JOURNAL OF MULTIDISCIPLINARY RESEARCH AND INTEGRATED SCIENCES

Volume 4, Issue 1, 2024 | Article ID: JMRIS-2024-006 | Pages: 58–66 | Article Type: Research Article

Received: 15-02-2024 | Revised: 23-02-2024 | Accepted: 05-03-2024 | Published: 15-03-2024

# Impact of Automated Data Engineering Systems on Data Quality and Operational Efficiency Among Small and Medium-Sized Enterprises in the United States: A Prospective Cohort Study

**Sonika Dwa[1], Md Sakib Ibn Jobaer[2], Mohammad Jashim Uddin[3], Israt Jahan[4], Mohammad Nazir[5]**

[1] *University of the Cumberlands, United States | sonikadwauoc@gmail.com*

[2] *University of Greenwich, United Kingdom*

[3] *Dhaka University of Engineering & Technology (DUET), Bangladesh*

[4] *East West University, Bangladesh*

[5] *Honda North America, United States*

## ABSTRACT

**Background:** Automated data engineering may improve the reliability and availability of enterprise information, but longitudinal evidence connecting implementation with measurable operational outcomes in United States small and medium-sized enterprises (SMEs) remains limited. This study evaluated changes in data quality and operational performance associated with the adoption of automated data engineering systems.

**Methods:** A longitudinal cohort of 240 US SMEs was observed monthly from January 2021 through December 2023, generating 8,640 firm-month records. The adoption cohort comprised 120 firms that initiated automation between January and December 2022; 120 firms constituted a non-adopting comparison cohort. Firm-level annual means were calculated for 2021 and 2023, treating 2022 as the implementation interval. Differences in within-firm changes were estimated using covariate-adjusted linear regression with heteroskedasticity-robust standard errors. Primary outcomes were data completeness and reporting turnaround time; secondary outcomes included accuracy, consistency, timeliness, error indicators, manual effort, pipeline latency, report punctuality, downtime and operating costs.

**Results:** Data completeness increased from 90.66% to 96.13% among adopters and from 89.03% to 90.35% in comparison firms; the adjusted between-cohort difference in change was +4.45 percentage points (95% confidence interval [CI] 4.17 to 4.74). Reporting turnaround decreased from 48.83 to 30.55 hours in adopters but increased from 43.69 to 44.64 hours in comparison firms (adjusted difference −18.88 hours; 95% CI −20.40 to −17.35). Adopters also had favourable adjusted changes in consistency (+5.94 percentage points), data timeliness (+8.41 percentage points), manual cleansing (−45.16 hours/month) and ETL latency (−23.50 minutes). Conversely, monthly data-operations costs increased by an additional US$446.79 (95% CI 305.37 to 588.21) relative to comparison firms.

**Conclusions:** Automation adoption was associated with substantial improvements in information quality and reporting efficiency during 2021–2023, alongside higher recorded operating expenditure. Baseline differences between cohorts and the observational design limit causal attribution. Evaluation of automated data engineering should combine multidimensional quality indicators, workflow outcomes and full economic accounting.

**Keywords:** automated data engineering; data quality; small and medium-sized enterprises; cohort study; difference-in-differences; extract-transform-load; operational efficiency; United States.

## 1. Introduction

Small and medium-sized enterprises (SMEs) increasingly depend on heterogeneous digital systems to coordinate transactions, customer relationships, fulfilment, financial reporting and regulatory obligations. An individual business may operate separate accounting, e-commerce, enterprise resource planning, customer management and operational applications, each with its own record structure and update frequency. The resulting information environment can be difficult to reconcile, particularly when engineering teams are small and staff rely on manually maintained spreadsheets or repeated extraction and validation tasks. The digital transformation literature therefore treats data architecture and organisational capabilities as foundations of enterprise performance rather than merely supporting infrastructure (Bharadwaj et al., 2013; Vial, 2019; Verhoef et al., 2021).

Data quality is multidimensional. Accuracy alone cannot determine whether a dataset is useful: completeness, consistency, timeliness, accessibility and fitness for its intended context also matter. Wang and Strong (1996) established a user-centred conception of information quality, while Pipino, Lee and Wang (2002) distinguished objective indicators from contextual assessments. Subsequent methodological work emphasised that measurement should be embedded in organisational processes and linked to clearly defined information requirements (Batini et al., 2009). Consequently, a seemingly technical improvement, such as fewer duplicate records, may have practical consequences only when it improves an operational process or reduces the effort required to produce trustworthy outputs.

Automated data engineering systems bring together scheduled ingestion, extract-transform-load (ETL) or extract-load-transform processing, schema enforcement, validation rules, exception

handling, data reconciliation and delivery of curated information to reporting applications. Integration and cleaning research has long recognised that heterogeneous source systems create semantic, structural and procedural errors that persist without explicit quality controls (Lenzerini, 2002; Rahm and Do, 2000; Vassiliadis, 2009). Automation can make these controls repeatable at scale, and contemporary error-detection methods can identify missing fields, duplication, implausible values and broken constraints before downstream use (Fan and Geerts, 2012; Abedjan et al., 2016). Nevertheless, automation may also replicate incorrect rules rapidly if governance and human oversight are inadequate.

The business benefits of technology adoption depend on more than the installation of software. Adoption, the intensity of use, training, prior digital maturity and complementary organisational assets can influence whether information systems create value (Sambamurthy, Bharadwaj and Grover, 2003; Zhu and Kraemer, 2005; Mithas, Ramasubbu and Sambamurthy, 2011). Technology–organisation–environment models and diffusion theory similarly suggest that enterprises differ in their readiness and incentives to implement new infrastructure (Tornatzky and Fleischer, 1990; Rogers, 2003). In particular, digitally mature SMEs may adopt advanced data engineering because they already have a greater capacity to benefit from it, potentially confounding simple comparisons with firms that retain manual processes.

