Received: 23 July 2026 | Revised: 12 September 2026 | Accepted: 14 September 2026 | Published online: 16 September 2026

Journal of Finance and Financial Management, Volume 2, 2026
https://h-tsp.com/
https://doi.org/10.5281/zenodo.22791260

Outlier or Signal? An Auditable Framework for SEC Filing-Derived Financial Data Cleaning

Lubo Zhou*

University of California, Berkeley, CA, USA
**Author to whom correspondence should be addressed.*

Abstract: ***Researchers often clean financial-statement panels before the research question is fully visible. Extreme SEC XBRL values can be construction errors, one-time accounting events, operating signals, or unresolved cases, but common preprocessing rules often collapse them into a single outlier category. This paper develops an auditable alternative. The pipeline downloads official SEC submissions and companyfacts data, constructs a provenance-preserving firm-year panel, identifies anomaly candidates without deleting them, and links each candidate to filing evidence used for review. In a 20-company retail and consumer discretionary sample covering 156 unbalanced firm-years within a 2016-2024 target window, the pipeline identifies 73 anomaly candidates across 50 firm-years. Human review classifies 38 candidates as business signals, 24 as ambiguous, 10 as accounting events, and 1 as a construction issue. The central case, Costco 2019, shows why the audit trail matters: a gross-profit reconciliation failure is not an operating anomaly but a line-item definition mismatch. In a within-sample inference demonstration, mechanical winsorization shifts the operating-margin persistence estimate and inflates its standard error, while audited cleaning preserves the evidence-supported estimate. The contribution is methodological: data lineage is tied to outlier-treatment decisions, so a researcher can see why a value was retained, corrected, set aside, or flagged.***

Keywords: SEC filings; XBRL; Data cleaning; Outliers; Financial statement data; Data provenance; Audit trail.

1. Introduction

Empirical finance and accounting research often begins after the dataset has already been cleaned. The consequential decisions are made earlier: which observations are treated as errors, which values are capped, which line-item definitions are treated as comparable, and which extreme observations remain as potential signals. These choices matter when researchers construct data directly from SEC filings. The data are machine-readable, but the construction process still requires judgment about tags, periods, units, accessions, amendments, and metric definitions.

This paper makes those decisions visible. An extreme value first becomes a reviewable anomaly candidate, not an automatic deletion. The candidate carries the evidence needed to interpret it: XBRL tag, accession number, period, unit, triggering rule, related quality checks, and filing-context snippets where available. The final treatment is then recorded as part of the dataset rather than hidden inside preprocessing.

The claim is deliberately methodological. The sample is not designed to prove predictive superiority over mechanical winsorization. It is designed to show how an auditable workflow can preserve the

evidence behind each treatment decision and separate mechanisms that would otherwise be collapsed into a generic outlier category. The framework is especially relevant for as-filed data, where a selected fact can be technically valid but inappropriate for a particular empirical construct.

The Costco 2019 case is the clearest example. The pipeline flags a gross-profit reconciliation failure because selected gross profit does not equal total revenue minus cost of revenue. The audit trail shows that the values themselves are not wrong. Total revenue includes membership fees, while Costco's gross-margin discussion is based on net sales less merchandise costs. The anomaly is therefore a construction-quality issue: the analyst's formula is misaligned with the company's line-item definition.

The paper contributes in three ways. First, it provides an end-to-end pipeline for constructing a SEC filing-derived firm-year panel from official sources while retaining provenance. Second, it connects data lineage to cleaning decisions: existing provenance practices document where a value came from, while this framework also records what happens after the value is flagged as anomalous. Third, it proposes a review codebook and triage workflow that distinguish construction issues, accounting events, business signals, and unresolved cases before a researcher changes the data.

This framing is useful for information systems research because it turns a familiar data-cleaning operation into a governed decision process. The pipeline is not only a way to collect XBRL facts; it is a way to expose the sequence of decisions that convert official filing data into empirical variables. That sequence matters because the same number can be treated as a data error in one construct, as a real event in another, and as a business signal in a third. A reproducible data pipeline is therefore incomplete unless it also records the treatment logic attached to flagged observations.

