



Data Lineage as the New GxP Frontier

WHITE PAPER

Why Traceable Is Becoming ALCOA's Most Important Letter

Abstract

Data integrity has driven the majority of FDA drug GMP warning letters for years, and the underlying regulatory expectation keeps getting more demanding. The original ALCOA framework, requiring data to be attributable, legible, contemporaneous, original, and accurate, has expanded into ALCOA+, adding completeness, consistency, endurance, and availability, and the most recent guidance adds a further letter: traceable. That single addition reframes data integrity from a set of static record-keeping properties into a continuous requirement that every data point be traceable forward and backward, from its origin through every change to its final report. For organizations that have spent years satisfying ALCOA+ through point-in-time controls, this is a meaningfully harder standard, and one that AI-assisted data pipelines make both more urgent and more achievable. This white paper applies established data reconciliation and traceability best practices, already proven across fragmented compliance data sources, to the discipline of data lineage: knowing not just that a data point is accurate today, but exactly where it came from, what touched it, and how it changed along the way.

A Well-Known Problem That Keeps Getting Worse

Data integrity is not a new area of FDA scrutiny. It has been a dominant theme in drug GMP enforcement for over a decade, and the numbers have not moved in a reassuring direction. Roughly 60 to 80% of FDA drug GMP warning letters cite data integrity deficiencies as a primary or contributing violation, and in fiscal year 2024 alone the FDA issued 105 warning letters for human drug quality issues, an 11% increase over the prior year.ⁱⁱ

The most common root causes are familiar to anyone who has worked through a data integrity remediation: shared login credentials that make individual attribution impossible, audit trail functionality that was disabled or never enabled, backdated or falsified records, quality systems running on unvalidated spreadsheets, and improper electronic signatures. None of these are exotic failure modes. They are basic control gaps that keep recurring across the industry.

What has changed is the standard organizations are being held to. Data integrity used to be evaluated largely at the level of individual records: was this specific entry attributable, accurate, contemporaneous. Increasingly, regulators are asking a harder question: can you trace this data point's entire life cycle, from the instrument or system that generated it through every transformation, aggregation, and report it eventually informed.

ALCOA+ Gets a New Letter, and a Much Bigger Job

The original ALCOA principles, attributable, legible, contemporaneous, original, and accurate, were extended years ago into ALCOA+ with the addition of complete, consistent, enduring, and available. The most recent guidance goes further, adding traceable as an explicit, distinct requirement.ⁱ

Traceability overlaps with attributability but demands considerably more: a documented chain of custody for data as it moves between systems, and automated logging of every modification a data point undergoes, not just its final state. In practice, this means an organization needs to be able to answer a question like where did this batch record's yield figure originate, what system calculated it, what values fed that calculation, and has anything in that chain changed since, not as a forensic reconstruction project during an inspection, but as a standing, queryable capability.

This is a materially different engineering problem than point-in-time data integrity. A record can be accurate and attributable while still living inside a system where its lineage, everything that produced it and touched it, is effectively invisible without a dedicated reconstruction effort.

Applying Best Practices: A Data Lineage Practice for GxP Systems

A foundational best practice in data governance transfers directly to this problem: maintaining one reconciled, traceable record of data as it moves across multiple systems and sources. Data lineage for GxP data is the same underlying discipline, applied to the full life cycle of a manufacturing, quality, or clinical data point rather than to spend or transparency data.

1. **A Data Origin Map for Critical Data Elements.** Identify the specific data elements that matter most for regulatory purposes, batch release values, stability results, critical process parameters, and document exactly which system originates each one.



2. **Automated Lineage Capture, Not Retrospective Reconstruction.** Build lineage tracking into the systems and integrations themselves, so the chain of custody is captured as data moves rather than reconstructed after the fact during an inspection.
3. **Lineage Visibility Across System Boundaries.** Extend lineage tracking through every system a critical data element passes through, including the points where data moves between validated systems and spreadsheets, LIMS, MES, and reporting layers.
4. **AI Pipeline Traceability as a First-Class Requirement.** Where AI or automated tools touch regulated data, whether flagging anomalies, transforming values, or generating summaries, require those systems to produce their own audit trail of how they processed the data, not just their output.

None of this requires replacing validated systems that already work. It requires building the reconciliation and traceability layer already applied to spend and reporting data, extended to the specific problem of proving where a regulated data point came from and everything that happened to it along the way.

What a Data Lineage Practice Buys, and Where It Gets Hard

Advantages

- **Inspection readiness that does not require reconstruction:** when lineage is captured continuously, answering an investigator's origin question is a query, not a project.
- **Faster root cause investigations:** a queryable lineage record dramatically shortens the time needed to trace a data anomaly back to its source.
- **A defensible foundation for AI adoption in regulated processes:** traceable AI pipelines are easier to validate and defend than ones whose internal processing is a black box.
- **Fewer data integrity findings tied to reconstruction gaps:** many data integrity citations stem not from bad data but from an inability to prove where good data came from.

