

Data Cleaning in Clinical Trials: Process & Best Practices

By Adrien Laurent, CEO at IntuitionLabs • 11/17/2025 • 40 min read

- data cleaning
- clinical trials
- clinical data management
- gcp
- data quality
- edc systems
- discrepancy management
- 21 cfr part 11



[Revised April 1, 2026]

Executive Summary

Data cleaning in clinical trials is a fundamental component of clinical data management (CDM) and a cornerstone of research integrity. This comprehensive report details the methodologies, workflows, and best practices by which clinical data managers identify and correct errors in trial data. We examine the historical evolution of data management—from paper-based case report forms (CRFs) to modern electronic data capture (EDC) systems—and highlight the regulatory and technological context driving data-cleaning activities. Modern trials demand high-quality, reliable data, as flawed data can lead to invalid conclusions, regulatory setbacks, or compromised patient safety. Indeed, studies have reported data error rates ranging from as low as 0.14% with double data entry up to over 6% with more manual methods ⁽¹⁾ ([pmc.ncbi.nlm.nih.gov](https://pubmed.ncbi.nlm.nih.gov)), underscoring the risks of inaccurate data.

The report covers the full scope of data cleaning in clinical research: we define core concepts, differentiate data cleaning from central monitoring, and outline a detailed data-cleaning workflow. We discuss the roles and responsibilities of data managers, highlighting techniques such as edit checks, range checks, consistency checks, duplicate detection, and discrepancy management. Advanced tools—ranging from SAS and R scripts to specialized [CDISC-compliant EDC/CTMS](#) platforms—are described. Case studies and empirical findings illustrate the impact of robust data cleaning: for example, introduction of real-time validation in the ASPREE trial dramatically reduced data-entry errors from 0.3% to 0.01% ⁽²⁾ (trialsjournal.biomedcentral.com). Supporting evidence is drawn from peer-reviewed literature, regulatory guidelines, and industry reports.

We analyze quantitative findings (e.g. error rates and their effects on statistical power) and present evidence-based arguments. Tables summarize key processes, such as common types of data errors and their remediation strategies, and comparative error rates by data-entry method. Perspectives from industry, academia, and regulatory frameworks are included, along with historical context and discussion of future directions (including automation and artificial intelligence). The conclusion reinforces that meticulous, continuous data cleaning is essential to ensure trial validity, minimize costs, and ultimately protect patient safety. All assertions are supported by citations to authoritative sources.

Introduction and Background

Clinical trials rely on the integrity of collected data to assess the safety and efficacy of interventions. By definition, *good data management practices* involve ensuring that clinical data are “collected, protected, cleaned, and managed” according to [regulatory standards \(e.g. 21 CFR Part 11\)](#) ⁽³⁾ ([pmc.ncbi.nlm.nih.gov](https://pubmed.ncbi.nlm.nih.gov)). Data cleaning—correcting errors and inconsistencies in trial data—is a critical part of this process. Its importance stems from the fact that **poor data quality undermines confidence in trial results** ⁽³⁾ ([pmc.ncbi.nlm.nih.gov](https://pubmed.ncbi.nlm.nih.gov)). As Oronsky *et al.* note, high-quality data are the “fuel” of [drug development](#) engines; conversely, datasets with systematic or random errors can threaten both patient safety and the validity of conclusions ⁽³⁾ ([pmc.ncbi.nlm.nih.gov](https://pubmed.ncbi.nlm.nih.gov)) ⁽⁴⁾ ([pmc.ncbi.nlm.nih.gov](https://pubmed.ncbi.nlm.nih.gov)).

The phrase “data cleaning” (sometimes called data editing or data management) has evolved over time. Historically, early clinical trials recorded information on paper or punch cards. In those days, it was common to enter data only after trial completion and perform large-scale cleaning just before analysis ⁽⁵⁾ ([pmc.ncbi.nlm.nih.gov](https://pubmed.ncbi.nlm.nih.gov)). With the advent of computers and digital CRFs, the concept of cleaning emerged to correct transcription and consistency errors (e.g. via double data entry to uncover typos) ⁽⁵⁾ ([pmc.ncbi.nlm.nih.gov](https://pubmed.ncbi.nlm.nih.gov)) ⁽¹¹⁾ ([pmc.ncbi.nlm.nih.gov](https://pubmed.ncbi.nlm.nih.gov)). Over time, real-time electronic data capture (EDC) allowed sites to enter data directly into databases, enabling “front-loaded” quality checks. In modern trials, data are often subject to immediate edit checks and range checks as they are entered into EDC systems, reducing some errors at the point of entry ⁽⁶⁾ ([pmc.ncbi.nlm.nih.gov](https://pubmed.ncbi.nlm.nih.gov)) ⁽⁷⁾ ([pmc.ncbi.nlm.nih.gov](https://pubmed.ncbi.nlm.nih.gov)). Nonetheless, additional centralized cleaning is still necessary, as many types of errors only become apparent after data aggregation (for example, logical inconsistencies or duplicate forms) ⁽⁶⁾ ([pmc.ncbi.nlm.nih.gov](https://pubmed.ncbi.nlm.nih.gov)) ⁽⁸⁾ (www.ncbi.nlm.nih.gov).

This report reviews *how* data managers currently find and fix errors in clinical trial data, based on historical practices, current standards, and empirical evidence. We cover definitions (e.g. of terms like “data cleaning”, “discrepancy”, “database lock”); discuss regulatory and technical environments; detail common error types and cleaning techniques; and present real-world examples and metrics. Finally, we explore emerging trends (such as automation and [artificial intelligence](#)) that may shape the future of data cleaning. Each section draws on scientific literature, guidelines, and case studies, with extensive citations to support the analysis.

1. Context and Definitions

Clinical data management (CDM) is a multi-step process encompassing data collection, entry, cleaning, and analysis. It is foundational to [good clinical practice \(GCP\)](#) and regulated by guidelines (e.g. ICH E6(R3), finalized in January 2025, and FDA’s 21 CFR Part 11). Data quality, broadly defined as data “free of deficiencies” and “support [ing] the same conclusions as error-free data,” is paramount ⁽⁹⁾ [pmc.ncbi.nlm.nih.gov](#)). Within CDM, **data cleaning** has been defined as “the identification and correction of errors in the data prior to analysis” ⁽¹⁰⁾ [pmc.ncbi.nlm.nih.gov](#)). This definition underscores that data cleaning is explicitly post-collection and pre-analysis, focusing on removing noise (errors, duplicates, inconsistencies) without deleting the underlying signal ⁽¹¹⁾ [pmc.ncbi.nlm.nih.gov](#)).