2. Related Work and Contribution

The paper relates to five bodies of work. The first is research on influential observations and outlier treatment in accounting. Leone et al. (2019) show that winsorization and truncation are common responses to influential observations, while also emphasizing that these methods can alter legitimate observations and affect inference. This paper does not reject mechanical baselines. It argues that mechanical transformations should be compared with an auditable classification process rather than treated as a default answer.

The second is research on XBRL, restatements, and structured filing data quality. Debreceny et al. (2010) document calculation and consistency issues in early SEC XBRL filings. Hennes et al. (2008) show why error-like and irregularity-like accounting events should not be collapsed into a single category in restatement research. Dechow et al. (2011) similarly treat financial-statement signals as evidence that requires interpretation rather than automatic exclusion.

The third is research on standardization choices in financial-statement data. Du et al. (2023) show that discrepancies between as-filed financial statements and standardized databases can affect empirical inference. This project takes the as-filed route and treats the researcher's own construction choices as part of the object to be audited.

The fourth is data cleaning and data wrangling. Rahm and Do (2000) describe data cleaning as part of a broader transformation and integration problem, and Kandel et al. (2011) show how interactive wrangling systems can make transformations inspectable. The present paper applies this logic to financial-statement construction, where the transformation is not merely technical: the analyst must decide whether an extreme value is an error, an event, a signal, or unresolved.

The fifth is data provenance and dataset documentation. Database research distinguishes forms of provenance that explain why and where data appear in query results (Buneman et al., 2001). Workflow-provenance work treats derivation history as metadata about how a data product is produced (Simmhan et al., 2005). Datasheets for datasets extend this concern to documentation that helps dataset creators and users communicate assumptions, limitations, and intended uses (Gebru et al., 2021). The present paper builds on that logic but applies it to a specific financial-data decision point: once a filing-derived value is flagged as anomalous, should it be corrected, set to missing, retained as an accounting event, retained as a business signal, or left for further review? The increment is not generic lineage; it is lineage tied to outlier-treatment decisions.

The practical gap is that financial-statement researchers often inherit data-cleaning rules from convenience rather than from an explicit audit model. A winsorization threshold can be reported in a methods section, but the individual values affected by that threshold are rarely linked back to filing evidence. Conversely, a hand correction can be justified in a notebook but vanish from the published dataset. The framework here treats the correction, retention, or review-needed decision as part of the data product. In that sense, the contribution is closer to decision provenance than to extraction alone.

3. Data and Pipeline

The study uses official SEC data sources: company submissions metadata, XBRL companyfacts data, and filing text snippets downloaded only for anomaly candidates (U.S. Securities and Exchange Commission, 2025). The configured demonstration sample covers 20 retail and consumer discretionary firms with a target window of fiscal years 2016 through 2024. The pipeline restricts the firm-year panel to annual 10-K and 10-K/A filings and excludes 10-Q filings. The current processed panel is unbalanced: selected annual XBRL filings under the current rules produce 156 firm-year observations, with 11 companies observed across all nine target years and 9 companies observed for a partial window.

Table 1 summarizes the working sample and pipeline outputs. Full company-level coverage breakdowns are available in an anonymized repository that will be provided upon acceptance.

Table 1: Study sample and pipeline outputs

Output	Current count
Companies in configured sample	20
Firm-year observations in processed panel	156
Fiscal-year target window	2016-2024
Companies with full 2016-2024 coverage	11
Companies with partial coverage	9
Anomaly candidates	73
Firm-years with at least one anomaly candidate	50
Full-sample human review rows completed	73

The pipeline uses a declared SEC User-Agent, conservative rate limiting, and local caching. Raw SEC JSON responses are preserved separately from processed outputs, so later revisions can rerun metric selection without overwriting source data.

