

Auditable Economic Data Analytics for Acquisition Premiums: A Provenance-Aware SEC Filing Study

Dezun Zhao*

Independent Researcher, Beijing, China

**Corresponding author: Dezun Zhao.*

Abstract

An acquisition premium is difficult to verify after a target delists or when a database no longer states its price benchmark. We assembled a cash-deal sample from SEC filings and the public MarkupMnA release. The 152 index entries correspond to 151 released agreement files; cash-term, filing, and premium checks retain 21 transactions. Premium sentences use distinct comparators. The primary analysis therefore keeps the ten observations tied to a dated pre-announcement close and reports the remaining eleven only as a sensitivity set. On identical leave-one-out folds, we compare the training-fold median and mean with OLS, Ridge, and Huber, using cash offer price and agreement length. None of the regressions beats the median for MAE and the mean for RMSE. The result is an auditable benchmark with a limited negative finding, not a causal or model-superiority claim.

Keywords

acquisition premium, public M&A transactions, SEC filings, data provenance, predictive benchmarking

1. Introduction

The meaning of an acquisition premium depends on the price used as its reference. That price is not always recoverable from public data: a target may delist, its ticker may change, or the definition used by a data vendor may no longer be available. A premium stated in a filing avoids some reconstruction, but the filing and its stated comparator still have to be preserved. This paper addresses that evidentiary task rather than ordinary data cleaning.

Our starting data are the U.S. cash deals in the public MarkupMnA release, which draws on SEC EDGAR agreement filings [11, 10, 13]. Its index lists 152 entries, whereas 151 agreement files are available in the release. We examined those files one at a time and logged the inclusion decision. Sixty-six contain both a cash term and an agreement date. An eligible transaction filing was located for 59 of them, and 21 filings state a premium that matches the cash offer. The analytical sample is narrower: it contains the ten filings that identify a dated closing price before the announcement. The other eleven observations cite another closing price, a volume-weighted average price, enterprise-value language, or an unclassified comparator. They are retained for sensitivity analysis, not pooled into the main outcome.

Two analytical choices follow from those records. A premium remains tied to a named price point and event [1, 4, 7]. A prior close, a VWAP, and an enterprise-value comparison consequently do not describe the same outcome, even when the percentages appear similar. Combining them would change what is being measured, not simply add cases. The second choice concerns evaluation with only ten primary observations. A flexible regression can appear helpful in such a small sample when a constant forecast would be safer. We therefore place OLS, Ridge, and Huber beside loss-aligned training-fold constants, using identical leave-one-out splits and no selection on held-out observations [14, 3]. The result is a row-level provenance audit and a homogeneous primary cohort. None of the tested regressions improves on either constant baseline. That finding applies only to the filing-derived sample described here.

2. Related Work

Prior M&A research places the premium within the wider transaction setting. Reference-point prices can affect merger behavior [1], while termination fees, acquirer type, and perceived hostility describe other deal conditions [8, 2, 12]. Work on target insiders similarly considers completion, premiums, and deal characteristics together [5]. This literature supports careful measurement; it does not make filing-stated premiums with different reference-price definitions economically equivalent. We therefore do not interpret a coefficient on a contract field as an economic effect.

Public agreement corpora make the document side of the task repeatable. MAUD provides expert-reviewed questions over merger agreements [15], and MarkupMnA supplies markup-based structural annotations for 151 SEC-sourced agreements [11, 10]. Our use is narrower than legal-question answering or section segmentation. The agreements supply the public starting point, and transaction filings supply the premium language and the comparison named by the filer. Each audit row therefore includes agreement and filing hashes, not just a model score.

The error measure also needs a clear benchmark. MAE is associated with a median forecast, whereas squared-error measures are associated with a mean forecast [6]. Both constants are retained in the evaluation. The regressions use fold-local scaling in scikit-learn [9]; this paper reports point estimates, not an estimator search over the fixed public table. Before fitting, we assign MAE to the median constant and RMSE to the mean constant. A regression is thus not credited because a favorable error measure was chosen after the results.

3. Provenance-Aware Method

3.1 Source Flow and Deterministic Audit

MarkupMnA has 152 index rows, but one indexed agreement has no released local contract CSV. The public release therefore supplies 151 canonical agreement files, and that is the denominator used here. The crosswalk retains the mapping-only entry and identifies its missing payload rather than hiding the one-row difference.