It is important to distinguish data cleaning from monitoring activities. As Love *et al.* explain, “data cleaning” and “central monitoring” are conceptually distinct: data cleaning is applied to the data elements themselves (e.g. correcting a date that was entered outside an allowable range), whereas monitoring (whether on-site or centralized) refers to the oversight processes ensuring compliance with protocol and GCP ⁽¹²⁾ [pmc.ncbi.nlm.nih.gov](#)). Historically, 100% source data verification (SDV) was the standard approach to assure data accuracy, but this has given way to risk-based and centralized monitoring approaches that may rely more on reviewing aggregated data for anomalies. Nevertheless, even with advanced monitoring strategies, the specific tasks of finding and correcting data errors typically fall to the data management team.

Another related term is **discrepancy management** (or query resolution). Here, any value that “deviates from what is expected”—be it an out-of-range value, a missing required field, or an inconsistency—is flagged as a query or discrepancy ⁽¹³⁾ [pmc.ncbi.nlm.nih.gov](#)). Resolving reminders (queries) with the site or investigators is essentially part of data cleaning, sometimes termed “cleaning the data on the CRFs” ⁽¹⁴⁾ [pmc.ncbi.nlm.nih.gov](#)). In practice, data cleaning often proceeds as an iterative cycle: automated checks flag issues, data managers review and classify them, and then queries are issued to sites to clarify or correct the values. This loop continues until data quality reaches target levels, after which the database is “locked” for analysis ⁽⁸⁾ [www.ncbi.nlm.nih.gov](#)).

As a practical note, data cleaning is not perceived neutrally. Oronsky *et al.* remark that “data cleaning often carries a negative connotation, as if it were a euphemism for post hoc data manipulation” ⁽¹⁰⁾ [pmc.ncbi.nlm.nih.gov](#)). However, this stigma is misplaced. Data cleaning is an essential *quality assurance* step: it aims to remove extraneous “noise” (duplicates, miscoded entries, etc.) while preserving legitimate variation (the true signal). The success of cleaning is evidenced by an improved “signal-to-noise” ratio in the dataset ⁽¹¹⁾ [pmc.ncbi.nlm.nih.gov](#)).

Definition (Data Cleaning): *The process of identifying, investigating, and correcting or annotating errors and inconsistencies in clinical trial data, typically through programmed checks, queries, and manual review, in order to prepare a high-quality dataset for statistical analysis* ⁽¹⁰⁾ [pmc.ncbi.nlm.nih.gov](#) ⁽⁸⁾ [www.ncbi.nlm.nih.gov](#)).

Key quality attributes include accuracy, completeness, validity, and consistency. For example, Shaheen *et al.* (2019) emphasize domains such as completeness (no missing data), accuracy (values reflect true observations), validity (values conform to format/rules), and consistency (logical coherence across variables) ⁽¹⁵⁾ [pmc.ncbi.nlm.nih.gov](#)). Data cleaning addresses each of these: missing values are queried or imputed, formatting errors are corrected, and logical inconsistencies are resolved by returning to source documents.

2. Historical Evolution of Clinical Data Management

The practice of data management in clinical trials has evolved significantly since the mid-20th century. Early trials relied on paper case report forms (pCRFs) or even punch cards, with data entry often occurring late in the study. Data cleaning then was typically a batch process: data were double-entered or verified, and any discrepancies were reconciled before final analysis. For instance, in the CHART radiotherapy trials (circa 1994), the necessity of double data entry was examined as a means to reduce errors (^[16] [pmc.ncbi.nlm.nih.gov](#)).

With the proliferation of computers, trialists gradually shifted to centralized data capture. A landmark shift occurred with the introduction of Electronic Data Capture (EDC) systems. These platforms allowed data to be entered at site level directly into a central database, often with built-in edit checks for immediate validation. The Love *et al.* (2021) letter notes that as EDC supplanted paper forms, site staff began entering data in real time with automated prompts, and central trial teams performed further cleaning using range and logic checks (^[6] [pmc.ncbi.nlm.nih.gov](#)). This “real-time” data cleaning was a major advance: problems could be detected and addressed while the trial was ongoing, greatly reducing the post-hoc cleaning effort.

In parallel, the role of the Data Manager has expanded. In earlier eras, data managers primarily oversaw manual entry and verification tasks. Today, they design electronic CRFs (eCRFs), implement edit checks in EDC systems, and manage queries through specialized software. Technological innovations like Clinical Data Management Systems (CDMS) (e.g. Oracle Clinical, Medidata Rave) and analytics platforms have become standard tools for data managers (^[17] [pmc.ncbi.nlm.nih.gov](#)).

However, despite automation, new complexities have emerged. Modern trials often collect vast amounts of data from disparate sources: EDC systems, lab and biomarker vendors, imaging, patient-reported outcomes (ePRO), wearable devices, and sometimes direct EHR integration (^[18] [pmc.ncbi.nlm.nih.gov](#)) (^[19] [pmc.ncbi.nlm.nih.gov](#)). Integrating and cleaning across these sources can be daunting, as illustrated by Farnum *et al.* (2019): assembling an “error-free” dataset may involve data from dozens of disparate systems, historically managed through disjointed processes that were “resource intensive and error prone” (^[20] [pmc.ncbi.nlm.nih.gov](#)). In response, some sponsors have adopted unified data warehousing solutions (e.g. Oracle Data Management Workbench or Xcellerate), which centralize data and cleaning workflows (^[21] [pmc.ncbi.nlm.nih.gov](#)) (^[22] [pmc.ncbi.nlm.nih.gov](#)).

In summary, the historical trajectory moved from post-hoc cleaning of late-entered data to continuous, integrated cleaning. Yet error still “creeps in and propagates”, as one review observed—no amount of process safeguards can eliminate mistakes entirely (^[23] [pmc.ncbi.nlm.nih.gov](#)). Thus, modern data cleaning builds on a long tradition: leveraging new tools while reinforcing foundational practices like double data entry and audit trails.

3. Regulatory Framework and Standards

Clinical data cleaning is conducted within a strict regulatory environment. The cornerstone guideline is ICH E6, which governs Good Clinical Practice. **ICH E6(R3), finalized on January 6, 2025**, replaced the prior E6(R2) version that had been in effect since 2016. The FDA issued its final guidance adopting E6(R3) on September 8, 2025, while the EMA’s adoption became effective July 23, 2025, and Health Canada’s implementation date was April 1, 2026 (^[24] [about.citiprogram.org](#)). E6(R3) introduces a quality-by-design approach and risk-based quality management, with technology-neutral language that supports digital tools and decentralized trial designs. Annex 2, which will specifically address decentralized and pragmatic trials, underwent public consultation from November 2024 through February 2025 and is expected to be finalized in mid-2026 (^[25] [acrpnnet.org](#)). While no guideline explicitly prescribes *how* to clean data, they set requirements that implicitly necessitate thorough cleaning practices. ICH E6(R3) continues to mandate trial

monitoring and data verification by sponsors to ensure data accuracy (^[26] ichgcp.net). GCP also requires that all trial data be attributable, legible, contemporaneous, original, and accurate (ALCOA) (^[27] [pmc.ncbi.nlm.nih.gov](https://pubmed.ncbi.nlm.nih.gov)). In practice, meeting these standards means data managers must document all changes via audit trails and must correct any deviations from ALCOA principles.

