Australian Journal of Science and Technology

ISSN Number (2208-6404)
Volume 10; Issue 1; March 2026

Review Article

Significance of source data quality for the functioning of predictive algorithms in product management

Elena Levi*

Product Director, Business Applications & Ecosystem Payoneer, Petah Tikva, Israel

ABSTRACT

This article examines how the quality of source data affects predictive algorithms used in product-oriented information systems and the reliability of decisions derived from them. The growing reliance on experimentation platforms, recommendation engines, and pricing models underscores the study's relevance. The novelty lies in integrating data-centric artificial intelligence, machine learning operations (MLOps) practices, and data governance into a unified conceptual framework for predictive decision-making flows in product-oriented information systems. The study summarizes recent publications on data quality dimensions, management practices and operational pipelines, including research on data-centric improvement cycles and MLOps maturity models. Special attention is paid to modeling behavior under controlled data defects and drifts, as well as to feedback loops between data problems, algorithm outputs, and product outcomes. The goal is to develop an integrated conceptual model that connects data quality characteristics, pipeline controls and product workflows. Analytical and comparative methods, along with a targeted literature review and conceptual synthesis, are applied. The conclusion offers practical guidelines for product leaders and cross-functional teams designing predictive products and internal processes that rely on data-driven decisions.

Keywords: Data quality, decision support, drift, experimentation, governance, machine learning operations, predictive models, product management, product systems, reliability

Submitted: 14-02-2026, **Accepted:** 03-03-2026, **Published:** 30-03-2026

INTRODUCTION

Predictive algorithms increasingly guide product decisions across discovery, prioritization, launch, and growth by shaping opportunity sizing, experimentation readouts, targeting, pricing, and churn or retention interventions. Decision quality depends on the integrity of upstream events, identifiers, timestamps, and governance rules; incompleteness, delay, bias, or semantic inconsistency can produce outputs that appear coherent in dashboards yet fail under real-world market feedback, pushing teams back toward intuition after repeated "data-driven" disappointments. A comparable dependence appears in telemetry-intensive environments (smart metering, industrial internet of things, anomaly detection), where identifier gaps, time misalignment, latency, and schema evolution propagate through feature computation and online inference into operational actions.

The article analyses how source data quality shapes predictive behavior in product-oriented information systems and specifies the conditions under which outputs remain dependable for strategic and day-to-day decisions. Three tasks are addressed:

1) Systematizing data-quality dimensions that matter for predictive product work and linking them to recurring decision types
2) Synthesizing recent evidence on model sensitivity to defects and drift
3) Proposing an integrative framework connecting data-quality practices, machine learning operations (MLOps) controls, and product workflows, with emphasis on decision stability, latency, and organizational trust.

MATERIALS AND METHODS

A systematic literature review was conducted in accordance with preferred reporting items for systematic reviews and meta-analyses 2020 reporting guidelines. The review focused on evidence on how source data quality influences predictive models used in product-oriented information systems and

Address for correspondence: Elena Levi, Petah Tikva, Israel. E-mail: elena.frum@gmail.com

telemetry-driven settings. Eligible records were peer-reviewed journal articles and full conference papers (January 2020–December 2025) that examined data-quality dimensions (accuracy, completeness, consistency, timeliness, validity, uniqueness) or operational defect types (missingness, label noise, identifier/timestamp faults, schema drift, distribution shift/drift) and reported extractable outcomes on model behavior, operational stability, monitoring/controls, or decision consequences. Exclusions covered short items without full text, non-peer-reviewed opinion pieces lacking reproducible methods, papers limited to generic data cleaning without linkage to predictive pipelines, and studies with insufficient outcome or pipeline-stage detail. The included studies were organized by defect category, affected pipeline stage (collection/labeling, ingestion/extract, transform, load [ETL], feature computation, training, serving/inference, monitoring), and operational setting (product analytics/experimentation, recommendation/pricing, streaming telemetry/monitoring).

