The measurement programme, its gates, and the catalogue

Each unit begins with a preregistration frozen before any statistic exists. The document fixes the population, the outcome definition, the direction of any effect, the bar the primary statistic must clear, and the consequence that follows from each outcome. The programme applies the Harvey, Liu and Zhu (2016) argument about multiple testing as its default significance discipline, so directional claims are held to a threshold well above the conventional two-standard-error rule, and several of the units discussed below carry permutation controls alongside the analytic bar.

The mechanism that produces most of the settlements catalogued here sits ahead of the statistical bar. Validity gates test whether the instrument is capable of the measurement at all, and they are written to fire before the outcome is observed. A blind hand-validation gate draws a fixed sample, usually twenty records, and requires an analyst reading the underlying document to agree with the extraction at a stated rate. A population gate requires the detected event count to fall inside the range the design expected, on the reasoning that a count an order of magnitude below expectation indicates a detection failure rather than a rare phenomenon. A prevalence gate requires the control arm to exhibit the background rate the design assumed. A coverage gate requires the fraction of the population that can be priced or scored to exceed a floor, and where selection is plausible it additionally requires the coverage differential between outcome arms to stay below a ceiling.

> **Verdict class.** `settled as measurement-limited` records that a limit of the measurement prevented the frozen question from being answered as posed, and that the hypothesis is neither confirmed nor refuted. The limit is usually detected by a validity gate ahead of the statistical bars, though the precursor signature study reached its bars and failed a calibration requirement its design had defined in advance as a measurement limit. Sealing is a separate consequence applied per unit: the insider precursor study sealed its statistics under its own kill language, the desk replay sealed its primary, and the precursor signature study's statistics remained reportable because its leakage tripwire held. The class is distinct from `settled as null`, which records that the statistic was read and failed its bar, and from `settled as failed-validity`, which records an estimator that failed its own construction bars on real data.

A sealed or non-citable number appears below only where the register records it, and it is identified as such at the point of use. The insider precursor study's case prevalence is quoted in section 5 under that marking, and the spread study's single trough observation is repeated in section 3 as an illustration only. The odd-lot tender study's per-event profit figures and the desk replay's sealed Brier differences are held in their archived results files and are not reproduced in this paper. None of these figures supports a claim about returns.

The programme also scores its own forecasts. Each preregistration states a prior probability that the unit will settle positive, and settlement converts that prior into a Brier score, so that a sequence of confident designs that fail leaves a visible cost in the record. Those priors appear in the table below alongside the gate that fired and the defect class developed in sections 3 through 8. Every stated prior sits between 0.30 and 0.45, so each design was priced as more likely to fail than to succeed, and the settlements largely confirmed guarded forecasts rather than overturning confident ones.

UNIT	GATE THAT FIRED	NUMERIC OUTCOME	DEFECT CLASS	PRIOR → BRIER
Odd-lot self-tender, first pass	Post-hoc instrument audit	n = 48; 21 of 48 events carried implausible spreads above 30%; 8 of 48 from a single quarterly repurchaser	Terms extraction; price adjustment	—
Odd-lot self-tender, second pass	Blind extraction gate	18 of 77 filtered events (23.4%) outside the plausibility band against a bar below 10%	Terms extraction	0.40 → 0.16
Candidate payoff legs	Band-drop and coverage gates	Leg A: 12 in band against a floor of 15, band drop 29.4%. Leg B: 245 creditable of 400 against a floor of 300	Delisting and price; entity scoping; filer keying	—
Insider precursor	Control prevalence gate	control prevalence 6.75% against a designed floor of 8%	Control-pool composition	0.35 → 0.1225
Spread-implied outcome	Blind extraction gate; population floor	8 of 20 blind samples were junk tickers; 83 eligible against a design range of 120 to 250, with 13 failures in total	Terms extraction; delisting and price	0.30 → 0.09
Post-termination drift	Population gate	35 terminations against a design expectation of 150 to 350; 2 cash and 33 stock against 90 to 210 per arm	Entity versus episode scoping; terms extraction	0.40 → 0.16
Desk replay	Reproduction gate	one of six reproducible cards differed by 30.2 points against a 2-point ceiling; replay coverage 16.1% against a 60% gate	Point-in-time inputs; delisting and price; unfailable selftest	0.45 → 0.2025
Precursor signature	Calibration bar	top decile predicted 0.6658 against a realised 0.6028	Miscalibration; control admissibility	0.30 → 0.49; a defect-corrected reading scoring 0.09 is disclosed and not taken (section 9.1)

Four further units bear on the same boundary without carrying that verdict class. The strategic-review language study settled null after four outcome-blind repairs to its classifier. The quality time-conditioning study settled null at a statistic of 2.776 against a bar of 3.0. The filing-typed cause study settled null with its effect running opposite to the frozen direction. The hierarchical-floor estimator failed its own validity bars on real data. Each is discussed where its defect class arises.