For each firm-year, the pipeline selects annual XBRL facts for revenue, cost of revenue, gross profit, operating income, net income, SG&A, capital expenditures, assets, liabilities, equity, cash, inventory, current assets, and current liabilities. Derived variables include revenue growth, gross margin, operating margin, net margin, SG&A ratio, capex ratio, asset growth, leverage, current ratio, inventory

ratio, and year-over-year margin changes.

Metric selection follows configured tag priority lists. For duration metrics, the selector requires annual-length periods and prefers fiscal-year facts. For instant metrics, it requires a balance-sheet-date fact near the annual report date. The selected values retain taxonomy, XBRL tag, unit, period start and end, filing accession, selected filing URL, selection reason, and audit information about candidate facts and missing values.

The anomaly detector is intentionally conservative in what it does to the data. It creates a queue rather than a cleaned panel. The rule set includes economic threshold screens, year-over-year change screens, cross-metric reconciliation checks, robust distribution-tail screens, operating-margin sign-transition checks, and tag-discontinuity checks. Examples include unusually large revenue-growth changes, negative or unusually high margins, SG&A ratios above revenue, large operating-margin changes, nonpositive assets or revenue, and gross-profit reconciliation failures. These rules are designed to over-identify cases that deserve interpretation. They are not designed to decide treatment by themselves.

Each anomaly candidate receives a stable identifier, triggering rule, severity, raw value, prior value where relevant, change value where relevant, selected tag, filing accession, filing URL, and initial issue hint. The initial hint is deliberately weak. For example, a reconciliation warning can suggest a possible construction problem, and the presence of impairment or restructuring tags can suggest a possible accounting event. The final label still depends on review evidence. This separation between screen and treatment is the core safeguard against treating every tail observation as an error.

4. From Lineage to Treatment Decisions

The pipeline treats a cleaning decision as a data object rather than a hidden preprocessing step. This is the operational distinction between ordinary lineage and the audit framework in this paper.

The selected-fact table records where each metric came from: ticker, fiscal year, metric name, value, taxonomy, tag, unit, accession number, period dates, filing date, frame, and selection reason. This answers the ordinary lineage question of where the value originated and why that fact was selected over alternatives.

The anomaly-candidate table records why the selected value needs review: candidate identifier, trigger rule, severity, raw value, prior value, change value, quality-check hint, selected tag, accession, and filing URL. This step turns numerical extremeness into an inspectable case rather than a deletion.

The review template binds evidence to a treatment choice. The reviewer supplies a label, confidence level, evidence note, cleaning decision, and corrected value if correction is chosen. The review row links the original value, the anomaly trigger, the filing evidence, and the proposed treatment in one record. The cleaning decision log then records what actually changed: candidate identifier, raw value, cleaned value, reviewer label, cleaning decision, evidence note, and a changed flag. The cleaned panel is therefore never the only artifact. Every changed or retained anomaly remains connected to its raw source, trigger rule, evidence, and treatment decision.

The review codebook has four primary labels. A construction or extraction error means that the value is not valid for the intended empirical construct because of tag selection, unit mismatch, period mismatch, accession conflict, duplicate fact, parsing error, or metric-definition mismatch. An accounting event means the value is real but event-driven, such as impairment, restructuring, discontinued

operations, acquisition, litigation charge, inventory write-down, or tax valuation allowance. A business signal reflects real operating performance, business-model economics, demand shock, distress, or macro shock. Ambiguous means that available evidence is insufficient for confident classification.

The boundary between construction error and a possible fifth category, comparability issue, is important. The codebook keeps comparability inside the construction-error family because the treatment problem is the same: the value is not wrong as filed, but it is wrong for the analyst's current construct. The review template therefore records subtypes, such as tag selection, period selection, unit, accession conflict, formula definition, and line-item comparability.