FDA's 21 CFR Part 11, which applies to electronic records, demands that systems are validated and that changes are tracked. A significant related development came in **September 2025**, when the FDA issued its **final Computer Software Assurance (CSA) guidance**, superseding the decades-old "General Principles of Software Validation." The CSA guidance endorses risk-based approaches to validation—including unscripted and exploratory testing, continuous monitoring, and leveraging supplier evidence—and introduces new definitions for cloud computing, SaaS, IaaS, and PaaS (^[28] [federalregister.gov](https://www.federalregister.gov)). This modernizes the compliance landscape for EDC systems and data cleaning platforms. The NCBI Field Trials guidance stresses validating and testing every data system to comply with GCP (^[29] www.ncbi.nlm.nih.gov). Once data cleaning is complete and the database is locked, the rules require that no further changes occur before analysis (^[8] www.ncbi.nlm.nih.gov). The locked database is, in effect, a "screenshot" of the final dataset, which regulators may inspect or reference during auditing.

In addition, the FDA issued its first draft guidance on AI in drug development in **January 2025**—"Considerations for the Use of Artificial Intelligence to Support Regulatory Decision-Making for Drug and Biological Products"—introducing a 7-step credibility assessment framework for AI models used in submissions. In **January 2026**, the FDA and EMA jointly released "Guiding Principles of Good AI Practice in Drug Development," further shaping how AI-assisted data cleaning tools may be evaluated in a regulatory context (^[30] [aihnet.com](https://www.aihnet.com)).

Sponsors also follow industry guidelines, such as the Society for Clinical Data Management (SCDM) Good Clinical Data Management Practices (GCDMP), which define error rate calculations and best practices (^[31] [pmc.ncbi.nlm.nih.gov](https://pubmed.ncbi.nlm.nih.gov)) (^[9] [pmc.ncbi.nlm.nih.gov](https://pubmed.ncbi.nlm.nih.gov)). Notably, the SCDM opened its updated GCDMP chapter on "Clinical Trial Database Lock" for public review through March 2026, reflecting continued refinement of these foundational practices (^[32] [scdm.org](https://www.scdm.org)). The SCDM has also launched a beta test for the new **Certified Clinical Data Scientist (CCDS)** credential in 2025, with formal launch expected in 2026, signaling the growing importance of data science skills in clinical data management (^[33] [scdm.org](https://www.scdm.org)). These references emphasize that error counting methods must be consistent and that error rates should ideally be very low (approaching 0% in critical variables) (^[31] [pmc.ncbi.nlm.nih.gov](https://pubmed.ncbi.nlm.nih.gov)). The House of Commons guidance (ISO 9000, mentioned in [44]) reinforces treating data as a "product" requiring fit-for-purpose quality (^[34] [pmc.ncbi.nlm.nih.gov](https://pubmed.ncbi.nlm.nih.gov)) (^[9] [pmc.ncbi.nlm.nih.gov](https://pubmed.ncbi.nlm.nih.gov)). In essence, multiple layers of regulations and standards converge to ensure a systematic approach to data cleaning in trials.

4. Data Cleaning Workflows and Processes

Data cleaning in a clinical trial is a structured, often iterative process. A typical workflow includes: planning (during study start-up), data entry and validation (during the trial), discrepancy resolution (ongoing and end-of-study), and final database lock. Key components of these processes include: CRF design, edit checks, query management, and database archival.

4.1 Data Management Planning

Good data cleaning begins before the trial starts. The Data Management Plan (DMP) and Data Validation Plan are drafted during study design, detailing procedures for data collection, coding, cleaning, and QC (^[35] www.ncbi.nlm.nih.gov). Investigators design the CRF (paper or electronic) to minimize ambiguity; for example, duplicate fields are avoided or used deliberately and consistently to ease cross-checking (^[36] www.ncbi.nlm.nih.gov). All variables and code lists (especially for adverse events and medications) are defined in annotated CRFs or data dictionaries. As Field Trials

guidance recommends, variable names should be meaningful and consistent across visits, facilitating automated checks later ([37] www.ncbi.nlm.nih.gov).

The team also identifies “key variables”—critical outcomes or eligibility criteria that must be complete and accurate. Automated edit checks (range checks, required fields, consistency checks) are programmed based on the protocol. For example, logical rules might ensure a male subject’s pregnancy status is automatically set to “not pregnant” ([8] www.ncbi.nlm.nih.gov). These checks are validated through self-testing of the database prior to go-live. Crucially, “data cleaning should be an ongoing process, rather than something done at the end of the study,” and the plan for how data will be cleaned should be documented from the outset ([35] www.ncbi.nlm.nih.gov). This forward planning includes scheduling regular quality reviews, defining query types, and training sites on eCRF entry.

Table 1 outlines *common categories of data errors* and typical detection/fix strategies (many of which are decided in the planning stage). This table is not exhaustive but illustrates the breadth of data issues:

Error Type	Example	Detection Methods	Cleaning/Resolution Approach
Missing Data	A required lab result not entered	Edit checks (required-field violations); reports of missing fields	Issue query to site to determine if data exist in source; if truly missing, document reason or impute (per Statistical Analysis Plan) ([2] trialsjournal.biomedcentral.com) ([38] www.ncbi.nlm.nih.gov).
Entry Error (Type)	Date of birth entered as 13/45/1980	Range checks (date out of realistic range); logic checks	Review against source documents, correct via query. If date is clearly wrong (e.g. month=13), query site for correct value ([39] www.ncbi.nlm.nih.gov) ([38] www.ncbi.nlm.nih.gov).
Inconsistency	Mismatch between related fields (e.g., weight inconsistent with height)	Consistency checks across forms; medical logic checks	Investigate discrepancy. Query site for correct measurement or confirm correct form. Update data accordingly.
Out-of-Range / Invalid	Lab value outside normal assay limits	Range checks (system alerts if value outside plausible range)	Confirm whether value is true (possibly extreme result) or data entry error. If error, correct; if true outlier, accept but flagged for analysis.
Duplicate Records	Same subject entered twice (e.g. once per site)	Unique identifiers check; subject tracking lists	Identify duplicates and merge records or coordinate with sites to reconcile.
Coded Data Errors	MedDRA term incorrectly entered or misspelled	Automated coding reviews; manual review of verbatim text	Use standardized medical coding. Recoder verbatim terms and recode, with query if ambiguity.
Skip Pattern Violations	Question answered when preceding condition not met	Edit checks on CRF design (skip logic); manual review	Instruct site to blank answer (if not applicable) or adjust preceding response.
Data Entry System Bug	Values truncated or misaligned	Data audits after migration or system changes	Work with informatics to correct system issue; reimport or fix affected records with audit trail.