3. The terms-extraction wall

Transaction terms live in filing prose. A tender offer states its price per share in a cover page and again in the offer to purchase; a merger states consideration in a definitive proxy statement, often alongside superseded figures from a background-of-the-transaction narrative. Extracting a price from that prose is the most load-bearing operation in filing-based event research, because almost every downstream quantity is a ratio against it. Four units here were limited or killed by the extraction step, and the register's verdicts on the step differ across them: the odd-lot pass's blind gate accepted the sampled extractions while the population-level plausibility test that followed found extraction quality inadequate to certify; the spread study's blind gate failed outright; the post-termination study found an extracted price for three of thirty-five episodes, so its accuracy was never tested; and the replay's deal lane combined the same

junk-ticker defect with a coverage shortfall. Each unit failed the joint requirement of accuracy and coverage across the population its design needed.

The second odd-lot self-tender pass extracted 106 offers. Its blind validation gate drew twenty of them and compared the extracted price against the filing by hand: the median absolute percentage error was zero and at least seventeen of twenty were exact, with two recorded suspects, one a set of warrant terms and one an offer denominated in Canadian dollars. The gate the unit failed was the population-level plausibility test that followed. Eighteen of seventy-seven text- and issuer-filtered events, 23.4%, produced an extracted offer outside the band from 0.90 to 2.50 times the pre-filing close, against a frozen ceiling below 10%. The statistical bars were never read. Two diagnostics were recorded for any successor design: forty-one of the seventy-seven events had no spread at all at the following close, entry exceeding 99% of the offer, so a forward instrument in this class must price on the same day; and sixteen of the seventeen scored events resolved as completed through amendment parsing, so the exit machinery worked.

The spread-implied outcome study asked whether the entry spread on an announced deal predicts completion. Its blind gate returned eight of twenty samples that were not prices at all but fragments of cover-page text captured by the ticker bridge, the literal tokens `symbo`, `code` and `of` among them; thirty-three of eighty-three eligible rows carried such tokens. The same reading found a wrong-quantity class, including a deal price of 2.75 recorded against a stock trading near 29 and a note-tender price attached to common shares. The unit was independently terminal on population: eighty-three eligible episodes against a design range of 120 to 250, and thirteen failures in total, so the failure arm could not reach the twenty-per-bucket citable bar at any price coverage. The mechanism behind that shortfall is the extraction wall itself. Terms extraction concentrates on completed deals, because a completed deal files a definitive proxy statement containing the consideration, and a deal that dies often never does.

The post-termination drift study inherited the same bottleneck at one remove. Its design classified terminated deals as cash or stock, following Malmendier, Opp and Saidi (2016), and its frozen cash test required an extracted per-share price. Of thirty-five detected terminations, three had one. The cash arm therefore held two events against a design expectation of 90 to 210, and the population gate fired before any return was computed and before any price was fetched.

The fourth instance is the deal lane of the desk replay, discussed in section 4, where the same ticker-bridge tokens reappear and coverage lands at 16.1% against a 60% gate. Together the four units support a specific statement of the wall. Free filing text supports accurate extraction of a transaction price for the subset of episodes that file a definitive terms document, and that subset is selected on completion, so any design whose failure arm requires an extracted price is measuring the wrong population before it starts.

> **Numbers in this section are in-sample and none is forward-proven.** The odd-lot profit figures in the archived results file are sealed by the gate and are not cited here or anywhere else. The +32.8% spread observed at the March 2020 trough on a deal that subsequently closed is recorded in the spread study as a single non-citable observation and is repeated here only as an illustration of the class the forward ledger is expected to measure.

4. The delisting-and-price wall

A merger that completes removes its target from the exchange, which is at once the event the research wants to measure and the event that terminates the free price series. The takeover-premium leg of the candidate payoff study reached this wall by a route worth recording in full, because each mechanical repair to the join moved the failure to a deeper layer until the remaining failure was the data source itself.