The review protocol asks the reviewer to use three evidence layers. The first is numerical: the triggered metric, its prior-year value, and related ratios. The second is provenance-based: tag, unit, period dates, accession, annual filing selection, and quality-check warnings. The third is context-based: event tags and short filing snippets around terms such as impairment, restructuring, discontinued operations, acquisition, litigation, inventory write-down, goodwill, and store closure. A high-confidence label requires consistency across these layers. If the available evidence does not establish a mechanism, the row remains ambiguous and is retained with a review-needed flag rather than forced into a stronger category.

The current review should not be read as an independent blind inter-rater reliability study. A separate adjudication-support file was used to check that the completed labels were consistent with the recorded evidence, but the paper does not report a Cohen's kappa statistic because the process was not designed as a blind reliability experiment. A larger version of the study should separate seed-label construction, independent reviewer validation, and model-assisted classification. This distinction avoids overstating the present evidence while preserving the review output that is already complete.

5. Central Audit Case: Costco 2019

Costco 2019 is the central empirical case because it is directly auditable from selected facts and the filing's reported line-item definitions. The pipeline selects total revenue of $152.703 billion, cost of revenue of $132.886 billion, and gross profit of $16.465 billion from accession 0000909832-19-000019. A generic cross-metric rule compares selected gross profit with total revenue minus cost of revenue. That formula produces $19.817 billion, creating a $3.352 billion reconciliation gap.

The audit trail points to the filing context. Costco reports net sales of $149.351 billion and membership fees of $3.352 billion. Total revenue equals net sales plus membership fees, and the parsed membership-fee amount equals the reconciliation gap. The filing's gross-margin discussion is tied to net sales less merchandise costs. Using the net-sales denominator resolves the apparent inconsistency: net sales minus cost of revenue equals $16.465 billion, the selected gross-profit fact.

This case is not an extraction failure in the simple sense. The selected XBRL facts are internally meaningful. The failure is in the analyst's reconciliation formula: it assumes total revenue is the correct input for a gross-profit check, but Costco's gross-margin presentation separates membership fees from net sales economics. The correct audit conclusion is therefore not to delete an outlier. The pipeline surfaced a line-item comparability problem: the raw filing values should remain traceable, and the reconciliation rule should use a company-aware or metric-aware definition when membership fees are material.

The case is not surprising to analysts who already know Costco's membership-fee economics, but it is

useful as a hard audit example because any reader can reproduce the arithmetic from the accession and selected fact table. It also illustrates a broader point: an anomaly can be caused by a researcher's construction choice rather than by a bad SEC fact.

The case also clarifies why the framework uses the term construction issue rather than extraction error. If a researcher used total revenue as the denominator for a gross-margin construct, the resulting ratio would mix merchandise economics with membership-fee economics. If the researcher used net sales, the gross-profit reconciliation would align with the filing. Both choices can be implemented mechanically, but only the audit trail explains which construct is being estimated. In this setting, the treatment decision is not to overwrite the SEC fact. It is to flag the construct and make the line-item definition visible to downstream users.

This matters for broader panels because company-specific revenue structures are common. Membership fees, platform commissions, franchise revenue, licensing revenue, fuel sales, reimbursement revenue, and pass-through costs can all create cases where a generic formula is formally computable but substantively misaligned. A pipeline that only selects facts can miss this issue. A pipeline that records the reconciliation failure and the filing context can surface it before the mismatch becomes part of an empirical design.

6. Review Results and Mechanical Benchmark

The full-sample review is complete: 73 of 73 anomaly candidates have labels and treatment decisions. These labels show how heterogeneous the anomaly queue is once filing evidence and metric construction are taken into account.

Table 2: Full-sample review label distribution

Review label	Anomaly count	Share
Accounting event	10	13.7%
Ambiguous	24	32.9%
Business signal	38	52.1%
Construction error	1	1.4%

Table 3: Full-sample cleaning decision summary

Cleaning decision	Count	Changed count
Keep raw and flag	49	0
Needs more review	24	0

The cleaning-decision summary is itself informative. The current review does not overwrite any raw financial-statement values. The one construction issue is a metric-definition problem rather than a wrong XBRL fact, and the event and signal rows are retained with flags. The contribution in this sample is therefore governance, traceability, and classification rather than a large number of numerical corrections.