Table 1: Common clinical trial data errors and typical cleaning actions (examples compiled from literature) ([11] pmc.ncbi.nlm.nih.gov) ([38] www.ncbi.nlm.nih.gov) ([35] www.ncbi.nlm.nih.gov).

4.2 Data Entry and Immediate QC

Once the trial is underway, data entry (manual or electronic) commences. In a paper-based trial, double data entry (two operators separately transcribe the same CRF) was historically the “gold standard” for minimizing entry errors ([5] pmc.ncbi.nlm.nih.gov) ([39] www.ncbi.nlm.nih.gov). In modern EDC-based trials, site staff enter data directly through electronic CRFs, and the system immediately runs programmed checks. For example, a field may be configured as “required,” preventing save unless filled in, or a range check may flag an out-of-bound value. Immediate edit checks (automatic discrepancy flags at entry) greatly reduce simple entry errors ([6] pmc.ncbi.nlm.nih.gov) ([7] pmc.ncbi.nlm.nih.gov).

Despite these safeguards, errors inevitably slip through. Some errors arise from user misunderstandings or data quirks that checks cannot catch; others come from mis-sync in data transfers (e.g., lab vendors uploading data). Therefore, data managers regularly export data for review. Using statistical software or query management tools, they inspect accumulated records for patterns: missing data trends, outliers, or data consistency issues. Reviewers may produce listings or dashboards that show metrics like query counts per site or percent completeness. Farnum *et al.* illustrate that data undergo “a significant amount of scrutiny” at multiple levels—timeliness, completeness, accuracy, and consistency—

where outputs of programmed checks are reviewed by data managers for suspected discrepancies (^[40] [pmc.ncbi.nlm.nih.gov](https://pubmed.ncbi.nlm.nih.gov)).

When discrepancies are found, they generate queries (often through the EDC's query module) directed to sites. A typical query asks the site to confirm or correct a specific data point (e.g. "Please verify the complete date of birth for Subject 123; month was entered as 13"). The site typically responds by updating the data in the EDC or annotating the CRF. As the Field Trials guide notes, some simple queries (like an obvious date error) can be resolved by the data management team internally, but most require site intervention (^[41] [www.ncbi.nlm.nih.gov](https://pubmed.ncbi.nlm.nih.gov)). The site's answers effectively *clean* that issue – either by correcting the entry or confirming its accuracy. The data manager then marks the query resolved in the system.

It is critical that all data corrections occur in a single **master database** with robust version control. After each batch of queries, data managers maintain an up-to-date master copy with every change tracked (^[42] [www.ncbi.nlm.nih.gov](https://pubmed.ncbi.nlm.nih.gov)). Even anticipated *post hoc* discoveries (e.g. errors found during statistical analysis) must lead to corrections in the master, keeping it "always up to date" (^[42] [www.ncbi.nlm.nih.gov](https://pubmed.ncbi.nlm.nih.gov)). Version control logs who made the change, when, and why, satisfying audit requirements.

4.3 Ongoing Data Cleaning and Discrepancy Management

Data cleaning is not a one-time event but a continuous process throughout the trial. Field Trials guidance emphasizes, "Data cleaning should be an ongoing process... planned and documented before any significant volume of data has been collected" (^[35] [www.ncbi.nlm.nih.gov](https://pubmed.ncbi.nlm.nih.gov)). In practice, data managers follow a schedule of cleaning activities (often weekly or biweekly) where all collected data are run through the suite of programmed checks, and any resulting queries are addressed.

As data accrues, data managers may also conduct ad hoc analyses to catch errors not foreseen by the initial checks. For example, spotting an implausible spike in lab values or an unexpectedly high number of protocol violations at one site could trigger an inquiry. Central statistical monitoring (CSM) tools or simple graphical reviews can reveal site anomalies or systematic errors. Modern systems often support **query tracking dashboards**, which display metrics like number of open queries per site, average resolution time, and error severity. Farnum *et al.* mention a "Query Tracker" app that reports query counts, cycle times, and root-cause analyses, highlighting that query volume and turnaround are key performance indicators (^[43] [pmc.ncbi.nlm.nih.gov](https://pubmed.ncbi.nlm.nih.gov)).

The team resolves queries iteratively, usually continuously. For example, if an inconsistency between two lab values is flagged, the data manager generates a query to clarify which one is correct. The site corrects the appropriate CRF entry, and the manager updates the database. The objective is that by the time of database lock, all critical queries are closed and all remaining data values are validated or explained. An approximation of data readiness can be measured: some operations teams have a "Subject Tracker" to evaluate "subject data cleanliness and readiness for database lock" (^[44] [pmc.ncbi.nlm.nih.gov](https://pubmed.ncbi.nlm.nih.gov)), indicating how much of the anticipated data work remains.

By the final cleaning phase (months after last patient last visit), the aim is for 100% resolution of outstanding discrepancies. The Field Trials handbook describes this final sweep: once all checks have run and queries closed, the data are declared "clean" and the database is locked (^[45] [www.ncbi.nlm.nih.gov](https://pubmed.ncbi.nlm.nih.gov)). At lock, the data are frozen (no further edits allowed) and prepared for analysis. In GCP terms, lock marks the point after which the randomization code can be unblinded and definitive analyses performed (^[45] [www.ncbi.nlm.nih.gov](https://pubmed.ncbi.nlm.nih.gov)). A clean, locked database thus represents the culmination of the data cleaning process: "the data are as of high quality as possible, before they are analysed" (^[8] [www.ncbi.nlm.nih.gov](https://pubmed.ncbi.nlm.nih.gov)).