The financial implications are similarly ambiguous. Labour-intensive reconciliation and reporting may decrease after standardised automation, yet cloud services, orchestration platforms, implementation support and oversight may increase recurring expenditures. Cloud computing offers scalable resources but introduces service, governance and pricing decisions that differ from traditional capital investment (Armbrust et al., 2010; Marston et al., 2011). International evidence has also described persistent SME-specific barriers involving skills, finance, infrastructure and organisational capacity (OECD, 2021). Thus, an evaluation limited to time savings would risk presenting an incomplete account of the organisational consequences of automation.

Longitudinal analysis is useful because implementation effects may develop gradually and because firms can differ systematically before adoption. However, the timing of technology adoption also complicates causal estimation: traditional two-way fixed-effects comparisons may be difficult to interpret when different groups receive an intervention at different times and effects change over time (Goodman-Bacon, 2021; Sun and Abraham, 2021; Callaway and Sant'Anna, 2021). An endpoint comparison in which all eventual adopters have completed implementation provides a straightforward descriptive contrast, although it still requires careful attention to pre-existing differences and common-trend assumptions. Against this background, the present study examined three questions: whether automation adoption was associated with improvements in data quality; whether those improvements coincided with more efficient operational reporting and processing; and whether recorded data-operations costs moved in the same direction as technical and workflow performance.

# 2. Methods

## 2.1 Study design and observation period

This prospective cohort analysis used monthly firm-level observations spanning 36 consecutive months from 1 January 2021 to 31 December 2023. Enterprises were identified by unique longitudinal codes, allowing repeated observations to be linked without direct business names in the analytic records. Calendar year 2021 represented the pre-adoption period. Automation adoption occurred during 2022, which was treated as the implementation and transition period. Calendar year 2023 represented the post-adoption period for all adopting enterprises. Firm-level annual means were calculated separately for the baseline and endpoint periods so that each enterprise contributed equally to the principal analysis, irrespective of the magnitude of its monthly transaction volume.

## 2.2 Cohort definition and longitudinal coverage

The cohort consisted of 240 US SMEs distributed across eight industry groups: retail and e-commerce, professional services, healthcare services, manufacturing, technology services, hospitality, logistics and transport, and financial services. Enterprise locations were classified into the Northeast, Midwest, South and West US Census regions. The adoption cohort contained 120 enterprises with an observed automation initiation date from January through December 2022. The comparison cohort contained 120 enterprises with no recorded automation activation throughout the observation interval. Each firm contributed 12 observations in every calendar year, producing a balanced panel of 8,640 enterprise-months. There was no recorded firm-level attrition during follow-up.

## 2.3 Exposure assessment

The main exposure was adoption of an automated data engineering system, defined operationally by the longitudinal indicator of automation activity and the recorded start month. A separate measure of automation coverage represented the percentage of relevant data engineering activities covered by automation in a given month. These variables distinguish the binary transition to an active system from the subsequent depth of its deployment. All adoption-cohort enterprises were inactive during 2021 and active by December 2022. Their observed duration of active deployment through December 2023 ranged from 13 to 24 months, with a median of 19 months. Mean automation coverage among adopters during 2023 was 77.37%. The comparison group remained at 0% automation coverage. Coverage was described as an implementation characteristic rather than entered as an adjustment covariate because it occurs after adoption and may be part of the mechanism connecting adoption with outcomes.

## 2.4 Outcomes and measurement domains

Two outcomes were designated as the primary measures of performance: data completeness, expressed as the percentage of expected fields populated, and reporting turnaround, expressed as hours required to deliver reporting outputs. These outcomes were selected to represent complementary dimensions of information quality and operational service delivery. A lower reporting-turnaround value indicates better performance, whereas a higher completeness percentage indicates better performance. Their joint evaluation avoids equating faster reporting with higher-quality information or treating quality gains as operationally meaningful without observing their delivery.

Secondary data-quality measures comprised record accuracy, consistency and timeliness percentages; duplicate-record rate; schema-error rate; and monthly critical data incidents. Operational measures included monthly manual cleansing hours, ETL batch latency in minutes, hours of unplanned downtime, the percentage of reports delivered on time, total reports delivered each month, and order-error rate. Monthly data-operations cost, measured in US dollars, was examined as an economic outcome. These recorded indicators provide a multidimensional measurement framework consistent with the distinction between intrinsic, contextual and process-related characteristics of information quality (Wang and Strong, 1996; Pipino, Lee and Wang, 2002; Cai and Zhu, 2015). No composite score was created, because combining dissimilar units would obscure the nature and direction of individual associations.

## 2.5 Baseline enterprise characteristics

Baseline covariates were measured in 2021 or recorded as stable enterprise characteristics. They included number of employees, firm age, annual revenue, digital-maturity score, data-engineering team size in full-time equivalents (FTEs), cloud utilisation, integration complexity on a 0–10 scale, industry, regulatory-exposure category and US Census region. The number of connected source systems was additionally summarised descriptively. These measures capture organisational scale, legacy complexity, sectoral structure and readiness to use digital

infrastructure. Selection of these characteristics was informed by the technology-adoption and information-management literature, which describes the importance of prior capability and organisational context (Tornatzky and Fleischer, 1990; Zhu and Kraemer, 2005; Khatri and Brown, 2010; Mithas, Ramasubbu and Sambamurthy, 2011).

### 2.6 Statistical analysis

Continuous baseline characteristics were summarised using mean and standard deviation; categorical characteristics were reported as counts and percentages. For each outcome and each enterprise, the mean of the 12 available monthly observations in 2021 and the corresponding mean in 2023 were computed. The within-enterprise change equalled the 2023 mean minus the 2021 mean. The unadjusted difference-in-differences (DiD) contrast was the mean change in the adoption cohort minus the mean change in the comparison cohort. This contrast measures divergence in observed changes rather than the simple difference between cohorts at the 2023 endpoint. The 2022 observations were excluded from this primary contrast because adoption was staggered across the transition year.