The leg's purpose was to replace a prior of positive 30% on the announcement-premium payoff with a measurement. Its extraction gate passed cleanly, with a median absolute percentage error of zero, seventeen of twenty exact, and the extractor correctly preferring

final terms over the superseded figures in background narratives. The first join keyed on filer identity and dropped 78.6% of matched rows out of the plausibility band. The second, after rekeying to the subject company, dropped 60%. That iteration exposed two further defects. An announcement date taken as the first filing in the deal family lags the press release by weeks, so the supposedly unaffected close was already post-announcement, and delisting notices of the Form 25 note class gave live acquirers false closed episodes. The third iteration required an event-report announcement anchor, a filer that stops filing, and a price series that dies. Those three requirements are the correct definition of a provably-real completed target, and they left sixty-one candidates, seventeen joined episodes and twelve in band against a floor of fifteen, with a band-drop rate of 29.4%. Fourteen candidates were excluded because the price series survived, thirty because no usable price existed. The free bulk source returned no delisted series at all; the delisted-capable source was missing twenty-nine of sixty-one windows.

> **Statement of the constraint.** The requirements that establish a target really completed its transaction are the same requirements that guarantee the free price sources cannot serve it. A mechanically correct join therefore converts a coverage problem into a smaller and cleaner coverage problem, and never into an adequate sample.

The second leg of the same study, measuring the forward drift of companies that announce a strategic review and never convert, failed a coverage gate at 245 creditable episodes of a 400-episode sample against a frozen floor of 300 and an override of 270. Its 155 unpriced residual is survivorship-skewed by construction, so running the leg thin would have biased the loss side of the payoff optimistic. The frozen consequence was that the priors of positive 30% and negative 35% stand in the expected-value proxy with their settled status stated on the card, and that measured legs must arrive from the platform's own forward ledger rather than from reconstruction.

The deal lane of the desk replay reached the same wall from a different direction and measured it most cleanly. Of 273 eligible episodes, forty-four were priceable at the decision horizon, an overall coverage of 16.1% against a 60% gate. Coverage on completed deals was 14.0% and on failed deals 77.8%, a differential of 63.8 percentage points against a ceiling of 25. The unpriceable residual decomposes into 194 episodes with no resolvable ticker, twenty-nine with a junk ticker and six with no cached window. Ignoring the differential clause would have produced a spread-conditional completion curve estimated overwhelmingly on deals that failed, which is the selection the gate exists to catch.

5. Control-pool composition

The insider precursor study asked whether non-routine open-market purchases by officers and directors precede deal announcements, using the routine-filer classification of Cohen, Malloy and Pomorski (2012) applied at the filing level. Its harvest was sound: 5,595 ownership-change filings cached, a parser selftested against planted purchases, planted sales and planted routine filers, and six of six selftests passing. Four hundred deal episodes were matched against four hundred seeded registrant-quarter controls.

The prevalence gate fired, with control non-routine purchase prevalence at 6.75% against a designed floor of 8%, and the statistics were sealed; the case prevalence of 10.25% is recorded in the archived results file and is explicitly non-citable under the preregistration's own kill language. The shortfall has a straightforward diagnostic explanation. Thirty-one per cent of control windows contained any ownership-change filing at all, against seventy-nine per cent of case windows. The seeded control pool draws from all registrants, and that population contains funds, shells and inactive issuers in quantity. A contrast built on it therefore measures the separation between operating companies and entities that barely file, rather than the separation between companies that are and are not about to be acquired.

The class generalises to any feature that counts filings, which is most of what free filing data offers. The named revival was activity-matched controls, and the programme executed it. The precursor signature study, which pooled the pre-deal signals into a joint model,

drew 8,523 controls matched on filing activity for 2,853 cases, and its matching variable behaved as intended: ownership-change filing presence ran 88.0% in cases against 81.7% in controls, with a standardised mean difference on the matching variable of negative 0.037, against the unmatched study's 79% versus 31%. The confound was removed. What the matched study then found is treated in section 9, and it does not vindicate the original hypothesis.

6. Entity-level versus episode-level scoping

A company has a filing history; a transaction has a filing episode. A classifier written against the first when the design intends the second will mislabel every company that has done more than one deal, and the error is silent because both scopes return plausible-looking labels.

The post-termination drift study carried the frozen rule that a terminated deal is a stock deal if the episode contains a business-combination prospectus of the Form 425 class. Implemented against each company's entire filing family, the rule labelled any company that had ever run a stock deal as a stock deal forever. Forty-two of forty-three detected events were mislabelled, with named counter-examples: one company's terminated 2008 transaction was a cash leveraged buyout and another's was a tender offer, both carried as stock. The correction was recorded as a pre-outcome amendment rather than applied silently after the fact. Scoping the rule to the episode window moved the split from one stock label in forty-two events to two cash and thirty-three stock, which repaired the labels without populating the cash arm, since that arm was independently blocked by the extraction wall of section 4.