5. Types of Errors and Detection Methods

Clinical data can exhibit many error types. Understanding these helps data managers decide which checks to apply. Table 2 (below) summarizes common error categories (expanding Guidebook Table 1 ideas) and typical detection/cleaning strategies. Each trial may experience some or all these issues:

- **Transcription and typographical errors:** Mistyping digits or text (e.g. “135” instead of “1350”). Often caught by range checks (flag an out-of-range value) or by double-entry consistency checks. The only fix is to return to source (the patient chart) and correct.
- **Missing values:** Either due to skipped CRF fields or uncollected tests. Detected by edit checks that enforce mandatory fields, or reports of “missing for required field.” Requires querying to fill in a value (actual or coded as legitimately missing).
- **Logic/inconsistency errors:** Contradictory answers (e.g. female subject marked pregnant = Yes, but with male sex). Detected by cross-field logical rules. Resolved by clarifying the source answer.
- **Duplicate records:** The same subject data entered twice (possibly via data entry error or duplicate submission). Detected via matching algorithms on key identifiers. Cleaned by merging or deleting duplicates per protocol.
- **Format/coding errors:** Non-compliant entries for coded fields (e.g. MedDRA adverse event terms, medication lists). Detected by validation of code lists. Cleaned by recoding to standard dictionaries and querying unclear verbatim.
- **Outliers:** Valid values that are extreme (e.g. a lab result many SDs from the mean). These may represent true but rare events or entry errors (e.g. decimal in wrong place). Detected by statistical outlier tests or visual review. Usually flagged for verification and either corrected or retained as a true extreme.
- **Non-adherence to data standards:** e.g. entering text into a numeric field, date format inconsistencies, units confusion. Detected by type/format checks at entry. Corrected by standardizing formats or querying.
- **Data integration errors:** Mismatches when merging data from different sources (e.g. lab name mismatches, different coding for the same question in two systems). Detected through reconciliation processes, often with SAS or ETL scripts. Resolving requires mapping tables and data transformation.

Table 2: Detection methods and cleaning strategies for typical clinical trial data errors (examples derived from clinical data management literature) (^[11] [pmc.ncbi.nlm.nih.gov](https://pubmed.ncbi.nlm.nih.gov/)) (^[46] [pmc.ncbi.nlm.nih.gov](https://pubmed.ncbi.nlm.nih.gov/)) (^[8] www.ncbi.nlm.nih.gov).

The toolset for detecting these errors includes both built-in EDC checks and external processes. Within an EDC, study builders configure **edit checks** (also called data validation rules). These can be of several types:

- *Range checks:* Ensure a numeric value falls within a biologically plausible interval (e.g. 0–100 for percent).
- *Format checks:* For date/time or numeric fields, confirm correct format and delimiter usage.
- *Consistency checks:* Across fields, e.g. *HAVING if age < 18 then eligible flag must be No.*
- *Required-field checks:* Insist certain fields be non-empty (especially “key” variables like adverse events or primary outcomes).
- *Cross-form logic:* Check that events occur in logical sequence across visits.

Systems also allow custom scripts: as Farnum *et al.* note, complex checks that require combining multiple data sources are often implemented externally using SAS, R, or Python (^[47] [pmc.ncbi.nlm.nih.gov](https://pubmed.ncbi.nlm.nih.gov/)). For example, a CRA might program a SAS script to compare oxygen flow rate on vital signs forms vs concomitant medications to confirm alignment.

Besides programmed checks, data managers rely on **visualization and statistical review** for “sanity checks.” Frequency listings (e.g. count of adverse events by site) can reveal anomalies, and scatterplots of lab values may highlight outliers. Plots of query trends or key data elements over time can also signal irregularities. As a modern practice, risk-based monitoring tools incorporate statistical metrics (e.g. standard deviations, Mahalanobis distances) to flag sites or subjects warranting attention (^[43] [pmc.ncbi.nlm.nih.gov](https://pubmed.ncbi.nlm.nih.gov/)).

Ultimately, human review is indispensable. For nuanced issues (ambiguous text entries, subtle coding dilemmas) experienced data managers and medical coders inspect data detail. Field Trials guidelines stress that “above average familiarity with clinical trials, industry acronyms, ... jargon ... is highly recommended for [coders]” converting unstructured data into clean, standardized terms ⁽⁴⁴⁸⁾ [pmc.ncbi.nlm.nih.gov](https://pubmed.ncbi.nlm.nih.gov/)). This often involves skilled review of verbatim text and medical coding, which is beyond automated cleaning.

6. Tools and Technologies for Data Cleaning

Data managers employ a range of software tools to automate and streamline cleaning. Typical components include:

- **Electronic Data Capture (EDC) Systems:** Platforms like Oracle InForm, Medidata Rave, Veeva Vault CDMS, OpenClinica, Castor EDC, or REDCap handle query management natively. They host edit checks on CRFs and maintain audit trails of changes ⁽⁴⁴⁹⁾ [pmc.ncbi.nlm.nih.gov](https://pubmed.ncbi.nlm.nih.gov/) ⁽⁵⁰¹⁾ www.ncbi.nlm.nih.gov). Modern EDCs often include query dashboards and data listings. As of 2025–2026, leading platforms are integrating AI capabilities: Medidata Rave has introduced AI-powered automation for EDC setup and automated study locking, reducing configuration time and accelerating interim and final analysis ⁽⁵¹⁾ medidata.com). Veeva announced Vault CDMS enhancements including SiteVault CTMS integration (April 2025) and plans for Veeva eSource (expected H2 2026), which aims to eliminate paper at sites and streamline EHR-to-EDC data flow ⁽⁵²⁾ castoredc.com). Castor EDC has expanded its decentralized trial platform with integrated eCOA and eConsent capabilities.
- **Clinical Data Management Systems (CDMS):** These are comprehensive suites (e.g. Oracle Data Management Workbench, SAS Data Network, or custom CTMS/DCM solutions) that integrate EDC with lab and image data. Farnum *et al.* highlight systems like Oracle Life Sciences Hub, Xcellerate, or others that centralize data from multiple sources and feed integrated discrepancy checks ⁽⁵³⁾ [pmc.ncbi.nlm.nih.gov](https://pubmed.ncbi.nlm.nih.gov/) ⁽²²⁾ [pmc.ncbi.nlm.nih.gov](https://pubmed.ncbi.nlm.nih.gov/)). For example, Oracle’s Data Management Workbench can connect to different EDCs, offering “configureable data management workflows, extensive query/discrepancy management and library management” ⁽⁵³⁾ [pmc.ncbi.nlm.nih.gov](https://pubmed.ncbi.nlm.nih.gov/)). Oracle has further enhanced its EDC with AI-enabled EHR interoperability as of 2025, focusing on automated patient recruitment and monitoring workflows.
- **AI-Powered Data Quality Tools:** A growing category of tools applies artificial intelligence directly to data cleaning tasks. Veeva announced AI Agents across its applications beginning December 2025 for commercial use, expanding to R&D and quality in 2026. Pilot studies by consulting firms such as Deloitte have reported 20–30% time savings in data cleaning cycles using AI-assisted validation tools ⁽⁵⁴⁾ mckinsey.com). AI-based NLP tools for handling free-text entries in EDC are becoming mainstream, flagging ambiguous inputs and suggesting structured formats in near real-time ⁽⁵⁵⁾ clinion.com). These tools complement traditional programmed checks by detecting subtle patterns and anomalies that rule-based systems miss.
- **Statistical and Programming Tools:** SAS, R, Python, or even Excel are widely used for custom checks and reporting. Analysts might write SAS programs to apply batch validations (e.g. confirming that laboratory units match expected lab tests). R or Python could build anomaly detection using statistical or machine learning methods. In practice, much data cleaning is done in such scripts; the outputs (e.g. lists of suspect records) then feed query generation.
- **Data Tracking Dashboards:** Many sponsors use specialized QC dashboards (sometimes part of their CDMS) that visualize metrics: e.g. number of open queries, median response times, site performance on data timeliness, etc. These tools help manage workloads and pinpoint bottlenecks. The Farnum *et al.* Xcellerate system includes four applications (Discrepancy Manager, Page Tracker, Query Tracker, Subject Tracker) specifically to “streamline and automate” data cleaning tasks ⁽²²⁾ [pmc.ncbi.nlm.nih.gov](https://pubmed.ncbi.nlm.nih.gov/)).
- **Coding and Dictionaries:** For unstructured text, standardized coding systems (MedDRA for adverse events, WHO-DDEX for drugs, ICD for diagnoses) and coding tools (WHO Drug Dictionary or MedDRA auto-coders) are used. Trained coders review verbatim terms and map them, a form of cleaning that converts messy text into analyzable data ⁽⁴⁴⁸⁾ [pmc.ncbi.nlm.nih.gov](https://pubmed.ncbi.nlm.nih.gov/)). After coding, coded datasets are reconciled with entries to ensure consistency.