To adjust for measured baseline differences, the within-enterprise change was regressed on adoption-cohort membership, log-transformed baseline employee count, log-transformed annual revenue, baseline digital maturity, firm age, data-team FTEs, integration complexity, cloud utilisation, industry indicators, Census-region indicators and regulatory-exposure indicators. The estimated coefficient on adoption membership represents the adjusted between-cohort difference in mean change between 2021 and 2023. Ordinary least-squares models were fitted with HC3 heteroskedasticity-robust standard errors, and two-sided 95% confidence intervals were calculated. Transformations of employee count and annual revenue were used to limit the influence of right-skewed enterprise size. Inference was performed at the firm level, using one change observation per firm rather than treating correlated firm-months as statistically independent.

The principal contrast used the common 2023 endpoint for all adoption firms rather than comparing already-treated enterprises with later-treated enterprises during implementation. This approach avoids the problematic weighting of staggered treatment comparisons highlighted in recent econometric research (Goodman-Bacon, 2021; Sun and Abraham, 2021). It does not, however, remove bias from unmeasured differences in expected outcome trajectories. To assess observable pre-adoption dynamics, a separate linear monthly slope was estimated for each firm using its 12 observations during 2021; mean slopes were compared between cohorts with Welch unequal-variance tests. These checks were treated as diagnostic rather than proof of a valid parallel-trends assumption. Findings were interpreted as associations, not experimentally identified causal effects (Bertrand, Duflo and Mullainathan, 2004; Callaway and Sant'Anna, 2021).

### 2.7 Missingness, quality checks and multiplicity

The 240-by-36 enterprise-month structure was complete. Missing individual outcome values were infrequent: 39 records (0.45%) for data accuracy, 72 (0.83%) for reporting turnaround, 74 (0.86%) for on-time report delivery and 47 (0.54%) for order-error rate. Data-refresh delay had 60 missing monthly values but was not included among the reported endpoint outcomes. Missing values in automation start dates were structural in the non-adopting cohort, while months since adoption were undefined before implementation. An enterprise annual mean was calculated from the observed months for that outcome; all 240 firms had sufficient non-missing records in both endpoint years, so no firm was removed from the principal models and no missing monthly value was imputed. Variables with no missing records were analysed directly.

Because outcomes spanned several interrelated quality and efficiency domains, the principal emphasis was placed on effect sizes and confidence intervals. As a conservative supplementary assessment, Holm-adjusted significance probabilities were calculated across the 15 modelled outcome measures. Graphical analysis displayed monthly cohort means and normal-approximation 95% intervals based on variation between enterprises in each month, with the adoption year visually identified. This display shows the observed trajectories but is not a substitute for model-based inference. The structure and reporting of this observational study were informed by the STROBE recommendations (von Elm et al., 2007).

## 3. Results

### 3.1 Cohort characteristics and baseline comparability

A total of 240 enterprises contributed all 36 planned observation months, corresponding to 8,640 enterprise-month records. The adoption and comparison cohorts each included 120 firms. Relative to comparison enterprises, adoption firms had more employees at baseline (87.79 versus 61.01), higher annual revenue (US$19.35 million versus US$12.74 million), a larger recorded data-team complement (1.69 versus 1.24 FTEs) and a higher digital-maturity score (58.84 versus 46.32). Cloud utilisation was reported by 75.0% of adopters and 50.0% of comparison firms. These imbalances support the need to adjust for initial organisational capacity when interpreting performance changes. The cohorts also differed in their industry and geographic composition, although all eight industries and all four Census regions were represented in both groups (Table 1).

### 3.2 Unadjusted changes in data quality and operational measures

Data completeness rose by 5.48 percentage points in adoption enterprises, compared with an increase of 1.32 points in comparison enterprises; the unadjusted difference in changes was +4.16 points. Accuracy increased by 3.90 and 0.82 points, respectively, and consistency by 6.97 and 1.39 points. Data timeliness showed the largest absolute improvement among the four positive quality dimensions: +9.90 points in adopters compared with +1.84 points in comparison enterprises. Duplicate-record and schema-error rates declined appreciably in adopters, whereas changes in comparison firms were comparatively small. These patterns indicate that the observed gains extended beyond a single quality metric.

Reporting turnaround decreased by 18.28 hours in the adoption cohort while increasing by 0.96 hours among comparison enterprises. Manual cleansing work decreased by 49.89 hours per month in adopters versus 4.72 hours in the comparison cohort. ETL batch latency fell by 17.31 minutes among adopters but rose by 4.45 minutes in comparison enterprises, while on-time report delivery improved by 9.59 and 1.32 percentage points, respectively. Unplanned downtime and critical data incidents also decreased more in the adoption cohort. However, the raw change in monthly data-operations expenditure was positive in both groups and was greater among adopters, producing a difference in changes of +US$486.60 (Table 2).

### 3.3 Covariate-adjusted differences in change

After adjustment for enterprise scale, revenue, digital maturity, sector, region, cloud use, integration complexity, data-team size, firm age and regulatory exposure, automation adoption remained associated with greater increases in completeness (+4.45 percentage points; 95% CI 4.17 to 4.74) and greater reductions in reporting turnaround (−18.88 hours; 95% CI −20.40 to −17.35). Both differences had two-sided p values below 0.001. The direction and magnitude of the estimates were consistent with the unadjusted findings, despite pronounced baseline differences between the cohorts.

The adjusted change estimates also favoured adopters for accuracy (+3.27 percentage points), consistency (+5.94 points), timeliness (+8.41 points), duplicate-record rate (−1.41 points), schema-error rate (−1.06 points), manual cleansing (−45.16 hours/month), on-

time reports (+8.75 points), ETL latency (−23.50 minutes), unplanned downtime (−2.12 hours/month) and critical incidents (−0.24 events/month). In contrast, adoption was associated with a greater increase in monthly operating cost (+US$446.79; 95% CI 305.37 to 588.21). The smaller event rate for critical incidents should be interpreted cautiously because the pre-adoption slope comparison indicated differing trajectories for that outcome (Table 3).