The premium leg described in section 4 was the first instance of the same class, in the form of a join keyed to the filer rather than to the episode's subject. The register records the recurrence explicitly, which is the reason the class is named here rather than left as two separate incidents. The operational rule that follows is to define the episode first, as a window with a start and an end, and to evaluate every classifier and every join inside it, treating any rule that consults a company's history outside the window as a design decision requiring its own justification.

7. Bidder-versus-target keying

The filing index carries tender-offer documents under both the bidder and the subject company. A pipeline that keys on the filer therefore acquires the acquirer's identity for a document about the target, and every price it subsequently fetches belongs to the wrong company. In the premium leg this produced a band-drop rate of 78.6% at the first join iteration, a rate large enough for any plausibility test to catch, and the plausibility test caught it.

The dangerous form of the same defect is the one the desk replay found, where resolution from subject identity to a trading symbol silently returns the acquirer or a symbol belonging to an unrelated company with a colliding identifier, and the archived results file names two large-capitalisation symbols returned this way. Inspection of the ticker cannot detect it, because the returned ticker is a real ticker for a real company. Only a price-ratio test against the extracted deal terms catches it, and that test is available exactly on the 16% of episodes that are priceable at all. Of the forty-four testable rows in that lane, eight were wrong-quantity wide and sixteen fell outside the premium band.

> **Detection rule.** Symbol resolution from a filing identifier must be validated against a quantity the filing itself asserts rather than against the plausibility of the symbol. Where no such quantity exists in the document, the resolution is unvalidated and should be treated as an uncontrolled defect channel rather than as a minor data-quality issue.

8. Selftests that cannot fail

Every unit in this programme runs a selftest suite against synthetic fixtures before it touches real data, on the principle that an instrument should be shown to detect a planted effect and to refuse a planted null. Two selftests in the post-termination drift study passed while being incapable of failing, and the pair is worth

recording precisely because the suite around them was otherwise sound and the unit reported five of five passing after repair.

The first was a permutation control whose fixture assigned every synthetic observation within a year the same label, so permuting labels within years permuted nothing and the control reproduced the observed statistic exactly. A permutation test whose null distribution is a point mass at the observed value reports agreement under any implementation, correct or otherwise. The second was a leakage tripwire comparing post-window exposure against a threshold, where both the measured exposure and the reference were zero for structural reasons unrelated to whether the function under test leaked, and comparing two zeros produces a pass carrying no information about the property being tested.

Both were rebuilt as positive controls, meaning fixtures on which the test is required to fire. The repaired leakage tripwire moves from 0.21 to 1.20 when run against a deliberately leaky variant of the function while remaining invariant on the correct one, so the tripwire is demonstrated to detect the thing it exists to detect. The discipline generalises to a requirement that every validity test in a suite be exercised in both directions, and that a test which has never been observed to fail be treated as unverified rather than as passing.

The desk replay contributes a structural variant of the same pathology. Its elapse-timing bar compared a conditional probability against a realised rate, and the two quantities were computed from the same two counts, so the bar is an algebraic identity that could not fail on any data: the maximum absolute difference over fifty randomly drawn cohorts is 0.0 percentage points. The same lane additionally replayed a code path that production does not execute, so the quantity it computed was not the quantity the desk uses.

A third instance in the programme is a stage-transition study whose engine compared a string against a date in its censoring branch; the selftest missed it because every planted episode resolved, so the censoring branch was never entered. The register records this as the third instance of the same lesson, which is that a suite must exercise every branch and not merely the branch the fixture author expected to matter.

9. Calibration and prior-contamination limits

Two units failed for reasons no data purchase would remedy, and belong in the catalogue because a replicator will otherwise attribute their failure to the data source.

9.1 Discrimination without calibration

The precursor signature study pooled the pre-deal filing signals into a penalised logistic model with a gradient-boosted secondary, evaluated by combinatorial purged cross-validation with a 180-day embargo on 2,853 cases and 8,523 activity-matched controls, 11,376 rows in total, fifteen paths, with 53,908 rows purged. Its leakage tripwire did not fire, at a skill drop of 0.0077 against a bar of 0.020, so its statistics were reportable rather than sealed. The primary bar passed as delivered, at a Brier increment of 0.01112 against a bar of 0.010 and an area-under-curve increment of 0.1635 against 0.05, and the permutation control passed at an observed joint skill of 0.1333 against a permuted 95th percentile of negative 0.0018.

The calibration bar failed, with the top decile predicting 0.6658 on a 95% interval of 0.6380 to 0.6933 against a realised 0.6028 on a Wilson interval of 0.5741 to 0.6308. The preregistration states in advance that a model which discriminates but is miscalibrated settles measurement-limited, because the operational consumer of the score sizes positions from a probability floor and an uncalibrated score is unusable however well it ranks.