We then apply four predeclared screens. An agreement must first state a cash-only per-share term and an agreement date (66 cases). It must next yield an eligible SEC transaction filing within the specified form-priority and date window (59 cases). We search DEFM14A, SC 14D-9, PREM14A, DEFA14A, and SC TO-T from three days before through 220 days after the agreement date, in that order, and retain the earliest eligible filing within a form. A numeric premium must appear in transaction context and match the cash consideration (21 cases). The comparator in that sentence then sets primary or sensitivity status. For each retained row, we store the contract identifier, agreement and filing URLs, cash-offer and premium sentences, reference-price type/date/value, cohort label, inclusion rationale, and SHA-256 hashes of local source artifacts. The complementary ledger shows why the other canonical files stop: 85 lack cash/date evidence, seven lack an eligible filing, and 38 lack a matched explicit premium.

Audit records are kept at the row level. The crosswalk captures release position, original filename, EDGAR agreement link, payload status, and the reason a record is included or excluded from the canonical denominator. For a retained deal, the agreement identifier connects the contract evidence with the transaction filing selected for review. We save the cash-consideration passage separately from the premium passage. This helps distinguish the final offer from an earlier bid, historical proposal, or another market-price comparison in the

same filing. The classification follows the words used in the premium sentence: a VWAP remains a VWAP; enterprise-value language is not turned into a stock-price comparator; and a closing price is primary only when the sentence provides both a pre-announcement cue and a date. These rules were fixed before regression errors were inspected.

The pre-specified comparator rule is shown in Figure 3. It routes the ten dated pre-announcement closing-price observations apart from the eleven sensitivity-only observations before any model is fitted.

Two conditions determine whether a row can enter the benchmark. The audit counts must follow the path from 151 canonical files to 66 cash/date records, 59 eligible filings, and 21 explicit premiums. The retained record also needs the stated provenance fields and a valid primary-cohort label. If its hash, evidence sentence, reference-price field, or label is absent, it is not fitted. The local hash identifies the public artifact reviewed in this project. The filing URL lets a reader find the source record, but its existence does not establish that the local snapshot is complete.

The first 15 rows came from a manual seed table. A parser-assisted recall pass identified six further candidates. All six passed the same four screens before entering the 21-row table. Candidate generation did not use model errors or fitted coefficients. The cited release and SEC filing URLs are enough to repeat the checks; retention does not depend on an unpublished source.

Figure 1 traces source lineage, audit gates, cohort assignment, and the separate evaluation step. Exclusion totals appear beside the transition at which they occur, leaving the retained-record flow readable.

3.2 Outcome cohorts and evaluation

The outcome is the filing-stated premium in percentage points. Ten rows enter the primary cohort because their premium sentence identifies a dated last unaffected or pre-announcement closing price. The audit table also has three other-closing-price, six volume-weighted-average-price, one enterprise-value- based, and one unclassified comparator. Those eleven rows appear only in the mixed-definition sensitivity analysis. A reference value is quoted when the filing provides it or, for an algebraically valid per-share stock-price comparator, recorded as $P_offer/(1+premium\%/100)$. No per-share value is derived for enterprise-value-based or unclassified comparators. The stored value is an audit field, not a reconstructed outcome.

Mixing these comparators would change the outcome being modeled. A five-day VWAP captures a market window unlike the last unaffected closing price, and an enterprise-value phrase need not identify a per-share reference price. The sensitivity analysis asks only whether the descriptive error ordering changes when all retained records are shown; it does not enlarge the primary estimand. An implied reference value is kept as audit evidence, never substituted for a missing premium or used to create an observation.

Only log cash offer price and log agreement document length enter the models. The go-shop and no-shop fields are reported descriptively: the 21-row table has just two go-shop cases. We do not scale termination fees because no verifiable deal-value denominator is available for a comparable calculation. Restricting the feature set prevents rare clauses or differently scaled contract fields from creating an apparent prediction gain.