Searches were executed in December 2025; the final search date was December 20, 2025. Databases comprised Scopus, Web of Science Core Collection, IEEE Xplore, ACM Digital Library, ScienceDirect, SpringerLink, plus Google Scholar for forward citation chasing; reference lists were hand-searched on 21–22 December 2025. Search strings were adapted per platform while retaining a consistent concept structure combining data quality/defects/drift, predictive modeling/machine learning (ML), production pipelines/MLOps, and product or operational decision settings; full syntaxes were archived for reproducibility.

Records were deduplicated, screened by title/abstract, and then assessed at the full-text level. Two reviewers screened independently; disagreements were resolved by discussion and, when required, third-reviewer adjudication. Automation support was limited to duplicate detection and prioritization; humans confirmed inclusion/exclusion decisions.

Data extraction used a piloted form completed independently by two reviewers. Extracted outcomes covered predictive metrics under defects/drift (e.g., area under the curve, accuracy, F1, calibration error, log loss) and operational indicators (false alert rate, detection delay, stability of ranking/pricing decisions), alongside defect induction method (synthetic perturbation vs. natural drift), evaluation protocol (temporal validation and leakage controls), pipeline stage, setting, and reported effects of quality controls (validation checks, monitoring, data contracts, lineage). When summary statistics were incomplete, pre-specified reconstructions from reported baselines were applied when feasible.

Two reviewers assessed the risk of bias. Prediction-model studies were evaluated using prediction model risk of bias assessment tool domains, supplemented by ML-specific checks for leakage, temporal validity, and drift handling; engineering pipeline reports were assessed using an adapted credibility checklist emphasizing defect definition clarity, perturbation/drift characterization, transparency of evaluation procedures, and adequacy of baselines/controls.

For synthesis, the primary quantitative effect was defined as relative performance change under degraded versus reference data conditions, standardized in direction; when transformation was not supported, findings were retained for structured qualitative synthesis. Studies were mapped into synthesis families aligned to defect categories and pipeline stages; summary tables and an evidence map were used to connect defect patterns to decision-relevant failure modes.

Synthesis of Findings

The analytical synthesis of the selected sources reveals a consistent pattern: Predictive algorithms used in product management amplify data quality faults originating in source systems, event tracking, experimentation platforms, and manual data entry. When quality defects go unobserved or unmanaged, model improvements and sophisticated experimentation schemes yield diminishing returns, while organizational trust in data erodes.

The synthesis of the reviewed sources reveals three recurring system-level mechanisms through which data quality affects predictive algorithms used in product management: amplification of upstream data defects, delayed detection of quality drift, and erosion of organizational trust in predictive outputs.

Mohammed *et al*.[3] provide the most explicit quantitative evidence about this pattern. Their large-scale benchmark perturbs six traditional data quality dimensions – such as accuracy, completeness, and consistency – across multiple datasets and model classes and then measures the resulting change in predictive performance. They report that performance degradation under controlled data pollution depends on both the dimension of corruption and the stage of the pipeline where pollution occurs, with polluted serving data sometimes harming outcomes more severely than polluted training data. For product management, the finding means that decisions based on real-time dashboards and online inference can be flawed even when the offline training data appears relatively clean.

At the same time, Nugroho's review[5] documents how missing values, inconsistent identifiers, erroneous timestamps, and misaligned units contribute to systematic bias in predictive analytics. Cases include churn models that silently drop customers with partial histories, uplift models that mix control and treatment events, and marketing attribution pipelines that conflate internal and external identifiers. For product teams, such defects translate into misestimated customer segments, misleading estimates of incremental value, and A/B experiments whose conclusions cannot be replicated once pipelines evolve.