Two further facts are recorded in the settlement. The primary was not robust: control admissibility never inherited the case builder's first-filing rule, so 1,095 of 8,523 control rows, 12.8%, were measured over a truncated window, and 854 of 2,853 matched sets contain at least one. The increment inside that stratum is 0.02747 on 3,416 rows against 0.00411 outside it on 7,960 rows, with 0.00367 on refit, and the clean-stratum statistic against the bar is negative 5.78, so the

delivered increment is arithmetically the defect. A placebo arm scoring the same case firms at deal-free moments five, seven and nine years from their anchors recovers 56% of the case-control gap, and the area under the curve separating a real case moment from the same firm's quiet moment is 0.600 against 0.758 separating a case from a different firm. Feature by feature, between 89% and 140% of the gap on total filing count, ownership-change filing count, event-report count and the officer-departure item is present with no deal in either direction. What the pooled signals mostly encode is a permanent attribute of the firm rather than a state of the firm before a transaction.

The settlement also records two structural defects in the frozen design. The benchmark rung the model had to beat, the disclosed-proposal class of section 10, has an area under the curve of 0.5011 on this panel, so the frozen success criterion was never testable as written. The panel's base rate of 0.2508 is in turn an artifact of one-to-three matching, so a perfectly calibrated model here would be calibrated to a fictitious prevalence. The consequence map had three branches and none covered the outcome that occurred, so the most conservative reading was adopted and nothing was wired. The forecast ledger scores the frozen prior of 0.30 against the mechanical passage of the primary bar, for a Brier contribution of 0.49; the register discloses that a defect-corrected reading would score 0.09 and declines to take it, on the recorded ground that adopting a reading discovered after the outcome would let a post-hoc finding flatter the very forecast ledger that preregistration exists to protect.

9.2 A prior that is itself contaminated

The hierarchical floor unit proposed to replace per-name Wilson lower bounds, computed on five to thirteen episodes each, with a Beta-Binomial posterior shrinking toward a class panel of 3,298 episodes. It failed its own validity bars, and two pre-outcome amendments were required before the real run. The first repaired a frozen bar that demanded the zero-concentration limit of the posterior reproduce the Wilson bound, which conflates a score interval with a Bayesian percentile. The second exposed a genuine hazard: the zero-concentration limit of a plain Beta-Binomial is the Haldane posterior, whose fifth percentile at five successes in seven trials is 0.418, above both the Clopper-Pearson bound of 0.341 and the Wilson bound of 0.359. The better-motivated prior would have delivered a looser floor exactly at the sample sizes where these cards live.

On the real panel the estimator still failed, with seven of seventy-three production-eligible names violating the first bar and the conservatism guard failing. The guard's failure mode was recorded at the time as confined to two or fewer episodes, and that reading was wrong. The successor unit established by algebra that at a perfect record the Jeffreys interval is looser than the score interval by construction, so the guard fails at every such record: three of three by 11.7 points and seven of seven by 12.2 points. Seven of seven is the exact shape of the name the line of work exists to price. The original reading came from inspecting the first twenty-five rows of a list sorted on floor lift rather than on sample size, so a claim about small samples was drawn from a sample selected on a different quantity. The decomposition of the floor lift is worth recording independently of that verdict, since it identifies where the lift came from. The median floor lift of 17.5 percentage points splits into 4.4 points attributable to the change of interval convention and 13.8 points attributable to shrinkage toward a class rate that is itself estimated on a current-membership panel and therefore survivorship-inflated. The extreme case is a single name with one episode, a loss, whose raw rate of 0.00 would have carried a floor of 0.4457. Adopting the estimator would have manufactured confidence on precisely the thinnest records, and the frozen consequence of no partial adoption was applied.

10. What free filing data does reach

The same programme, on the same free filings, produced measurements that cleared their bars, and which ones cleared is itself informative: the reachable measurements are those whose outcome is a filing event rather than a price.

The stage ladder measured completion probability conditional on the deal stage inferred from typed filing sequences, on 3,056 episodes of which 2,970 were decided and 86 censored. Completion runs 50.3% from the start of a deal family, 72.9% after a preliminary merger proxy, 77.6% after a definitive merger proxy, 50.8% on the tender path and 76.3% after a target's recommendation statement, each level read through the delisting proxy of section 11 and therefore a lower bound, tight for cash consideration and loose for stock-for-stock survivors. A rider to the ladder gives a sharper conditional. Among episodes reaching the definitive-proxy or tender stage, those still unresolved past the closers' 75th-percentile duration of 113 days complete at 34.8% against 68.6% for reachers overall, on 847 survivors, at a two-proportion statistic of 16.35.