As a mechanical benchmark, 1%/99% winsorization is applied to the standard ratio metrics. Among the 73 reviewed candidate rows, it changes 12 candidate rows, corresponding to 8 unique firm-year-metric observations. Those mechanically changed observations include business-signal, accounting-event, and ambiguous labels. The Costco gross-profit reconciliation row is not in the winsorized ratio set, so the mechanical baseline does not diagnose the construction problem.

This benchmark is useful precisely because it is simple and familiar. Winsorization gives a researcher a reproducible transformation, but it does not explain why a value was extreme. In the reviewed sample,

the mechanically affected rows are not all errors. Some represent event-period losses, some represent operating shocks, and some remain ambiguous. At the same time, the one construction issue is outside the benchmark's ratio set. The comparison therefore shows two limits of a purely mechanical rule: it can alter observations that review evidence supports retaining, and it can fail to identify construct-definition problems that are not tail values of the winsorized metric.

Table 4: Key examples under the mechanical benchmark

Case	Metric	Review label	Mechanical status	Raw value	Winsorized value
Costco 2019 reconciliation	gross_profit	construction_error	not in winsorized metric set		
Starbucks 2020 revenue decline	revenue_growth	business_signal	unchanged	-11.3%	-11.3%

The implication is narrower than a claim that the audited cleaning procedure improves prediction. In this sample, the review mainly prevents two silent mistakes: treating event and signal observations as generic tails, and missing a construction problem that a ratio winsorization baseline would not diagnose.

The result that changed count equals zero should not be treated as a null implementation. It is a substantive outcome of the review. The evidence did not justify replacing the accounting-event and business-signal rows, and the Costco issue required a construct flag rather than a numerical overwrite. This is different from a raw-data default. Raw data alone say nothing about why the values are retained. The audited panel retains the same numbers but attaches reasons, labels, and decision logs. That distinction is important for reuse: another researcher can keep the same values, revise the construct, or exclude flagged rows in a sensitivity analysis without rediscovering the evidence trail.

7. Inference Demonstration

To test whether cleaning choices affect an empirical conclusion, the paper estimates operating-margin persistence on the same panel under three arms: raw, mechanical winsorization, and audited cleaning. The specification regresses current operating margin on lagged operating margin with company and fiscal-year fixed effects, clustering standard errors by company. Because the panel has only 18 firm clusters, Table 5 reports conventional cluster-robust p-values and wild cluster bootstrap p-values following the small-cluster concern emphasized by Cameron et al. (2008).

The persistence regression shows the inference channel that the small sample can support. Raw and audited estimates coincide at beta = -0.138, because the review did not overwrite operating-margin values. The 1%/99% mechanical winsorized arm changes 4 operating-margin values and raises the cluster-robust standard error from 0.079 to 0.138. Wild cluster bootstrap p-values are 0.208 for raw/audited and 0.341 for mechanical winsorization. Thus the current sample does not show a three-way divergence or a robust significance flip. It shows that a mechanical default can change precision and move the estimate away from the evidence-supported raw/audited position. This raw-audited equivalence is a finding rather than a coding failure: the review evidence says those values should be retained.

Table 5: Operating-margin persistence by cleaning arm

Dataset version	Beta	SE	Cluster p	Wild bootstrap p	R2	N	Clusters	Changed values
Raw	-0.138	0.079	0.098	0.208	0.700	108	18	0
Mechanical winsorized	-0.142	0.138	0.317	0.341	0.782	108	18	4
Audited	-0.138	0.079	0.098	0.208	0.700	108	18	0

The threshold scan in Table 6 is the strongest evidence in this section. As the winsorization rule becomes

more aggressive, the lagged-margin coefficient moves from -0.142 toward -0.071. Across the reported thresholds, mechanical winsorization does not produce a robust rejection under the small-cluster bootstrap; beta is instead a smooth function of the analyst's threshold choice. Because the coefficient is negative, it should be described as a within-sample margin-reversal estimate rather than translated into a positive-persistence economic magnitude. With a short panel and firm fixed effects, dynamic-panel bias can affect the level of the coefficient; the relevant comparison here is the within-specification wedge across cleaning arms.