Leave-one-out cross-validation (LOOCV) holds out row i and computes the following training-fold constants:

$$\hat{y}_{-i}^{\text{median}} = \text{median}_{j \neq i}(y_j), \hat{y}_{-i}^{\text{mean}} = \frac{1}{n-1} \sum_{j \neq i} y_j. \quad (1)$$

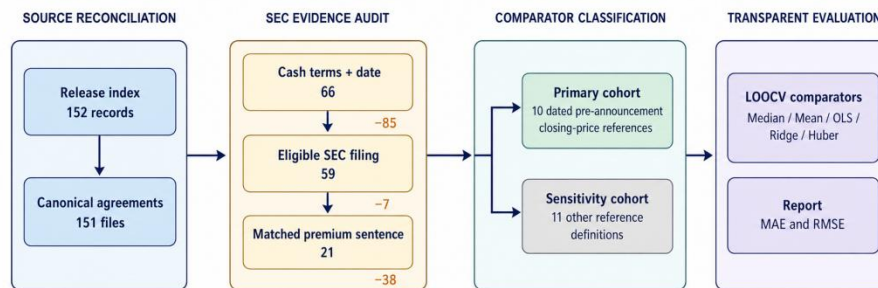
where y_j is the disclosed premium and n is the cohort size. OLS, Ridge with fixed $\alpha=1$, and Huber regression are transparent comparators fitted after fold-local standardization. Mean absolute error and root mean squared error are

$$MAE = \frac{1}{n} \sum_{i=1}^n |y_i - \hat{y}_{-i}|, RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_{-i})^2}. \quad (2)$$

MAE is evaluated against the training-fold median baseline and RMSE against the training-fold mean baseline. We report deterministic point estimates, but no fold-level t intervals, p values, or causal claims. A

regression can receive a predictive-improvement statement only if it beats both appropriate constants; otherwise the result is reported as a negative benchmark.

Figure 1: Audit trail used to construct the premium sample. The MarkupMnA index has 152 records, while the release provides 151 agreements. Cash/date, filing, and premium checks leave 66, 59, and 21 records. The stated comparator then separates ten primary observations from eleven sensitivity observations. The last panel summarizes the LOOCV comparison rather than a quantitative result



4. Audited Benchmark Results

4.1 Primary cohort

In the primary cohort, premiums range from 7.00 to 85.00 percentage points (mean 37.06; median 29.35). Re-running the recorded software environment with the same LOOCV splits returned the same numerical results. Table 1 reports the point estimates. The training-fold mean has a slightly smaller MAE, but the median is still the MAE benchmark because it is loss-aligned. Every regression is above the median on MAE and above the mean on RMSE. Huber is nearest to the constants, yet its MAE is 4.69 points above the median baseline and its RMSE is 9.19 points above the mean baseline.

The constants answer different loss questions. Within a training set, the fold median minimizes absolute deviation and the fold mean minimizes squared deviation. The mean happens to lower MAE by one point here, but it is not the MAE reference; likewise, the median's RMSE is not a defect of the median baseline. Each constant is assessed against its aligned loss. The relevant test is whether a regression with added features improves on both references. None does, so the result does not depend on a favorable summary statistic.

Table 1 is the only primary-model comparison. It reports the exact loss values and shows that none of the regressions benefits from selecting one error measure. The table describes the fixed ten rows; it is not a ranking of estimators for a broader population.

4.2 Mixed-reference sensitivity

Including all 21 retained rows changes the target because their premium comparators are heterogeneous. We therefore report this result only as a sensitivity check. The fold median and mean yield MAE/RMSE of 22.17/26.59 and 21.60/25.81; OLS, Ridge, and Huber yield 23.66/29.22, 23.56/29.04, and 23.23/29.14. No regression beats its relevant constant. This mixed outcome is not interchangeable with the primary cohort.

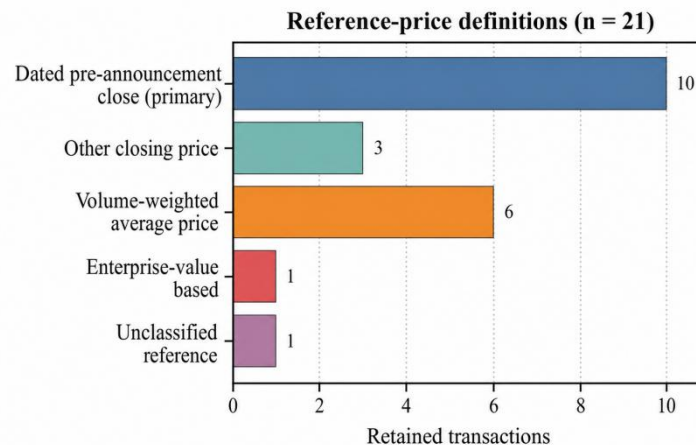
The sensitivity result asks whether the negative ordering is due solely to excluding eleven non-primary cases. It cannot establish that the primary rule is unnecessary. The mixed table combines several economic reference definitions, and its lower errors may reflect a different outcome distribution rather than added predictive signal. It uses the same observations, predictors, and LOOCV procedure as the primary analysis and is descriptive.