The break-predictor study measured which first-sixty-day filing patterns precede failure, on 2,625 episodes alive at day sixty with an overall not-completed rate of 56.19%. Completion is read through the same delisting proxy, so the levels quoted here inherit the proxy's bias against stock-for-stock survivors, and the contrasts between arms carry the finding. Two features cleared in their frozen direction: a proxy-path deal with no definitive merger proxy by day sixty fails at 62.2% against 45.6%, on 1,671 exposed episodes, at a statistic of 8.27; an unresolved tender at day sixty fails at 73.2% against 53.6%, on 343 exposed, at 6.80. Both exceed the permuted 95th percentile of 3.77. The study's largest effect ran opposite its frozen direction and was reported without being wired: a high count of additional soliciting material is associated with failure at 33.1% against 65.0%, at a statistic of negative 14.73, which reads as deal machinery operating rather than as vote trouble.

The proposal ladder measured the probability that a company disclosing receipt of an acquisition proposal enters a definitive process within twenty-four months. On 123 scored episodes with 81 conversions the rate is 65.85%, Wilson interval 57.11% to 73.64%, against 6.46% on 1,347 seeded control windows and against the strategic-review rung's 31.7%, at a statistic of 8.14. The conversion outcome is itself a filing event; confirmed completion rests on the delisting proxy, and the delisting-confirmed rate of 60.98% reported in section 11 is the corresponding lower bound. Time to the first deal-family filing runs 49, 114 and 228 days at the quartiles. Its instrument record belongs with the defect classes above. The first blind read scored ten of twenty on target identity, because the harvest was catching bidder-filed proposals, counterparty disclosures, fiduciary-out boilerplate and asset-only transactions. A document-level filter restricting the harvest to proposals the filer itself received took acceptance to nineteen of twenty. The same reading exposed background-narrative disclosures quoting a proposal at or after the definitive agreement, and the pre-definitive exclusion removed 53 of 176 accepted episodes, every one of which would have auto-scored as a conversion.

Two contrasts on the quality panel also cleared, on 3,298 episodes across 499 names; the panel is screened on current index membership, so the return levels quoted next are survivorship-inflated and only the contrasts, in which both arms carry the same bias, are citable. Entries in a constructive technical stage returned 9.31% over the 63-trading-day horizon against 5.36% for declining-stage entries, a gap of 3.95 percentage points at a Newey-West statistic of 5.175 with a permuted 95th percentile of 0.33 points. Entries within 190 days of the prior peak returned 10.13% against 5.48% for slower drawdowns, a gap of 4.65 points at 4.985, and there the permutation margin is thin, with a 95th percentile of 4.53 points against the observed 4.65, which the settlement discloses. Both contrasts had previously settled null on a 528-episode instrument, at statistics of 2.52 and 2.776, and both certified on roughly six times the sample without any change to the method. That pair is the clearest evidence in the programme that several of its nulls record limited power rather than absence of effect.

Not everything on the larger panel certified. Typing the cause of each drawdown from adverse event reports within a ten-day window found that adverse-typed episodes returned more, 8.92% against 7.69%, a difference of negative 1.23 points against the frozen direction at a statistic of negative 1.047, on 327 typed episodes of 3,295 scored. A study of whether the wording of a strategic-review announcement

conditions its conversion rate settled flat, with banker-named announcements converting at 36.38% on 492 episodes, sale-explicit at 34.44% on 540 and vague at 32.62% on 1,821, a top-to-bottom gap of 3.76 points at a statistic of 1.57 against a minimum detectable effect of roughly 4.4 to 6.2 points.

That last unit produced a side result bearing on the reliability of the reachable region. Its blind validation found that roughly half the review corpus was not an announcement at all, holding expense line items, non-standard financial-measure boilerplate, an investor-meeting title and acquirer-side news, and an eligibility gate cut 6,273 episodes to 2,853. On the cleaned population the conversion rate is approximately 33% against 31.7% on the raw corpus, so the strategic-review result survives population cleaning rather than being an artifact of phrase-match contamination.

11. Caveats that bind every figure above

Three qualifications attach to the numbers in this paper wherever they appear, and are stated here in one place because they are properties of the results rather than footnotes to them.

Completion is measured through a delisting proxy. An episode counts as completed when the target stops being a listed filer within the observation window, which means every completion rate quoted above is a lower bound. The bound is tight where consideration is cash, because a cash-acquired target delists reliably, and loose where consideration is stock and the surviving entity continues to trade. The stage ladder, the break predictors and the proposal ladder all inherit this, and the proposal ladder additionally reports a delisting-confirmed completion rate of 60.98% alongside its 65.85% definitive-process rate, the gap between the two being a direct reading of how much the proxy costs.