Table 6: Winsorization threshold scan

Threshold	Changed values	Beta	SE	Wild bootstrap p	R2	N
0.01/0.99	4	-0.142	0.138	0.303	0.782	108
0.025/0.975	8	-0.137	0.147	0.396	0.799	108
0.05/0.95	14	-0.111	0.162	0.508	0.826	108
0.10/0.90	26	-0.071	0.151	0.605	0.845	108

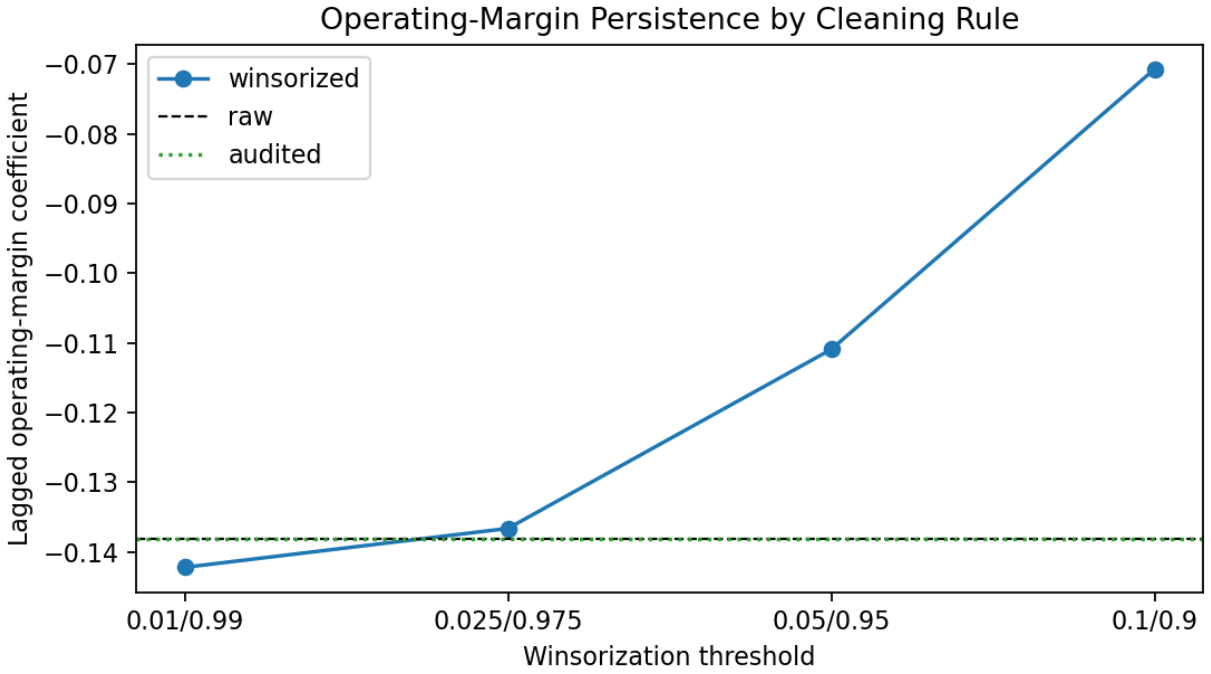

Figure 1: Winsorization threshold scan for the lagged operating-margin coefficient.

The small-sample wedge is driven mainly by a high-influence Macy's 2021 observation and by the lagged row that follows it. That concentration is not hidden here: it is both evidence for the paper's mechanism, because economically meaningful tail observations can carry identifying variation, and a reason the result should not be generalized before the expanded sample is reviewed.