Amudala Puchakayala[2] synthesizes practical strategies for maintaining high data quality across artificial intelligence (AI) projects, including data profiling, standardization, robust ETL patterns, and continuous monitoring. Although the article's primary focus is on engineering practices rather than product teams, it implicitly supports a product-oriented reading. Each recommended practice protects the reliability of the metrics product managers see when reviewing model performance, funnel breakdowns, or customer cohort behavior. For example, systematic handling of missing values and outliers stabilizes product key performance indicators derived from predictive models, reducing the temptation to "explain away" fluctuations as noise or seasonal anomalies instead of structural quality defects.

The emerging field of data-centric AI further shifts attention from model architecture to data quality as the main lever for improvement. Singh[7] reviews data-centric approaches across a broad range of AI and ML studies, highlighting techniques such as targeted relabeling, dataset pruning, synthetic data augmentation under strict constraints, and feedback-driven refinement of data collection. Within product management, these techniques correspond to processes in which teams invest more effort in clarifying event semantics, standardizing tracking schemas across surfaces, obtaining higher-quality labels from operations teams or customers, and prioritizing which segments deserve more precise data collection. Singh's review shows that many gains in model performance come not from "bigger" models but from more disciplined data work, suggesting that product organizations that over-invest in model experimentation and under-invest in data quality optimization risk plateauing quickly.

Sancricca and Cappiello[6] approach the same problem from an engineering pipeline perspective. Their work on "lightweight pipelines" argues that minimal yet thoughtfully designed data pipelines often suffice to provide high-quality input to AI systems, especially when combined with targeted quality checks. In product environments where time-to-market pressures are intense and "rapid, intuition-driven prototyping practices" are popular, this insight reframes pipeline engineering: teams do not need maximal automation and complexity at the outset, but they require clearly defined data contracts, versioning, and validation gates for the data that flows into predictive models. The trade-off identified by Sancricca and Cappiello – between pipeline sophistication and practical maintainability – mirrors the trade-off product leaders navigate between fast experimentation and sustainable quality in their predictive features.

Watson and Larson[8] and Zarour *et al*.[10] embed data quality within broader MLOps discussions. Watson's article describes MLOps as a "factory approach" to ML, where models move through repeatable stages of development, deployment, and monitoring. Data quality checks appear as routine activities rather than ad hoc firefighting. Zarour *et al*. in their systematic review of MLOps best practices and maturity models, show that higher-maturity organizations integrate data quality metrics, lineage tracking, and automated validation into their pipelines as first-class citizens. In such settings, product managers no longer receive static dashboards; instead, they participate in feedback loops in which data quality incidents are treated similarly to production outages. Predictive algorithms in those organizations obtain cleaner training and serving data, which translates into more stable decision support for product work.

Wu and Kästner[9] articulate the consequences of data quality problems for ML-enabled products that operate at scale. Their textbook chapter on data quality emphasizes that poor data quality not only reduces accuracy but also introduces subtle feedback loops: Dire predictions can produce user behavior that yields even lower-quality data, further degrading subsequent models. Product teams that rely on predictive algorithms for personalization, content ranking, or customer risk scoring stand at the center of such loops. Decisions about where to intervene – for example, by adding human-in-the-loop review or by adjusting exploration-exploitation strategies – depend heavily on whether teams understand the quality of incoming data and its evolution over time.

Industry-oriented work links these systemic observations back to concrete business outcomes. Msakni *et al*.[4] demonstrate in an automotive quality-control setting that model performance depends strongly on the completeness and accuracy of sensor data from the production line. Their case study shows that high-quality data enables earlier defect detection and reduced scrap. While the domain differs from software as a service (SaaS) product management, the mechanism remains similar: when input signals faithfully describe operational reality, predictive algorithms can detect risky patterns early enough for intervention. In digital products, analogous signals include event streams, transactional data, and user feedback; when polluted or delayed, they lead to late or misguided interventions in product roadmaps and experimentation.