These tools all generate audit trails, a key requirement in regulated trials. Any change (even missing-to-withdrawn) must be logged with “who, what, when, why” information (^[27] pmc.ncbi.nlm.nih.gov). Controlling and documenting these processes is as important as correcting the data itself.

7. Quality Metrics and Impact

Effective data cleaning can be quantified and its impact assessed. Common metrics include error rates, query rates, and query response times. The synthesized literature provides benchmarks and warning signs:

- **Error Rates by Method:** Garza *et al.* (2023) conducted a meta-analysis of data entry errors across methods. They found that double-data entry yields ~0.14% errors, single-entry ~0.29%, and paper-based medical record abstraction up to 6.6% (^[1] pmc.ncbi.nlm.nih.gov). Such figures imply that poor entry methods (e.g. chart abstraction without verification) can introduce substantial noise. Moreover, the distribution of error rates was extremely wide (2 to 2784 errors per 10,000 fields across studies (^[1] pmc.ncbi.nlm.nih.gov)), emphasizing variability. These data support the use of rigorous checks: e.g. moving from single-entry to double-entry or from transcription to direct EDC can reduce error rates dramatically.
- **Query Statistics:** Anecdotal and internal reports often track queries. Farnum *et al.* argue that “the number of queries and the elapsed time before a query is acknowledged and closed are important determinants of study and site risk and performance (^[43] pmc.ncbi.nlm.nih.gov).” For instance, if it takes many weeks for a site to answer queries, data lock is delayed and cost accrues. Similarly, a high number of queries per subject or form suggests either poor data quality or overly sensitive edit checks that may need adjustment. Some sponsors set targets (e.g. >90% of queries closed within 2 weeks).
- **Proportion of Clean Data:** Trials occasionally report the percentage of data free from issue. The ASPREE trial case study (Section 9) found that after system improvements, 96.6% of data values were accurate or within specified range (^[2] trialsjournal.biomedcentral.com). Another study (Shaheen *et al.*, 2019) measured the fraction of datasets with zero “defects”; initially only 13% of studied datasets were defect-free, but a quality initiative raised that to 81% (^[46] pmc.ncbi.nlm.nih.gov). These kinds of stats can motivate and guide data cleaning efforts.
- **Cost/Benefit:** Data cleaning is not free. Pronker *et al.* (2011) examined the cost of monitoring and data cleaning, concluding that certain traditional methods (like manual sponsor queries after double entry) were “time and cost inefficient” (^[56] pmc.ncbi.nlm.nih.gov). Hence, a data manager must justify the extent of cleaning: beyond regulatory compliance, the ultimate “cost” of residual data error is potentially even higher, possibly invalidating study conclusions or requiring larger sample sizes (^[57] pmc.ncbi.nlm.nih.gov) (^[31] pmc.ncbi.nlm.nih.gov).
- **Impact on Statistical Power:** Garza *et al.* note that high error rates could require up to 20% larger sample sizes to maintain power (^[57] pmc.ncbi.nlm.nih.gov) (^[58] pmc.ncbi.nlm.nih.gov). Additionally, data errors can bias results, affecting p-values and confidence intervals (^[59] pmc.ncbi.nlm.nih.gov). Therefore, rigorous cleaning translates directly into reliable science.

By systematically tracking these metrics, sponsors can balance thoroughness with efficiency. When error rates or unresolved query counts exceed expectations, it may trigger a process review or protocol amendment (for example, adding more edit checks or staff training). Data managers often document data-quality findings for the trial master file, both for continuous improvement and as evidence of adequate cleaning in regulatory submissions.

8. Perspectives and Case Studies

Empirical studies and real-world examples illustrate the stakes and successes of data cleaning in trials. We present a few representative cases:

8.1 ASPREE Trial (AWARD System)

The Aspirin in Reducing Events in the Elderly (ASPREE) trial was a large, community-based study. It developed a special web-based data system (AWARD) with built-in checks to minimize errors. Brown *et al.* (2019) report a striking outcome: after implementing user-driven data checks, error rates plummeted and query workloads shrank (^[2] trialsjournal.biomedcentral.com). Data-entry errors fell from 0.3% to 0.01%, out-of-range entries from 0.14% to 0.04%, and protocol deviations from 0.4% to 0.08% (^[2] trialsjournal.biomedcentral.com). Overall, 96.6% of the 39 million collected values were either within specified range or confirmed accurate upon query, with only 3.4% ultimately missing (mostly due to unpredictable reasons like participant non-attendance) (^[2] trialsjournal.biomedcentral.com).

The key lesson is that good system design can sharply reduce the data cleaning burden. AWARD's functionality empowered users to catch and correct issues as data were entered, demonstrating the **impact of preventive data quality measures**. ASPREE's modest cleaning costs (lower "than expected compared with other trials") further underscore the payoff of robust EDC design (^[2] trialsjournal.biomedcentral.com). This case exemplifies how re-architecting workflows (in this case, towards site empowerment) can slay common errors and streamline cleaning.

8.2 Systematic Analysis of Entry Errors

Mitchell *et al.* (2011) evaluated errors in a specific trial database (a prostate hyperplasia drug study using Target e*CRF) (^[60] [pmc.ncbi.nlm.nih.gov](https://pubmed.ncbi.nlm.nih.gov)). They found most errors were simple transcription mistakes from the original source to the EDC. Critically, they noted that as soon as direct data entry is used (eliminating a transcribing step), error rates "should go down dramatically" (^[7] [pmc.ncbi.nlm.nih.gov](https://pubmed.ncbi.nlm.nih.gov)). Out-of-range values and protocol violations can then be flagged at entry, enabling immediate correction. This study provides concrete evidence that eliminating redundant transcription is a high-value cleaning strategy. In modern terms, it reinforces the benefit of site-entered eCRFs (and even better, integration with source EHRs) for improving data accuracy (^[7] [pmc.ncbi.nlm.nih.gov](https://pubmed.ncbi.nlm.nih.gov)).