The inference demonstration therefore has two roles. It is not presented as final evidence that audited cleaning improves estimation. Instead, it demonstrates the channel through which cleaning rules can matter: a small number of tail observations can alter both the coefficient path and the estimated precision. The fact that wild bootstrap p-values do not produce a robust rejection is part of the evidence, not an inconvenience to be hidden. A publishable inference claim would require the expanded sample, preregistered specifications, and enough clusters to evaluate whether the wedge survives outside the retail demonstration.

The demonstration also explains why the audit framework is complementary to sensitivity analysis. Threshold scans show how mechanical cleaning changes estimates as the analyst turns a tuning parameter. Audit labels show what kinds of observations are being changed. A threshold scan without

labels can show instability but not mechanism; labels without an inferential exercise can show mechanism but not consequences. The combination is the paper's proposed diagnostic: show the estimate path and identify what evidence-supported cases drive it.

8. Discussion and Next Steps

The evidence supports a focused thesis: SEC filing data cleaning should be auditable before it is automated. The anomaly candidates in this sample arise from different mechanisms. Costco 2019 is a construction-quality issue. Other reviewed rows involve impairment, restructuring, and operating-shock classifications. Starbucks 2020 is a business-signal example. The same numerical screen can therefore produce cases that require different treatment.

This framing also disciplines the role of judgment. Mechanical rules remain useful for screening and for large-scale consistency, but a cleaning decision should preserve evidence about why a value is changed or retained. In financial statement data, that evidence often lives in tags, accessions, periods, units, and filing text. The framework is most useful where the cost of a silent mistake is high: deleting a construction issue can leave the metric definition unresolved, while capping an event-driven loss can make the panel smoother but less faithful to the firm's history.

The framework also gives a practical answer to the scalability objection. The proposed workflow is not to hand review every firm-year. It is to review the cases where the cost of a mistaken treatment is highest: high-severity candidates, reconciliation failures, nonpositive core metrics, major sign transitions, and observations with event-context evidence. Repeated low-risk patterns can be handled by validated rules only after those rules have been checked against reviewed examples. In a larger panel, the same architecture can support active triage: high-confidence model-assisted labels are retained for audit, while low-confidence or high-severity cases return to human review.

The immediate limitation is scale. Manual review of 73 candidates is feasible, but a broad multi-industry panel can produce thousands of candidates. The next empirical step is therefore a preregistered expansion. A 500-company SEC seed universe has already been processed into a candidate primary sample after excluding financial institutions, insurance firms, real-estate firms, REIT-like SIC groups, utilities, missing-SIC firms, and firms without SEC companyfacts coverage. That expanded pipeline produces 348 primary-sample companies, 1,825 firm-year observations, 1,049 anomaly candidates, and a 375-row Tier 1 review queue. Those expanded labels should not be used for inference until the three cleaning arms, fixed-effects specifications, wild-cluster bootstrap protocol, threshold scans, and decision rules are time-stamped.

A higher-ceiling version of the paper would show that raw data, mechanical winsorization, and audited cleaning lead to different coefficients, uncertainty estimates, or economic conclusions in at least one common accounting regression. Without that link, the present contribution remains a governance and reproducibility contribution rather than evidence that cleaning changes published-style inference.

The preregistered expansion should therefore answer two questions. The first is descriptive: across industries, what share of anomaly candidates are construction issues, accounting events, business signals, or unresolved cases? The second is inferential: conditional on those labels, when does audited cleaning diverge from raw data and from mechanical winsorization in standard empirical designs? If the answer is that the divergence is rare, that is still useful evidence for researchers who worry that winsorization changes conclusions too often. If the answer is that divergence appears in specific designs or industries, the audit framework helps identify where cleaning choices deserve the most attention.

9. Conclusion

This paper reframes outlier handling in SEC filing data as a data-quality audit problem. The implementation detects anomalies without deleting them, attaches filing evidence, distinguishes construction issues from event and signal candidates, and preserves an audit trail for each cleaning decision.