### 3.4 Monthly trajectories and implementation pattern

The monthly trends distinguish the transition year from the pre- and post-adoption periods. Mean completeness among adopters was 90.66% in 2021, 93.02% during implementation in 2022 and 96.13% in 2023; corresponding comparison means were 89.03%, 89.62% and 90.35%. The year-by-year divergence was therefore gradual rather than an instantaneous step at a common initiation date, consistent with variation in rollout timing and deployment coverage. Figure 1 displays the longitudinal pattern with enterprise-based 95% confidence intervals.

Reporting turnaround exhibited a complementary trend. Adoption-cohort mean turnaround was 48.84 hours in 2021, 41.41 hours in 2022 and 30.56 hours in 2023; comparison means were 43.62, 44.14 and 44.68 hours, respectively. While adoption firms started with slower reporting, their endpoint turnaround was substantially lower. The monthly trajectory indicates progressive separation during and after implementation rather than a single uniform shift. Figure 2 illustrates this divergence. These visual patterns support the aggregate analysis but cannot determine whether concurrent changes in management, staffing or reporting requirements contributed to the observed improvements.

### 3.5 Pre-adoption diagnostics and outcome sensitivity

The 2021 cohort comparisons of firm-specific monthly slopes did not identify a statistically significant difference for either primary outcome: data completeness ($p=0.489$) or reporting turnaround ($p=0.551$). Similar null slope comparisons were observed for accuracy, consistency, timeliness, manual cleansing and ETL latency. In contrast, unplanned downtime ($p=0.032$) and critical data incidents ($p=0.006$) differed in their pre-adoption slopes; associations for these outcomes merit particular caution. A failure to detect a difference for the primary outcomes should not be interpreted as verification of counterfactual parallel trends, especially given the substantial baseline differences in digital maturity and enterprise size.

All 240 firms contributed valid annual means for both primary measures, and the small proportion of isolated missing monthly observations did not cause complete-case firm exclusions. Reported associations were large relative to their robust standard errors. Application of the Holm familywise adjustment across the 15 outcome models retained statistical significance for the adjusted contrasts discussed above, although multiplicity correction cannot address selection bias, measurement error or omitted time-varying influences. The 2023 comparison also captures a mixture of adoption durations, and therefore represents an average association after implementation rather than a uniform treatment effect at a fixed number of months since activation.

## 4. Discussion

### 4.1 Principal findings and interpretation

Across three years of monthly observation, adoption of automated data engineering systems was associated with better recorded information quality and more efficient operational reporting among US SMEs. The principal adjusted associations comprised a 4.45-percentage-point larger improvement in completeness and an 18.88-hour larger reduction in reporting turnaround relative to non-adopting firms. These were not isolated statistical patterns: accuracy, consistency, timeliness, duplicate-record control, schema validation, cleansing effort and ETL latency all moved in directions consistent with improved processing reliability. The findings suggest that systematic automation of recurring engineering activities can coexist with improvements across several stages of the information lifecycle, from ingestion and validation to reporting delivery.

The distinction between raw endpoint values and changes over time is important. Adoption firms entered the observation period with greater digital maturity, larger data teams, greater revenue and more widespread cloud use. They also had slower reporting turnaround and higher manual cleansing effort at baseline, which could create greater opportunities for process improvement. A comparison restricted to 2023 would conflate these initial differences with post-adoption performance. Analysing each enterprise's annual change and adjusting for initial capacity produces a more informative contrast, although not a fully causal one. In this setting, the common post-adoption calendar year is particularly useful because enterprises adopted at different months in 2022; comparisons during rollout would combine different stages of implementation and risk inappropriate weighting of already-treated and later-treated firms (Goodman-Bacon, 2021; Sun and Abraham, 2021).

### 4.2 Relationship to data quality and information systems research

The observed pattern is compatible with established theoretical accounts of multidimensional information quality. Increased completeness and accuracy can reflect more systematic validation of fields and source records; higher consistency can reflect harmonisation of definitions and schemas across applications; greater timeliness can reflect more predictable ingestion and delivery cycles. These relationships are conceptually consistent with the frameworks of Wang and Strong (1996), Pipino, Lee and Wang (2002), Batini et al. (2009) and Fan and Geerts (2012). However, the available measures describe recorded outcomes rather than individual configuration rules, and therefore cannot identify which component of automation accounted for each quality gain. Automated monitoring, reconciliation, scheduling and governance may act together rather than as mutually exclusive mechanisms.

Engineering research has demonstrated that integration complexity and data errors cannot be solved solely by moving information faster between systems. Source heterogeneity, inconsistent business rules and ambiguous identifiers require quality-aware transformation and exception management (Rahm and Do, 2000; Lenzerini, 2002; Kimball and Caserta, 2004; Vassiliadis, 2009). The parallel improvement in duplicate-record and schema-error indicators is therefore substantively relevant. It suggests that enterprises were not merely increasing reporting throughput while tolerating a greater error burden. Nevertheless, the study cannot independently audit the underlying records against authoritative ground truth; percentages and error rates depend on the definitions and monitoring practices used by the participating enterprises. Future evaluation should triangulate operational metrics with audited reference samples and explicit validation standards, as advocated by data-cleaning and governance scholarship (Abedjan et al., 2016; Khatri and Brown, 2010).

Information systems success models distinguish system quality, information quality and organisational net benefits (DeLone and McLean, 2003). The current results offer indicators in each of these domains: ETL latency and downtime relate to system performance, completeness and consistency to information quality, and reporting turnaround and manual work to operational consequences. This separation is analytically preferable to relying exclusively on a broad digital-transformation score. The improvement in report punctuality also indicates that faster workflows may translate into more dependable service commitments, but the study does not capture whether managers made better decisions, customers received better services or revenue increased because of the technology.