Sparse contractual fields remain descriptive. Five of the 21 rows contain a no-shop flag and two contain a go-shop flag. No termination-fee scaling or clause effect is estimated. With two reproducibly extracted features and public, cohort-defined observations, the regressions do not justify added complexity.

Table 1: Primary-cohort LOOCV results for 10 dated pre-announcement closing-price comparators. Errors are percentage points; lower is better. Deltas use the median for MAE and the mean for RMSE

Method	MAE	Delta MAE	RMSE	Delta RMSE
Fold median	24.05	0.00	30.54	+1.25
Fold mean	23.05	-1.00	29.28	0.00
OLS	29.70	+5.65	39.17	+9.89
Ridge	29.02	+4.97	38.17	+8.88
Huber	28.74	+4.69	38.47	+9.19

Figure 2: Counts by filing-stated reference-price definition. Ten dated pre-announcement closing-price observations form the primary cohort; the other eleven observations are used only for sensitivity analysis



5. Conclusion

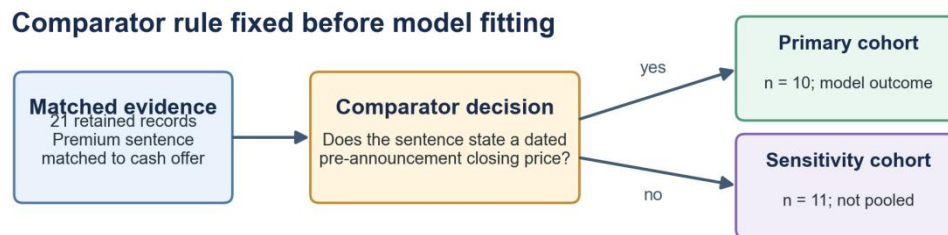
Public M&A agreement releases can support a filing-derived acquisition-premium benchmark only when the source trail remains attached to each record. In this case, the index has 152 entries and the public release has 151 agreement files. We recorded every retention decision and did not merge unlike reference-price definitions into one primary outcome. The 21-row audit table began as a 15-row pilot; a parser-assisted pass supplied six further candidates. Ten rows meet the dated pre-announcement-close rule used for primary modeling.

Within this data, the empirical result is straightforward. On the common LOOCV splits, OLS, Ridge, and Huber all have higher MAE than the median baseline and higher RMSE than the mean baseline. The comparison does not establish that deal characteristics never matter. It shows that this source-defined public table is too limited for a stronger predictive statement. A larger pre-specified filing collection, harmonized reference prices, and an external test set would be required for that purpose.

Figure 2 separates the cohorts directly. Ten primary rows use a dated pre-announcement closing-price definition; the other eleven are sensitivity rows. This is a classification rule, not a claim that the transactions represent every public acquisition. Later work should preserve the row-level provenance fields, choose one premium comparator before fitting, and reserve an independent collection for any performance claim.

The protocol documents how this sample was reported; it does not show that the selected filings exhaust the market. The 2020--2021 window, cash-only restriction, and explicit-premium requirement determine which rows enter the table. A changed MarkupMnA release, a corrected EDGAR document, or a newly located filing should produce a new hash manifest and source-flow record rather than silently replace a row. Future extensions should fix the reference-price rule, keep candidate generation separate from outcome modeling, and report exclusion categories with the retained cases. Higher-capacity predictors or quantified out-of-sample uncertainty are appropriate only after the collection supports a pre-specified train/test division. Until then, further extraction or modeling would not strengthen the stated evidence.