Finally, Alabi[1] brings the perspective back to product managers themselves. His work on ML for product management surveys uses cases such as demand forecasting, churn prediction, and user segmentation, and repeatedly underlines that data quality, governance, and cross-functional collaboration set the boundary conditions for success. Product managers who treat datasets and labels as opaque infrastructure have fewer levers to influence model reliability. Those who actively shape tracking strategies, data contracts with engineering, and

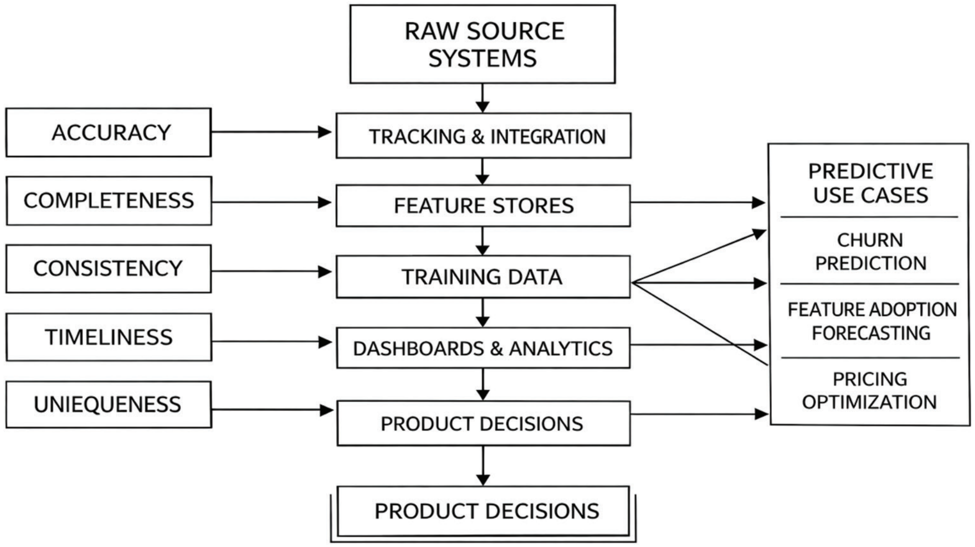

Figure 1: Data quality dimensions and their influence on product-level outcomes (Author's conceptual synthesis based on the reviewed literature)

Table 1: Data quality dimensions and their relevance for predictive product use cases (based on[1,5,7,9])

Data quality dimension	Typical defects in product settings	Impacted predictive use cases in product management
Accuracy	Mis-logged events, incorrect labels, miscalibrated metrics	Churn prediction, conversion models, attribution, pricing elasticity estimation
Completeness	Missing events, partial histories, and absent attributes	Customer lifetime value, segmentation, demand forecasting, experimentation analysis
Consistency	Conflicting schemas, divergent definitions across systems	Cross-channel analytics, omni-channel product journeys, cohort analyses
Timeliness	Latency in ingestion, backfills, and delayed feature computation	Real-time recommendations, alerting, and operational dashboards
Validity	Schema drift, unexpected values, unit mismatches	Feature engineering for ML, anomaly detection, and KPI calculation
Uniqueness	Duplicated users or events, identity resolution failures	Active user metrics, funnel conversion, experimentation, and traffic allocation

ML: Machine learning, KPI: Key performance indicator

Table 2: Organizational practices for sustaining data quality in predictive product environments (based on[1,2,6,10])

Organizational practice	Description in the literature	Interpretation for product management
Data contracts and schema governance	Formal specifications of fields, types, and semantics for shared datasets	Treat event schemas and feature stores as versioned product APIs
Continuous data quality monitoring	Automated checks, alerts, and dashboards for quality metrics	Monitor quality alongside product KPIs on PM dashboards
Data-centric improvement cycles	Iterative refinement of datasets, labels, and collection processes	Allocate development prioritization resources to improve tracking and labels
Integrated MLOps pipelines	End-to-end automation with validation at each stage	Align feature launches with pipeline readiness and quality gates
Cross-functional data stewardship	Shared accountability between engineering, data, and product	Nominate product-side owners for critical datasets and metrics

MLOps: Machine learning operations

governance forums for data issues can directly affect the quality of predictive inputs and, consequently, the trustworthiness of model-driven decisions.