### 4.3 Operational efficiency and the economics of automation

Automation-associated reductions in manual cleansing and reporting turnaround are operationally meaningful because they describe work that must otherwise be performed repeatedly. The adjusted difference of approximately 45 hours per month in manual cleansing represents a substantial change in recurring workload at the enterprise level. Lower ETL latency and fewer downtime hours point to improvements in processing continuity and predictability. These findings align with the broader view that digital capabilities create value when technical resources are coupled with organisational processes and complementary human skills (Brynjolfsson, Hitt and Yang, 2002; Sambamurthy, Bharadwaj and Grover, 2003; Mithas, Ramasubbu and Sambamurthy, 2011). Time released from repetitive data maintenance might be reallocated to analysis, monitoring or higher-value activities, but staff redeployment and resulting productivity were not directly measured here.

A consequential counterpoint is the increase in recorded monthly data-operations expenditure. In the adjusted models, adopter firms experienced a US$446.79 greater cost increase than comparison enterprises, even though they reduced processing time and cleansing effort. This finding is consistent with a potential trade-off between higher spending on infrastructure or managed services and lower manual work, but the separate cost components were not identifiable. Total data-operations expenditure should not be interpreted as a full return-on-investment calculation, and the observed increase does not establish that automation was economically disadvantageous. An appropriate evaluation would distinguish implementation capital, licence and platform subscriptions, cloud consumption, external support, internal labour, avoided errors and downstream benefits. The technology investment and cloud business-model literature makes clear why recurring expenditure and organisational value need not move in parallel (Armbrust et al., 2010; Marston et al., 2011; OECD, 2021).

### 4.4 Organisational capabilities and implementation timing

The adoption cohort's greater baseline digital maturity, employee count, revenue and cloud uptake is informative in its own right. It indicates that the organisations initiating automated data engineering differed substantially from those that did not, raising the possibility that implementation was enabled by prior investment, specialised staff or stronger internal governance. Diffusion and organisational adoption models predict that technology use depends on organisational readiness and the surrounding technical environment (Tornatzky and Fleischer, 1990; Rogers, 2003). Evidence from enterprise digitalisation similarly suggests that the amount and manner of post-adoption use may be more consequential than an initial adoption decision alone (Zhu and Kraemer, 2005). This is relevant because measured automation coverage reached an average of 77.37% in 2023, rather than universally reaching complete automation.

The observed gradual separation in outcomes also has implications for evaluation design. Adoption started in different months, so a common time-zero intervention model would mask heterogeneity in exposure duration. The 2023 comparison provides a transparent mature-period snapshot for all adopters, but it does not isolate short-run implementation costs, adaptation delays or any nonlinear effect of months since adoption. More detailed event-time analyses, ideally incorporating comparison units that remain untreated and estimators developed for staggered adoption, would be valuable if source-level implementation histories, prospective analysis protocols and sufficiently detailed time-varying organisational variables were available (Callaway and Sant'Anna, 2021; Sun and Abraham, 2021).

### 4.5 Practical implications for SMEs

For SME decision-makers, the findings support assessing proposed automation with a balanced, measurable performance framework. Data owners should define quality requirements before implementation, record baseline performance in each critical source, and monitor completeness, accuracy, consistency, timeliness, duplication and schema failures independently. Engineering teams should pair process measures such as pipeline latency, failure recovery and manual cleansing with service outcomes such as report punctuality and turnaround. Governance arrangements are necessary to determine who is accountable for exceptions, rule changes and source-of-truth disputes; automated processing does not eliminate the need for review and stewardship (Khatri and Brown, 2010; Otto, 2011).

Implementation plans should also recognise that a smaller organisation may face disproportionate skill constraints and capacity pressures. Training time, source-system complexity and integration effort may influence how quickly a pipeline produces reliable outputs. A staged deployment that begins with high-frequency, business-critical data streams may allow operational indicators to be tested before expanding coverage. As digitalisation research emphasises, the value of technological capability depends on the coordination of infrastructure, work practices and strategic objectives rather than the acquisition of automation software alone (Bharadwaj et al., 2013; Vial, 2019; Verhoef et al., 2021). Budget assessments should therefore include recurring operating expenses and an explicit appraisal of labour, quality assurance and reporting benefits.

### 4.6 Strengths and limitations

Several features strengthen the analysis. The balanced panel includes 240 enterprises from multiple industries and regions, with 36 sequential observations for every firm. Explicit monthly automation status identifies the transition year and avoids classifying the entire 2022 period as universally treated. The primary outcomes capture distinct quality and delivery dimensions, supplemented by a broad set of technical and operational indicators. Enterprise-level annual aggregation prevents large transactional firms from implicitly receiving greater statistical weight in the primary estimates, while covariate adjustment addresses measured baseline differences. The 2021 slope diagnostics offer a limited test for obvious pre-existing divergence, and the presentation of both unadjusted and adjusted contrasts makes the magnitude of model adjustment visible.

### 4.7 Directions for future research

Future studies should combine operational telemetry with auditable definitions of data-quality outcomes, detailed expenditure categories and structured information on organisational change. Prospective collection of implementation milestones, staff training, governance procedures, workflow redesign and the degree of automation would permit investigation of mechanisms and heterogeneous benefits. Longer follow-up would clarify whether cost increases are temporary, whether quality improvements plateau, and whether freed engineering time translates into decision quality, financial returns or customer service. Designs using randomised or plausibly exogenous rollout timing, where feasible, would strengthen causal identification. In non-randomised settings, transparent cohort definitions and modern staggered-adoption estimators can reduce the risk of misleading comparisons (Goodman-Bacon, 2021; Callaway and Sant'Anna, 2021).

## 5. Conclusion

In this longitudinal cohort of 240 US SMEs, adoption of automated data engineering systems during 2022 was associated with appreciable improvements in completeness, accuracy, consistency, timeliness and error control by 2023, together with shorter reporting turnaround and less manual cleansing work. The favourable operational patterns coexisted with a higher monthly data-operations cost, emphasising the need to evaluate both performance and expenditure. Because adopting enterprises differed from comparison firms before implementation, the results support a strong association rather than definitive causal attribution. A comprehensive assessment of enterprise automation

should integrate validated information-quality indicators, delivery metrics, governance capability and full economic measurement over an adequate implementation horizon.