Risks and Barriers

- **Legacy systems were not built for this:** many validated systems still in active use were designed around point-in-time accuracy, not continuous lineage capture, and retrofitting them is real engineering work.
 - **Spreadsheets remain a persistent gap:** critical calculations still living in unvalidated spreadsheets are functionally lineage dead ends that require remediation, not just documentation.
 - **Lineage tooling adds its own validation burden:** any system built to track lineage is itself a GxP system that needs validation and its own audit trail.
 - **Cross-functional ownership is required:** critical data elements span manufacturing, quality, and IT, and building one lineage map requires agreement across functions that do not always share priorities.
-

Case in Point: The Yield Figure Nobody Could Fully Explain

A mid-size manufacturer preparing for a pre-approval inspection ran an internal audit of its batch record data and selected a routine yield calculation to trace end to end as a test case. The final figure in the batch record was accurate. It matched the underlying weights and had passed quality review. What the audit team could not quickly



reconstruct was the full path: the calculation had originated in the MES, been exported to a spreadsheet for a secondary check that had become informal standard practice, and the spreadsheet's output, not the MES's own value, was what had actually been transcribed into the final record.

No data had been falsified and no result was wrong, but the actual lineage of the number in the batch record did not match what the validated system of record implied. Reconstructing that path took the audit team the better part of a week, pulling export logs, interviewing the analysts involved, and confirming the spreadsheet's calculation matched the MES logic. Had an FDA investigator asked the same question during an inspection, a week was not an available answer.

The manufacturer's response was not to ban the spreadsheet check outright but to bring it inside the lineage map: the spreadsheet was validated as a documented step in the calculation chain, its inputs and outputs were logged automatically, and the batch record now captured which system's value was actually used. The next internal audit of the same data element took minutes, not a week.

Recommendations for Quality and Data Governance Leaders

1. **Identify your organization's critical data elements first.** rather than attempting to map lineage for every data point in every system simultaneously.
2. **Build lineage capture into system integrations.** so the chain of custody is generated automatically as data moves rather than reconstructed after the fact.
3. **Bring informal spreadsheet calculations inside the validated boundary.** either by validating them as documented process steps or by eliminating the informal step entirely.
4. **Require audit trails from any AI or automated tool touching regulated data.** as a condition of deployment, not as a retrofit after go-live.
5. **Run a lineage reconstruction test on a real data element.** before an inspection does it for you, to find where your organization's actual gaps are.
6. **Assign clear cross-functional ownership of the lineage map.** spanning quality, manufacturing, and IT, so it survives beyond the initial mapping project.

Measuring Success: KPIs for Data Lineage Maturity

- **Critical data element lineage coverage:** the percentage of identified critical data elements with a documented, automated lineage map, tracked toward completeness.
- **Lineage reconstruction time:** how long it takes to trace a given data point's full origin and change history, tracked toward minutes rather than days.
- **Unvalidated calculation touchpoints remaining:** the count of informal spreadsheet or manual calculation steps still sitting inside a critical data element's lineage path.
- **AI pipeline audit trail coverage:** the percentage of AI or automated tools touching regulated data with their own documented, reviewable processing trail.



Conclusion

Data integrity enforcement has not softened, and the standard it is measured against keeps expanding. ALCOA+ adding traceable as an explicit requirement is not a minor guidance update. It asks organizations to prove not just that a data point is correct today, but where it came from and everything that happened to it along the way, on demand.

Applying disciplined best practices to a standing data lineage practice turns that requirement from a reconstruction exercise into a query. Organizations that build this capability now are not just reducing data integrity risk. They are building the traceability foundation that AI-assisted quality systems will require as they become a standard part of GxP operations.

About OmniTek & Contact Information

OmniTek Consulting is a trusted partner for organizations seeking to enhance governance, operational agility, and decision-making. We specialize in modernizing business processes with a blend of human-centered design, technology expertise, and hands-on change management.

For quality and data governance leaders facing the expanding demands of ALCOA+ traceability, OmniTek applies proven best practices to build a practical data lineage practice: a critical data element map, automated lineage capture built into system integrations, and audit trail standards that extend to AI-assisted processing. Whether you are preparing for your next inspection or building lineage capability ahead of AI adoption in regulated processes, OmniTek can help you build a data governance foundation that holds up under direct scrutiny.

Contact us today to explore how OmniTek can support your transformation journey and drive lasting results.



McLean, VA | Serving
clients nationwide



info@omnitekconsulting.com



omnitekconsulting.com

[Click here to book a 15-minute call to see how OmniTek can help your team put these insights into practice.](#)

Endnotes

ⁱ IntuitionLabs, "ALCOA+ Principles: A Guide to GxP Data Integrity," on the ALCOA++ Traceable principle and FDA/CDER warning letter trends, available at <https://intuitionlabs.ai/articles/alcoa-plus-gxp-data-integrity>

ⁱⁱ Klyverity, "FDA Warning Letters and Data Integrity: Lessons from Real Enforcement Actions," citing FY2024 warning letter volume and common data integrity root causes, available at <https://klyverity.com/blog/fda-warning-letters-data-integrity>