The Costco case shows why this matters. The flagged value is not a bad data point, but a mismatch between a generic formula and a company's reporting economics. Without provenance, that distinction is easy to miss. With provenance, the cleaning decision becomes inspectable.

Transparency and Availability

The pipeline is designed to be reproducible from official SEC submissions and companyfacts data, subject to a declared SEC User-Agent, rate limiting, and local caching. Raw SEC JSON files, selected fact tables, anomaly candidates, review templates, and cleaning decision logs are preserved as separate artifacts so that source data and treatment decisions are not overwritten. The replication package will be made available through an anonymized repository upon acceptance. Only verified review tables are reported as empirical results in this draft. Draft notes, intermediate queues, and unverified suggestions are excluded from the reported label counts and cleaning-decision summaries.

References

[1] Buneman, P., Khanna, S., & Tan, W.-C. (2001). Why and where: A characterization of data provenance. In J. Van den Bussche & V. Vianu (Eds.), *Database theory - ICDT 2001* (Lecture Notes in Computer Science, Vol. 1973, pp. 316-330). Springer. https://doi.org/10.1007/3-540-44503-X_20

[2] Cameron, A. C., Gelbach, J. B., & Miller, D. L. (2008). Bootstrap-based improvements for inference with clustered errors. *Review of Economics and Statistics, 90*(3), 414-427. https://doi.org/10.1162/rest.90.3.414

[3] Debreceny, R. S., Farewell, S., Piechocki, M., Felden, C., & Graning, A. (2010). Does it add up? Early evidence on the data quality of XBRL filings to the SEC. *Journal of Accounting and Public Policy, 29*(3), 296-306. https://doi.org/10.1016/j.jaccpubpol.2010.04.001

[4] Dechow, P. M., Ge, W., Larson, C. R., & Sloan, R. G. (2011). Predicting material accounting misstatements. *Contemporary Accounting Research, 28*(1), 17-82. https://doi.org/10.1111/j.1911-3846.2010.01041.x

[5] Du, K., Huddart, S. J., & Jiang, X. D. (2023). Lost in standardization: Effects of financial statement database discrepancies on inference. *Journal of Accounting and Economics, 76*(1), Article 101573. https://doi.org/10.1016/j.jacceco.2022.101573

[6] Gebru, T., Morgenstern, J., Vecchione, B., Vaughan, J. W., Wallach, H., Daume III, H., & Crawford, K. (2021). Datasheets for datasets. *Communications of the ACM, 64*(12), 86-92. https://doi.org/10.1145/3458723

[7] Hennes, K. M., Leone, A. J., & Miller, B. P. (2008). The importance of distinguishing errors from irregularities in restatement research: The case of restatements and CEO/CFO turnover. *The Accounting Review, 83*(6), 1487-1519.

[8] Kandel, S., Paepcke, A., Hellerstein, J. M., & Heer, J. (2011). Wrangler: Interactive visual specification of data transformation scripts. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (pp. 3363-3372).

[9] Leone, A. J., Minutti-Meza, M., & Wasley, C. E. (2019). Influential observations and inference in accounting research. *The Accounting Review, 94*(6), 337-364. https://doi.org/10.2308/accr-52396

[10] Palmrose, Z.-V., Richardson, V. J., & Scholz, S. (2004). Determinants of market reactions to restatement announcements. *Journal of Accounting and Economics, 37*(1), 59-89. https://doi.org/10.1016/j.jacceco.2003.06.003

[11] Rahm, E., & Do, H. H. (2000). Data cleaning: Problems and current approaches. *IEEE Data Engineering Bulletin, 23*(4), 3-13.

[12] Simmhan, Y. L., Plale, B., & Gannon, D. (2005). A survey of data provenance in e-science. *ACM SIGMOD Record, 34*(3), 31-36. https://doi.org/10.1145/1084805.1084812

[13] U.S. Securities and Exchange Commission. (2025, April 8). *EDGAR application programming interfaces*. https://www.sec.gov/search-filings/edgar-application-programming-interfaces