## References


Abedjan, Z., Chu, X., Deng, D., Fernandez, R.C., Ilyas, I.F., Ouzzani, M., Papotti, P., Stonebraker, M. and Tang, N. (2016) 'Detecting data errors: Where are we and what needs to be done?', Proceedings of the VLDB Endowment, 9(12), pp. 993–1004.

Armbrust, M., Fox, A., Griffith, R., Joseph, A.D., Katz, R., Konwinski, A., Lee, G., Patterson, D., Rabkin, A., Stoica, I. and Zaharia, M. (2010) 'A view of cloud computing', Communications of the ACM, 53(4), pp. 50–58. https://doi.org/10.1145/1721654.1721672.

Batini, C., Cappiello, C., Francalanci, C. and Maurino, A. (2009) 'Methodologies for data quality assessment and improvement', ACM Computing Surveys, 41(3), article 16, pp. 1–52. https://doi.org/10.1145/1541880.1541883.

Bertrand, M., Duflo, E. and Mullainathan, S. (2004) 'How much should we trust differences-in-differences estimates?', The Quarterly Journal of Economics, 119(1), pp. 249–275. https://doi.org/10.1162/003355304772839588.

Bharadwaj, A., El Sawy, O.A., Pavlou, P.A. and Venkatraman, N. (2013) 'Digital business strategy: Toward a next generation of insights', MIS Quarterly, 37(2), pp. 471–482. https://doi.org/10.25300/MISQ/2013/37:2.3.

Brynjolfsson, E., Hitt, L.M. and Yang, S. (2002) 'Intangible assets: Computers and organizational capital', Brookings Papers on Economic Activity: Microeconomics, 2002, pp. 137–198.

Cai, L. and Zhu, Y. (2015) 'The challenges of data quality and data quality assessment in the big data era', Data Science Journal, 14, article 2. https://doi.org/10.5334/dsj-2015-002.

Callaway, B. and Sant'Anna, P.H.C. (2021) 'Difference-in-differences with multiple time periods', Journal of Econometrics, 225(2), pp. 200–230. https://doi.org/10.1016/j.jeconom.2020.12.001.

DeLone, W.H. and McLean, E.R. (2003) 'The DeLone and McLean model of information systems success: A ten-year update', Journal of Management Information Systems, 19(4), pp. 9–30. https://doi.org/10.1080/07421222.2003.11045748.

Fan, W. and Geerts, F. (2012) Foundations of Data Quality Management. San Rafael, CA: Morgan & Claypool.

Goodman-Bacon, A. (2021) 'Difference-in-differences with variation in treatment timing', Journal of Econometrics, 225(2), pp. 254–277. https://doi.org/10.1016/j.jeconom.2021.03.014.

Jagadish, H.V., Gehrke, J., Labrinidis, A., Papakonstantinou, Y., Patel, J.M., Ramakrishnan, R. and Shahabi, C. (2014) 'Big data and its technical challenges', Communications of the ACM, 57(7), pp. 86–94. https://doi.org/10.1145/2611567.

Khatri, V. and Brown, C.V. (2010) 'Designing data governance', Communications of the ACM, 53(1), pp. 148–152. https://doi.org/10.1145/1629175.1629210.

Kimball, R. and Caserta, J. (2004) The Data Warehouse ETL Toolkit: Practical Techniques for Extracting, Cleaning, Conforming, and Delivering Data. Indianapolis, IN: Wiley.

Lenzerini, M. (2002) 'Data integration: A theoretical perspective', in Proceedings of the Twenty-First ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems. New York: ACM, pp. 233–246. https://doi.org/10.1145/543613.543644.

Marston, S., Li, Z., Bandyopadhyay, S., Zhang, J. and Ghalsasi, A. (2011) 'Cloud computing—The business perspective', Decision Support Systems, 51(1), pp. 176–189. https://doi.org/10.1016/j.dss.2010.12.006.

Mithas, S., Ramasubbu, N. and Sambamurthy, V. (2011) 'How information management capability influences firm performance', MIS Quarterly, 35(1), pp. 237–256.

OECD (2021) The Digital Transformation of SMEs. OECD Studies on SMEs and Entrepreneurship. Paris: OECD Publishing. https://doi.org/10.1787/bdb9256a-en.

Otto, B. (2011) 'Organizing data governance: Findings from the telecommunications industry and consequences for large service providers', Communications of the Association for Information Systems, 29, article 3. https://doi.org/10.17705/1CAIS.02903.

Pipino, L.L., Lee, Y.W. and Wang, R.Y. (2002) 'Data quality assessment', Communications of the ACM, 45(4), pp. 211–218. https://doi.org/10.1145/505248.506010.

Rahm, E. and Do, H.H. (2000) 'Data cleaning: Problems and current approaches', IEEE Data Engineering Bulletin, 23(4), pp. 3–13.

Rogers, E.M. (2003) Diffusion of Innovations. 5th edn. New York: Free Press.

Sambamurthy, V., Bharadwaj, A. and Grover, V. (2003) 'Shaping agility through digital options: Reconceptualizing the role of information technology in contemporary firms', MIS Quarterly, 27(2), pp. 237–263.

Sun, L. and Abraham, S. (2021) 'Estimating dynamic treatment effects in event studies with heterogeneous treatment effects', Journal of Econometrics, 225(2), pp. 175–199. https://doi.org/10.1016/j.jeconom.2020.09.006.

Tornatzky, L.G. and Fleischer, M. (1990) The Processes of Technological Innovation. Lexington, MA: Lexington Books.

Vassiliadis, P. (2009) 'A survey of extract–transform–load technology', International Journal of Data Warehousing and Mining, 5(3), pp. 1–27. https://doi.org/10.4018/jdwm.2009070101.

Verhoef, P.C., Broekhuizen, T., Bart, Y., Bhattacharya, A., Dong, J.Q., Fabian, N. and Haenlein, M. (2021) 'Digital transformation: A multidisciplinary reflection and research agenda', Journal of Business Research, 122, pp. 889–901. https://doi.org/10.1016/j.jbusres.2019.09.022.