8.3 Meta-Analysis of Error Rates

The Garza *et al.* (2023) systematic review is essentially a "study of studies" on data error rates (^[4] [pmc.ncbi.nlm.nih.gov](https://pubmed.ncbi.nlm.nih.gov)). It confirms that the mode of data capture deeply influences error rates. For a hypothetical trial, relying on retrospective chart abstraction (MRA) could yield errors on the order of 6–7% of all data fields, whereas double data entry would only produce about 0.14% errors (^[1] [pmc.ncbi.nlm.nih.gov](https://pubmed.ncbi.nlm.nih.gov)). Translating this: a trial with millions of data points under MRA could harbor tens of thousands of wrong values, potentially altering conclusions. The authors explicitly warn that such error rates *could necessitate increasing sample size* to preserve statistical power (^[61] [pmc.ncbi.nlm.nih.gov](https://pubmed.ncbi.nlm.nih.gov)).

This analysis underlines a broader implication: **data cleaning does not just maintain data; it affects study design and resources**. Choosing a more accurate capture method (even if more expensive during conduct) might lower the overall cost by reducing cleaning and sample sizes later. It also implies that data managers should pay attention to "data processing method" as a determinant of quality. The wide range of observed error rates (2 to 2784 per 10,000 fields) suggests that inconsistencies in counting error complicate comparisons (^[62] [pmc.ncbi.nlm.nih.gov](https://pubmed.ncbi.nlm.nih.gov)). Nonetheless, the clear hierarchy (double entry << single entry << scanning << MRA) stands out as actionable guidance.

8.4 Trial Harmonization Project (OUD Studies)

Balise *et al.* (2023) document the mess of harmonizing data from multiple opioid use disorder trials (^[63] [pmc.ncbi.nlm.nih.gov](https://pubmed.ncbi.nlm.nih.gov)). Though not about a single trial, it is an instructive example of data cleaning in a collaborative research context. The authors report that aligning the data from different trials (each designed independently) required

“an unexpectedly large amount of time” for data janitorial work (^[63] pmc.ncbi.nlm.nih.gov). They ended up creating 23 normalized database tables from three trials, resolving innumerable coding and organization discrepancies (^[64] pmc.ncbi.nlm.nih.gov) (^[63] pmc.ncbi.nlm.nih.gov).

Their take-home message is a cautionary one: lack of standardization can inflate cleaning work to a prohibitive level. Concretely, they suggest building in “guidance and checks” into data collection systems to avoid this burden (^[63] pmc.ncbi.nlm.nih.gov). This suggests that at trial start-up, planning for interoperability and metadata standards (e.g. CDISC models for variables) can greatly simplify future cleaning. It also illustrates that data cleaning work extends beyond the single trial: within meta-analyses or registries, inconsistent coding saps enormous resources.

9. Best Practices and Strategies

Drawing on guidelines and expert recommendations, the following strategies are emphasized by multiple sources:

- **Continuous Cleaning:** Begin data cleaning at trial launch, not just at closeout. The Field Trials manual and Oronsky *et al.* alike stress that cleaning is ongoing (^[35] www.ncbi.nlm.nih.gov) (^[6] pmc.ncbi.nlm.nih.gov). Early detection (e.g. within days of entry) prevents error propagation.
- **Collaborative Query Resolution:** Maintain strong communication channels with sites. Queries should be issued promptly, and their resolution should involve the original data-collection team or investigator to ensure accuracy. Speedy turnaround on queries accelerates cleaning – as one report notes, “the faster the turnaround time on these queries, the sooner it is possible to start cleaning the data” (^[65] pmc.ncbi.nlm.nih.gov).
- **Training and Standardization:** Train site staff on correct data entry and emphasize immediate entry (which enables real-time validation) (^[66] pmc.ncbi.nlm.nih.gov). Standard codes and value lists reduce ambiguity. Encourage staff to resolve minor queries (like obvious typos) internally by empowering “supervisors” among data entry clerks (^[67] www.ncbi.nlm.nih.gov).
- **Use of Double Data Entry Selectively:** While double entry can greatly reduce data entry errors (^[39] www.ncbi.nlm.nih.gov) (^[11] pmc.ncbi.nlm.nih.gov), it is costly. Many teams use double entry for critical study data (primary endpoints, key demographics) and single entry elsewhere, balancing resources. Modern EDCs partially substitute for double entry by real-time error checking (^[7] pmc.ncbi.nlm.nih.gov).
- **Audit Trail Discipline:** Every change in data cleaning must be justified in the audit trail. The locked database should be fully reproducible from source by the audit log (^[42] www.ncbi.nlm.nih.gov). Unjustified or undocumented fixes can be flagged during inspection.
- **Plan for Retrospective Checks:** Recognize that some errors (e.g. systematic site misinterpretations) may only surface in retrospective review. The plan should allow for final data review after lock and possibilities for updates (with justification) if errors are found during analysis (^[42] www.ncbi.nlm.nih.gov).
- **Risk-Based Focus:** Apply more intense cleaning to high-impact data. Analytical plans often pre-specify variables most critical for primary efficacy or safety. Ensure these variables have minimal missingness and error. For less critical fields, a somewhat higher error rate might be tolerable.
- **Metrics and Feedback Loops:** Continuously monitor data quality metrics (as in Section 7). If, for example, a particular site consistently lags on data timeliness or query closure, address it through retraining or targeted monitoring visits. High-level dashboards should allow project teams to prioritize cleaning resources efficiently.
- **Technology Leverage:** Take advantage of new tools. For example, the integration of EHR-to-EDC interfaces can automate data transfer and reduce manual entry. Some sponsors use risk-based monitoring tools that include statistical data quality algorithms. In the future, machine learning may further aid anomaly detection (see Section 11).

In practice, this often translates to formal SOPs (Standard Operating Procedures) that codify the above points. Many organizations maintain Data Management charters that define responsibilities, timelines, and tools for cleaning.