The reviewed literature consistently converges on a data-centric conclusion: models tend to be more robust and transferable when teams prioritize data quality as a primary objective, rather than treating it as a one-off cleaning exercise. The main data quality dimensions that recur across predictive product settings are summarized in Table 1, while the organizational measures used to sustain quality across predictive pipelines are synthesized in Table 2. Targeted relabeling, curating "golden" datasets, designing pipelines around data contracts, and aligning governance forums with product decisions are repeatedly highlighted as high-leverage moves.[2,6,7,9,10] The conceptual relationship between source data quality dimensions, pipeline controls, predictive behavior, and product-level outcomes is presented in Figure 1. When product organizations adopt such practices, predictive algorithms become not only more accurate in offline benchmarks but more reliable in guiding high-impact product choices, from AI feature launches to pricing experiments and roadmap bets.

DISCUSSION

The reviewed evidence supports a practical claim: predictive performance in product environments is bounded more tightly by upstream measurement integrity than by incremental modeling sophistication.[14,16-18] When event semantics are weak, identifiers are unstable, timestamps drift, or instrumentation changes without version control, improvements in architecture and tuning tend to yield fragile gains that fail during deployment, A/B testing, or real-time decisioning.[12-15]

Findings align across domains that differ in data velocity (product analytics versus telemetry): Failures concentrate at interface points – collection and labeling, ingestion/ETL, feature computation, and serving – where silent corruption is most likely and where organizational checks are often weakest.[11-15] This explains why teams experience "metric confidence collapse" after repeated inconsistencies: The system continues producing outputs, yet decision feedback becomes erratic because the mapping from reality to data has degraded.[11,19,20]

The synthesis implies a data-first operating model for predictive product work: treat quality regressions, schema evolution, and drift signals as operational incidents; couple monitoring with clear ownership and rollback levers; and prioritize semantic correctness (definitions, units, identity resolution, time alignment) before expanding model complexity. Under this discipline, model iteration becomes cumulative rather than cyclic, and experimentation readouts remain interpretable under change.[13,14,17-20]

Limitations and Future Research

The work synthesizes published evidence and proposes a conceptual integration; it does not add new empirical validation. The framework was not tested through controlled experiments or longitudinal field studies, and the emphasis on digital/SaaS product environments constrains its transfer to other domains. Future studies can validate the model through multi-site case studies, controlled simulations of quality degradation, and practitioner surveys spanning product, data, and engineering teams.

CONCLUSION

Predictive decision support in product-oriented systems depends on disciplined source data quality across the full lifecycle of data collection, transformation, model training, and serving. The review systematized the main quality dimensions relevant to predictive product work, synthesized recent evidence on the sensitivity of models to defects and drift, and integrated these findings into a unified framework linking data-quality practices, MLOps controls, and product workflows. The analysis shows that failures in identifiers, timestamps, schemas, labels, and event completeness reduce decision stability, distort experimental outcomes, and weaken organizational trust, even when the model architecture appears technically sound. Reliable predictive use in practice, therefore, requires continuous monitoring, explicit data contracts, governance routines, and cross-functional ownership of critical datasets so that model outputs remain interpretable, operationally stable, and suitable for strategic and day-to-day decisions.

AVAILABILITY OF DATA AND MATERIALS

Data sharing does not apply to this article as no datasets were generated or analyzed during the current study.

COMPETING INTERESTS

EL declares that she has no competing interests.

FUNDING

No external funding was received for this work.

AUTHORS' CONTRIBUTIONS

EL conceived the study, performed the literature synthesis, developed the conceptual framework, and wrote the manuscript.

ACKNOWLEDGMENT

Not applicable.