Vial, G. (2019) 'Understanding digital transformation: A review and a research agenda', Journal of Strategic Information Systems, 28(2), pp. 118–144. https://doi.org/10.1016/j.jsis.2019.01.003.

von Elm, E., Altman, D.G., Egger, M., Pocock, S.J., Gøtzsche, P.C. and Vandenbroucke, J.P. (2007) 'The Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) statement: Guidelines for reporting observational studies', PLoS Medicine, 4(10), e296. https://doi.org/10.1371/journal.pmed.0040296.

Wang, R.Y. and Strong, D.M. (1996) 'Beyond accuracy: What data quality means to data consumers', Journal of Management Information Systems, 12(4), pp. 5–33. https://doi.org/10.1080/07421222.1996.11518099.

Zhu, K. and Kraemer, K.L. (2005) 'Post-adoption variations in usage and value of e-business by organizations: Cross-country evidence from the retail industry', Information Systems Research, 16(1), pp. 61–84. https://doi.org/10.1287/isre.1050.0045.

## Table 1. Enterprise characteristics during the pre-adoption year (2021).

| Characteristic | Automation cohort (n=120) | Comparison cohort (n=120) |
|---|---|---|
| Enterprises, n | 120 | 120 |
| Employees in 2021, mean (SD) | 87.79 (63.89) | 61.01 (45.98) |
| Firm age, years, mean (SD) | 13.62 (6.50) | 12.98 (5.71) |
| 2021 annual revenue, US$ million, mean (SD) | 19.35 (15.62) | 12.74 (12.15) |
| Digital maturity score, mean (SD) | 58.84 (12.80) | 46.32 (11.21) |
| Data team size, FTEs, mean (SD) | 1.69 (1.11) | 1.24 (0.86) |
| Integration complexity, 0–10, mean (SD) | 4.40 (1.89) | 4.56 (1.80) |
| Source systems, mean (SD) | 10.72 (3.11) | 10.02 (3.00) |
| Cloud use, n (%) | 90 (75.0) | 60 (50.0) |
| High regulatory exposure, n (%) | 18 (15.0) | 29 (24.2) |
| Retail/e-commerce, n (%) | 30 (25.0) | 21 (17.5) |
| Professional services, n (%) | 17 (14.2) | 25 (20.8) |
| Healthcare services, n (%) | 12 (10.0) | 17 (14.2) |
| Technology services, n (%) | 17 (14.2) | 10 (8.3) |
| South Census region, n (%) | 42 (35.0) | 37 (30.8) |
| Northeast Census region, n (%) | 30 (25.0) | 24 (20.0) |

SD, standard deviation; FTE, full-time equivalent. Each enterprise contributes one baseline observation. Industry and region rows are selected categories; both cohorts contain firms from all eight industries and all four Census regions.

## Table 2. Mean outcomes in 2021 and 2023 and unadjusted between-cohort differences in change.

| Outcome | Adopt. 2021 | Adopt. 2023 | Compar. 2021 | Compar. 2023 | Raw DiD |
|---|---|---|---|---|---|
| Completeness (%) | 90.66 | 96.13 | 89.03 | 90.35 | +4.16 |
| Accuracy (%) | 92.01 | 95.91 | 90.86 | 91.68 | +3.08 |
| Consistency (%) | 85.78 | 92.75 | 84.02 | 85.41 | +5.58 |
| Timeliness (%) | 79.08 | 88.98 | 77.29 | 79.13 | +8.06 |
| Duplicate records (%) | 2.91 | 1.46 | 3.27 | 3.16 | -1.34 |
| Schema errors (%) | 1.96 | 0.85 | 2.22 | 2.11 | -1.00 |
| Manual cleansing (h/month) | 152.02 | 102.13 | 131.76 | 127.04 | -45.17 |
| Reporting turnaround (h) | 48.83 | 30.55 | 43.69 | 44.64 | -19.24 |
| On-time reports (%) | 82.13 | 91.72 | 79.89 | 81.21 | +8.27 |
| ETL latency (min) | 64.98 | 47.68 | 62.68 | 67.13 | -21.76 |
| Unplanned downtime (h/month) | 4.21 | 2.23 | 4.26 | 4.33 | -2.05 |
| Critical incidents (/month) | 0.63 | 0.35 | 0.65 | 0.61 | -0.23 |
| Operations cost (US$/month) | 14,752.06 | 15,442.52 | 13,131.77 | 13,335.62 | +486.60 |

Values are the means of firm-specific annual averages. Raw DiD = (adoption 2023 − adoption 2021) − (comparison 2023 − comparison 2021); the sign must be interpreted in relation to the outcome. All 240 firms contribute to each row.

## Table 3. Adjusted differences in change, confidence intervals and 2021 trajectory diagnostics.

| Outcome | Adjusted difference | 95% CI | p value | 2021 slope p |
|---|---|---|---|---|
| Completeness (%) | +4.45 | 4.17 to 4.74 | <0.001 | 0.489 |
| Accuracy (%) | +3.27 | 3.05 to 3.48 | <0.001 | 0.810 |
| Consistency (%) | +5.94 | 5.60 to 6.28 | <0.001 | 0.979 |
| Timeliness (%) | +8.41 | 8.02 to 8.80 | <0.001 | 0.936 |
| Duplicate records (%) | -1.41 | -1.49 to -1.33 | <0.001 | 0.088 |
| Schema errors (%) | -1.06 | -1.12 to -1.01 | <0.001 | 0.722 |
| Manual cleansing (h/month) | -45.16 | -48.97 to -41.35 | <0.001 | 0.113 |
| Reporting turnaround (h) | -18.88 | -20.40 to -17.35 | <0.001 | 0.551 |
| On-time reports (%) | +8.75 | 8.13 to 9.37 | <0.001 | 0.870 |
| ETL latency (min) | -23.50 | -24.78 to -22.22 | <0.001 | 0.497 |
| Unplanned downtime (h/month) | -2.12 | -2.32 to -1.93 | <0.001 | 0.032 |
| Critical incidents (/month) | -0.24 | -0.34 to -0.14 | <0.001 | 0.006 |
| Operations cost (US$/month) | +446.79 | 305.37 to 588.21 | <0.001 | 0.900 |

All outcome models include 240 enterprises and adjust for log baseline employment, log annual revenue, digital maturity, firm age, data-team FTEs, integration complexity, cloud use, industry, region and regulatory exposure. Standard errors are HC3 robust. 2021 slope p compares enterprise-level pre-adoption monthly slopes between cohorts (Welch test). CI, confidence interval. Holm adjustment across 15 modelled outcomes did not change the statistical significance of the reported 13 adjusted contrasts.