Every figure in sections 2 through 10 is an in-sample historical measurement made after the fact on archived filings, and none has been validated forward. The programme's own forward ledger for deal completion holds 136 resolutions against its thirty-resolution quorum. The count carries a provenance caveat. The desk archive records 204 resolved deals, but the archive writer formerly discarded the entry spread, leaving zero usable spread-to-outcome pairs. A backfill from the desk's snapshot history recovered entry records for 136 of the 204; the remaining 68 predate the snapshot history and are unrecoverable. The quorum is met, and the operational expected-value calculation still runs on a literature prior of 91% completion with a standard deviation of 4 points rather than on the platform's measured rates, because the measured base rates are not adopted until the registered spread hypothesis settles. A forward ledger for disclosed-proposal detections exists and holds six entries. Until the registered hypotheses settle and the remaining quorums fill, the correct reading of every confirmed unit above is that a historical relationship was measured rather than a tradeable edge demonstrated.

Survivorship binds where the panel is screened on current membership. The 3,298-episode quality panel is constructed from current index constituents, so its return levels are inflated by the exclusion of names that left the index, and only the contrasts, in which both arms carry the same bias, are citable. The hierarchical-floor unit is the clearest illustration of the cost: its class prior is estimated on that panel, and 13.8 of its 17.5-point median floor lift is shrinkage toward a rate that is survivorship-inflated. Neither the stage-gate nor the time-conditioning result is an independent replication of its 528-episode predecessor, because the larger panel contains the smaller one.

12. What each wall would take to breach

The remedies divide into three kinds, and the division matters more than the individual entries because it determines whether a research programme should spend money, spend engineering effort, or wait.

WALL	REMEDY	KIND
Terms extraction	A commercial transactions database supplying consideration, form and dates per episode, removing the dependence on prose extraction and on completion-selected terms documents	Money
Delisting and price	A survivorship-free price history retaining delisted series, the single purchase that unblocks the premium leg, the fizzle leg, the spread curve and the replay's deal lane at once	Money
Bidder-versus-target keying	An identifier map from filing entity to security, validated against a quantity the filing asserts	Money, partially; engineering for the validation test
Control-pool composition	Activity-matched controls, built under the case builder's own admissibility rules	Engineering, already executed
Entity versus episode scoping	Episode-window definition applied to every classifier and join	Engineering, already executed
Unfailable selftests	Positive controls on which each validity test must fire, plus branch coverage over the fixtures	Engineering, already executed
Point-in-time card inputs	Archiving each decision's inputs at the time it is made, since fundamentals-based vetoes cannot be reconstructed from a later snapshot	Time, from the moment archiving starts
Miscalibration of a pooled score	A panel whose base rate is the operational prevalence rather than a matching artifact, plus a placebo-in-time validity gate	Engineering plus time
Forward validation of every result above	Accrual of resolutions in the platform's ledgers to their frozen quorums	Time only

The last three rows are the ones no subscription addresses. The desk replay established that production replaces a card's probability with a conservative lower bound whenever a solvency veto fires on point-in-time fundamentals, and that those fundamentals have no archived history. One of six reproducible cards therefore differed by 30.2 points against a 2-point reproduction ceiling. The gate as frozen is unsatisfiable by any offline reconstruction, since no vendor sells the state of a decision system on a past date. The remedy is to begin archiving inputs, at a cost measured in elapsed calendar time. Forward validation stands in the same position: the eight measurement limits catalogued here are limits on reconstruction, and each of the frozen consequences attached to them routes its measurement to a forward ledger for that reason.

One operational finding from the replay survives its own seal and belongs to the same family of failures. The elapse-conditional timing panel that decays flagged probabilities was an untracked production-only file, read under a bare exception handler with a silent empty-list fallback. Production was verified healthy on inspection, but any rebuild from the source repository would have reverted every flagged card to a flat 0.60 with an unreachable expiry branch. A silent fallback around a missing input has the same character at run time as a selftest that cannot fail, in that both report success in a state that carries no information about the property at issue.

13. Reproducibility

Every unit in the catalogue passes through the same seven stages, and the stages are described here in the order in which they run. Units differ in which stages they use and in where they stop, and each settlement in sections 3 through 9 records a stage that could not deliver what the stage after it required.

The harvester opens the pipeline. It consumes the filing system's daily form indexes and its full-text search endpoint and produces a local document cache keyed by company identifier, form type and filing

date, so that every later stage reads from disk and no result depends on the state of a remote service on the day it was produced. Harvest volumes range from a few hundred documents for the narrow event classes to the 5,595 ownership-change filings behind section 5.