Figure 3: Cohort rule used before modelling. A premium statement enters the primary cohort when it identifies a dated pre-announcement closing price. The other retained statements are considered only in sensitivity analysis



References

- [1] Malcolm Baker, Xin Pan, and Jeffrey Wurgler. 2012. The Effect of Reference Point Prices on Mergers and Acquisitions. *Journal of Financial Economics* 106(1), 49-71. <https://doi.org/10.1016/j.jfineco.2012.04.010>
- [2] Leonce Barger, Frederik P. Schlingemann, Rene M. Stulz, and Chad J. Zutter. 2008. Why Do Private Acquirers Pay So Little Compared to Public Acquirers? *Journal of Financial Economics* 89(3), 375-390. <https://doi.org/10.1016/j.jfineco.2007.11.005>
- [3] Gavin C. Cawley and Nicola L. C. Talbot. 2010. On Over-Fitting in Model Selection and Subsequent Selection Bias in Performance Evaluation. *Journal of Machine Learning Research* 11, 2079-2107.
- [4] Peter Dodd and Richard Ruback. 1977. Tender Offers and Stockholder Returns: An Empirical Analysis. *Journal of Financial Economics* 5(3), 351-373. [https://doi.org/10.1016/0304-405X\(77\)90043-5](https://doi.org/10.1016/0304-405X(77)90043-5)
- [5] Jan Fidrmuc and Chunling Xia. 2023. Target Insiders' Preferences When Trading Before Takeover Announcements: Deal Completion Probability, Premium and Deal Characteristics. *European Financial Management* 29(5), 1296-1324. <https://doi.org/10.1111/eufm.12308>
- [6] Rob J. Hyndman and Anne B. Koehler. 2006. Another Look at Measures of Forecast Accuracy. *International Journal of Forecasting* 22(4), 679-688. <https://doi.org/10.1016/j.ijforecast.2006.03.001>
- [7] S. P. Kothari and Jerold B. Warner. 2007. Econometrics of Event Studies. *Handbook of Empirical Corporate Finance*, 3-36. <https://doi.org/10.1016/B978-0-444-53265-7.50015-9>
- [8] Micah S. Officer. 2003. Termination Fees in Mergers and Acquisitions. *Journal of Financial Economics* 69(3), 431-467. [https://doi.org/10.1016/S0304-405X\(03\)00119-3](https://doi.org/10.1016/S0304-405X(03)00119-3)
- [9] Fabian Pedregosa et al. 2011. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research* 12, 2825-2830.
- [10] Sukrit Rao, Pranab Islam, Rohith Bollineni, Shaan Khosla, Tingyi Fei, Qian Wu, Kyunghyun Cho, and Vladimir A. Kobzar. 2023. MarkupMnA: Markup-Based Segmentation of M&A Agreements. Technical paper accompanying the MarkupMnA release.
- [11] Sukrit Rao, Pranab Islam, Rohith Bollineni, Shaan Khosla, Tingyi Fei, Qian Wu, Kyunghyun Cho, and Vladimir A. Kobzar. 2023. MarkupMnA. Zenodo, version 1.1. <https://doi.org/10.5281/zenodo.8040341>
- [12] G. William Schwert. 2000. Hostility in Takeovers: In the Eyes of the Beholder? *The Journal of Finance* 55(6), 2599-2640. <https://doi.org/10.1111/0022-1082.00301>
- [13] U.S. Securities and Exchange Commission. 2026. EDGAR---Search and Access. Public filing system documentation. <https://www.sec.gov/search-filings>
- [14] Sudhir Varma and Richard Simon. 2006. Bias in Error Estimation When Using Cross-Validation for Model Selection. *BMC Bioinformatics* 7, 91. <https://doi.org/10.1186/1471-2105-7-91>
- [15] Steven Wang, Antoine Scardigli, Leonard Tang, Wei Chen, Dimitry Levkin, Anya Chen, Spencer Ball, Thomas Woodside, Oliver Zhang, and Dan Hendrycks. 2023. MAUD: An expert-annotated legal NLP

dataset for merger agreement understanding. In Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, pages 16369-16382. <https://doi.org/10.18653/v1/2023.emnlp-main.1019>

Funding

This research received no external funding.

Conflicts of Interest

The authors declare no conflict of interest.

Acknowledgment

This paper is an output of the science project.

Open Access

This chapter is licensed under the terms of the Creative Commons Attribution-NonCommercial 4.0 International License (<http://creativecommons.org/licenses/by-nc/4.0/>), which permits any noncommercial use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license and indicate if changes were made.

The images or other third party material in this chapter are included in the chapter's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the chapter's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.