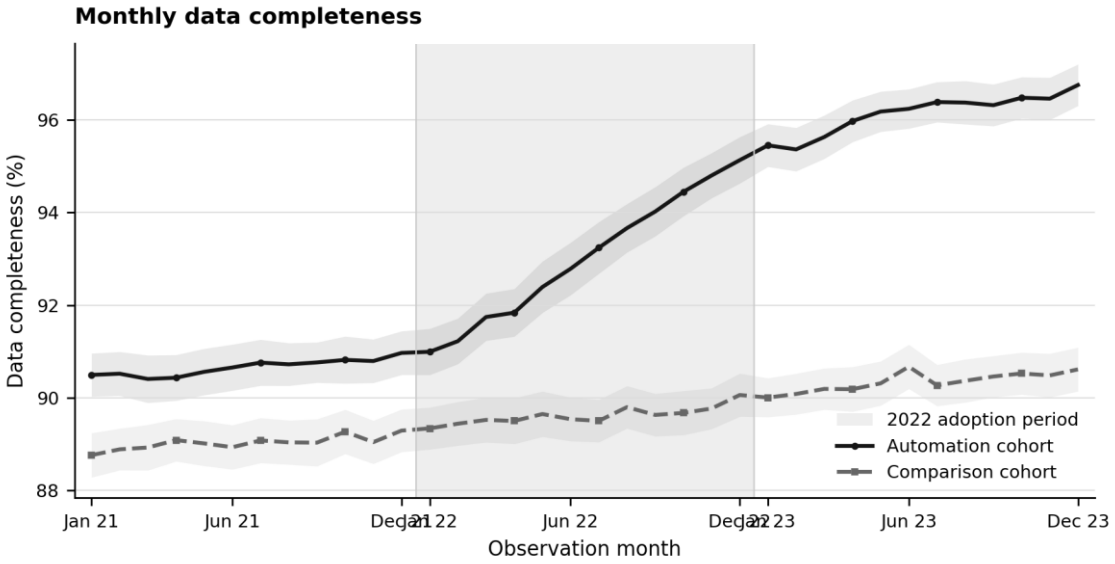


**Figure 1. Monthly mean data completeness in the automation and comparison cohorts (2021–2023). Shaded vertical interval identifies the 2022 adoption year; narrow bands represent approximate 95% intervals across firms within each month.**

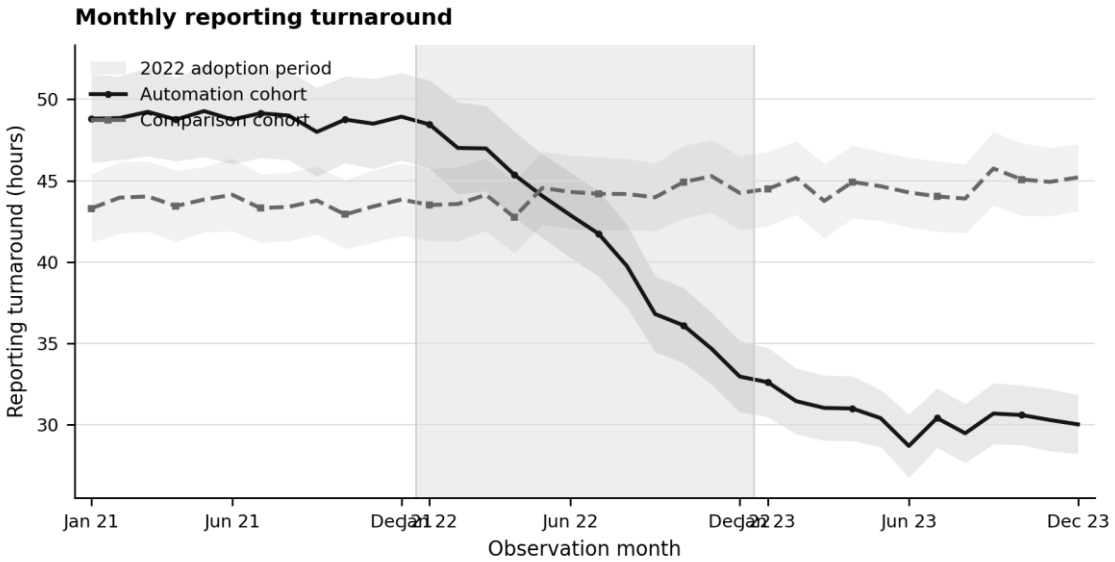


**Figure 2. Monthly mean reporting turnaround time in the automation and comparison cohorts (2021–2023). Lower values represent faster delivery. Shaded vertical interval identifies the 2022 adoption year; narrow bands represent approximate 95% intervals across firms within each month.**

**How to cite:** Dwa S, Jobaer MSI, Uddin MJ, Jahan I, Nazir M (2024). Impact of Automated Data Engineering Systems on Data Quality and Operational Efficiency Among Small and Medium-Sized Enterprises in the United States: A Prospective Cohort Study. Journal of Multidisciplinary Research and Integrated Sciences, 4(1), 58–66.

© 2024 Journal of Multidisciplinary Research and Integrated Sciences (JMRIS). Open access under CC BY 4.0.