AUTHORS' INFORMATION

EL is a product leader specializing in predictive decision flows within data-intensive information systems.

REFERENCES

1. Alabi M. Machine Learning for Product Management: Unlocking Insights from Data. ResearchGate Preprint; 2023. Available from: https://www.researchgate.net/publication/384729366_machine_learning_for_product_management_unlocking_insights_from_data [Last accessed on 2025 Dec 24].
2. Amudala Puchakayala PR. Data quality management for effective machine learning and AI modelling, best practices and emerging trends. Int Res J Innov Eng Technol 2022;6:327-40.
3. Mohammed S, Budach L, Feuerpfeil M, Ihde N, Nathansen A, Noack N, *et al*. The effects of data quality on machine learning performance on tabular data. Inf Syst 2025;132:102549.
4. Msakni MK, Risan A, Schütz P. Using machine learning prediction models for quality control: A case study from the automotive industry. Comput Manag Sci 2023;20:14.
5. Nugroho H. A review: Data quality problem in predictive analytics. Int J Appl Inf Technol 2023;7:79-91.
6. Sancricca C, Cappiello C. Lightweight Pipelines: Good Enough is Sometimes Better. In: Proceedings of the 2nd International Workshop on Data-Centric AI (DATAI 2025). VLDB Workshops; 2025.
7. Singh P. Systematic review of data-centric approaches in artificial intelligence and machine learning. Data Sci Manag 2023;6:144-57.
8. Watson HJ, Larson D. MLOps: From a cottage industry to a factory approach. Int J Bus Intell Res 2024;15:1-22.
9. Wu S, Kästner C. Data quality. In: Kästner C, editor. Machine Learning in Production: From Models to Products; 2024. Available from: https://mlip-cmu.github.io/book/16-data-quality.html [Last accessed on 2025 Dec 24].
10. Zarour M, Alzabut H, Al-Sarayreh KT. MLOps best practices, challenges and maturity models: A systematic literature review. Inf Softw Technol 2025;183:107733.
11. Sambasivan N, Kapania S, Highfill H, Akrong D, Paritosh PK, Aroyo L. Everyone Wants to do the Model Work, not the Data Work: Data Cascades in High-Stakes AI. In: CHI Conference on Human Factors in Computing Systems (CHI '21). Association for Computing Machinery; 2021. p. 1-15.
12. Grafberger S, Guha S, Stoyanovich J, Schelter S. MLINSPECT: A Data Distribution Debugger for Machine Learning Pipelines. In: Proceedings of the 2021 International Conference on Management of Data (SIGMOD '21). Association for Computing Machinery; 2021. p. 2736-9.
13. Grafberger S, Groth P, Stoyanovich J, Schelter S. Data distribution debugging in machine learning pipelines. VLDB J 2022;31:1103-26.
14. Priestley M, O'Donnell F, Simperl E. A survey of data quality requirements that matter in ML development pipelines. ACM J Data Inf Qual 2023;15:11.
15. Steidl M, Felderer M, Ramler R. The pipeline for the continuous development of artificial intelligence models-Current state of research and practice. J Syst Softw 2023;199:111615.
16. Gong Y, Liu G, Xue Y, Li R, Meng L. A survey on dataset quality in machine learning. Inf Softw Technol 2023;162:107268.
17. Jakubik J, Vössing M, Kühl N, Walk J, Satzger G. Data-centric artificial intelligence. Bus Inf Syst Eng 2024;66:507-15.
18. Zha D, Bhat ZP, Lai KH, Yang F, Jiang Z, Zhong S, *et al*. Data-centric artificial intelligence: A survey. ACM Comput Surv 2025;57:1-42.
19. Bernardo BMV, São Mamede H, Barroso J, Santos V. Data governance and quality management-Innovation and breakthroughs across different fields. J Innov Knowl 2024;9:100598.
20. Batool A, Zowghi D, Bano M. AI governance: A systematic literature review. AI Ethics 2025;5:3265-79.