Episode construction consumes that cache and produces the analytic unit of the programme, a window with a start and an end: an announcement anchor taken from an event report, and either a resolution date or a censoring date at the edge of the observation period. Every classifier and every join in the stages that follow is evaluated inside the window, which is the rule section 6 derives from the two units that did not observe it.

Terms extraction consumes the text of the terms documents inside an episode and produces the consideration per share and the form of consideration. Its output is validated by a blind hand read of a fixed sample, normally twenty records, in which an analyst reads the underlying document and compares it against the extracted value. Section 3 reports the outcomes of those reads.

Symbol resolution and the price join consume the subject company's filing identifier together with free daily price series, and produce the price windows at the horizons the design names. This stage carries the coverage gate, the coverage-differential clause where selection between outcome arms is plausible, and the price-ratio validation of section 7, which checks a resolved symbol against a quantity the filing itself asserts rather than against the plausibility of the symbol.

Control construction consumes the registrant universe and produces the comparison windows. Following section 5, it now also consumes the case builder's admissibility rules and the case arm's filing-activity distribution, so that controls are matched on activity rather than seeded at random from all registrants.

Scoring consumes the assembled panel and produces the primary statistic together with its controls: a permutation distribution built by relabelling within the design's blocking structure, Newey-West standard errors where the return windows overlap, and purged cross-validation with an embargo where the panel is used to fit a model. Settlement then consumes that statistic and the frozen preregistration and produces the verdict class, the Brier contribution scored against the stated prior, the consequence drawn from the frozen map, an entry in the research register and a dated report.

Each stage is exercised against synthetic fixtures before the unit is permitted to read real data, on the discipline of section 8: the fixture suite must show the instrument detecting a planted effect and refusing a planted null. Suites in this catalogue hold between four and nine fixtures. All of them passed before their runs began, the insider precursor suite at six of six and the post-termination suite at five of five once its two unfailable tests had been rebuilt as positive controls.

The exact file names for each stage of each unit, covering the code that implements it, the archived results file it writes, the preregistration that froze its bars and the dated report that narrates its settlement, are recorded in the repository and in that unit's entry in the research register.

The forward-ledger state quoted in section 11 is computed by the builder that assembles the opportunity desk. Its resolution counter counts usable spread-to-outcome pairs in the desk archive rather than asserting a count, and its expected-value routine states the basis it is running on and names the settlement it is waiting on. A replicator wanting to check that this catalogue is complete can query the register directly through its graph utility, which enumerates every closed line of work by category, or returns a single unit with its typed edges when given that unit's identifier.

> **On replication of the settlements themselves.** Re-running a unit that settled at a validity gate reproduces the gate failure rather than the sealed statistic, which is the intended behaviour. Where a successor design is named in a settlement, it is named with the specific change that would make the measurement possible, and those named revivals are the intended object of replication.

14. Conclusion

The measurements free filing data supports are those whose outcome is itself a filing event, such as whether a company that disclosed a received proposal later files a definitive agreement. The measurements it does not support are those whose outcome is a price on a security whose listing the event terminates, or whose inputs require transaction terms from episodes that never file a terms document, or whose control arm cannot be constructed from a free registrant universe without importing a composition confound.

Six defect classes account for the eight settlements catalogued here, and two recurred across units before being named. The response in each case was to record the wall, name the successor design that would breach it, and route the measurement to a forward ledger where reconstruction is not required. None of the confirmed measurements reported in section 10 has been validated forward; the deal-completion ledger holds 136 resolutions, recovered by backfill from the desk's snapshot history, with its measured base rates unadopted pending settlement of the registered spread hypothesis; and the appropriate confidence in every number above is that due an in-sample historical measurement carrying a documented proxy in its outcome definition.

References

[1] C. 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. Building diversified portfolios that outperform out-of-sample. *Journal of Portfolio Management*, 42(4):59–69, 2016.

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

Design sources named by the preregistrations above and recorded as such in the research register: Malmendier, Opp and Saidi (2016) for the cash-versus-stock revaluation asymmetry; Cohen, Malloy and Pomorski (2012) for the routine-filer classification; Palmrose, Richardson and Scholz (2004) and Hribar and Jenkins (2004) for the restatement-drift mechanism. Bibliographic details are held in the cited preregistrations.

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

A catalogue of the measurement limits reached by a preregistered event-research programme built on free corporate filings and free price series, grouped into named defect classes, with the remedies a data purchase supplies separated from those only forward accrual supplies. All figures are in-sample. Correspondence: M.A. Andreev, Quark Quantitative Research.