10. Implications and Future Directions

The landscape of clinical trials is rapidly evolving, and data managers must adapt their cleaning strategies accordingly. Key implications and trends include:

- **Growing Data Volumes and Complexity:** Trials increasingly collect high-dimensional data (genomics, wearables, continuous monitoring). This explosion heightens the cleaning challenge. Traditional manual query resolution cannot scale to millions of data points; automated data validation and sampling techniques become essential.
- **Real-World Data Integration:** The integration of real-world data (RWD) sources (insurance claims, electronic health records) introduces new cleaning tasks. RWD often lack the structure/monitoring of trial data. Alignment of RWD with trial definitions (e.g. reconciling diagnoses codes, temporal alignment of events) will require sophisticated cleaning and harmonization approaches.
- **Automation and Artificial Intelligence:** Machine learning and AI have moved beyond the research stage and are now actively enhancing data cleaning in clinical trials. ML-based anomaly detection can flag unusual patterns beyond simple range checks (^[68] pmc.ncbi.nlm.nih.gov). Automated imputation techniques (like k-nearest neighbors or deep learning) can handle missing data more intelligently. Major platform vendors have begun embedding AI directly into their products: Medidata Rave's AI-powered automation now handles study setup and locking, Veeva's AI Agents (launched December 2025) extend across R&D and quality workflows, and consulting firms report 20–30% time savings in data cleaning cycles with AI-assisted validation (^[54] mckinsey.com). AI-based NLP tools for handling free-text entries in EDC are also becoming mainstream, flagging ambiguous inputs and suggesting structured formats in near real-time (^[55] clinion.com). Regulatory frameworks are catching up: the FDA's January 2025 draft guidance on AI in drug development introduced a 7-step credibility assessment framework, and the joint FDA–EMA “Guiding Principles of Good AI Practice in Drug Development” (January 2026) establish expectations for transparency and validation of AI-assisted tools (^[30] aihnet.com). Early experiments (like Jarmakovica *et al.* using ML to improve completeness from 90% to nearly 100% in a dataset) continue to illustrate the promise (^[68] pmc.ncbi.nlm.nih.gov), and these approaches are now transitioning from pilot studies to production use.
- **Decentralized Trials and EDC Evolution:** The rise of decentralized or virtual trials (with remote monitoring and at-home data collection) further underscores data quality concerns. As sites fragment, centralized data review becomes even more critical. ICH E6(R3) Annex 2, currently in development and expected to be finalized in mid-2026, will provide the first harmonized regulatory framework specifically addressing decentralized and pragmatic trial designs (^[25] acrpnet.org). New platforms that integrate ePRO data from patients, telehealth records, and IoT devices will require expanded cleaning frameworks. Veeva's upcoming eSource platform (expected H2 2026) aims to eliminate paper at clinical sites entirely and streamline EHR-to-EDC data flow. The industry is shifting from measuring DCT success by technology features to measuring it by outcomes—shorter cycle times, fewer protocol deviations, and fewer missing endpoints (^[69] globalforum.diaglobal.org). Integration remains the primary challenge: platforms without robust APIs force manual processes that undermine the efficiency gains of decentralization. Data managers will need to become adept at handling diverse data formats and authentication mechanisms.
- **Regulatory Focus on Data Integrity:** Regulatory attention to data integrity has intensified significantly. The FDA's September 2025 Computer Software Assurance (CSA) guidance modernizes the validation landscape for clinical data systems, endorsing risk-based approaches and recognizing cloud-based platforms (^[28] federalregister.gov). ICH E6(R3)'s emphasis on quality-by-design and proportionate oversight directly impacts how data cleaning processes should be planned and documented. Any indication of inadequate data cleaning can delay approvals. Conversely, transparent cleaning processes (version-controlled and well-documented) can expedite reviews. Additionally, a 2025 JMIR systematic review of 851 clinical data quality papers proposed a comprehensive 4-stage data quality life cycle (planning, construction, operation, utilization) and identified the most critical quality dimensions as completeness, plausibility, concordance, security, currency, and interoperability (^[70] jmir.org).
- **Standardization and Interoperability:** Initiatives like CDISC (Clinical Data Interchange Standards Consortium) continue to shape cleaning practices. Standard data models (like CDASH and ADaM) when used prospectively

reduce discrepancies. CDISC published updated Controlled Terminology files in March 2026 with approximately 248 new QRS terms and 1,124 new terms across ADaM, CDASH, SDTM, and SEND files, reflecting the rapid evolution of data standards (^[71] [cdisc.org](https://www.cdisc.org)). Major new versions—SDTM v3.0 and ADaM v3.0—are in development to consolidate models and address complex study designs. Clean data at the point of collection, structured per standards, simplifies later pooling and analysis (^[19] [pmc.ncbi.nlm.nih.gov](https://pubmed.ncbi.nlm.nih.gov/)) (^[35] www.ncbi.nlm.nih.gov). Interoperability via FHIR APIs between EHRs and EDCs is progressing, with Oracle enhancing AI-enabled EHR interoperability in 2025 and Veeva planning native eSource integration for 2026—developments that could significantly lower transcription errors while raising new privacy and harmonization challenges.

- **Workforce and Training:** Data managers will need to adopt new skills (e.g. coding expertise, familiarity with big data tools) and to manage hybrid (AI-human) cleaning workflows. Training programs are evolving: the SCDM launched a beta test for the new **Certified Clinical Data Scientist (CCDS)** credential in 2025, with formal launch expected in 2026, recognizing that modern data managers need expertise spanning traditional CDM, statistics, and data science (^[33] [scdm.org](https://www.scdm.org)). Emphasis on data integrity may lead to integrated roles: e.g., statisticians and data managers collaborating on cleaning designs.
- **Ethical and Privacy Considerations:** As data cleaning often requires recontacting sites or linking data, privacy rules (HIPAA, GDPR) influence practices. Patient identifiers must be protected throughout cleaning; anonymized coding raises special issues if reidentification is needed. Advanced de-identification and secure query platforms may become standard.

In summary, the future will likely see more automation and stricter standards. Tools will shift data cleaning from laborious manual tasks to intelligent systems, but the fundamental principles remain: clear protocols, targeted checks, and documentation. Investments in cleaning – once seen as a mere cost – will increasingly be recognized as crucial to data integrity and trial success.

11. Conclusion

Data cleaning is an integral, labor-intensive, and ongoing activity at the heart of clinical trial data management. This report has detailed the processes by which data managers find and rectify errors, illustrating the myriad techniques from CRF design to query resolution. We have emphasized that high-quality data underpin all valid trial conclusions and regulatory decisions. As the literature shows, even small error rates can cascade into serious problems for trials (^[1] [pmc.ncbi.nlm.nih.gov](https://pubmed.ncbi.nlm.nih.gov/)) (^[58] [pmc.ncbi.nlm.nih.gov](https://pubmed.ncbi.nlm.nih.gov/)), and the clinical trial community has developed robust methodologies to address this challenge.

Key themes include: the necessity of planning and continuous cleaning (not deferring to the end), the importance of tailored checks for logical consistency, and the central role of queries in resolving uncertainties (^[8] www.ncbi.nlm.nih.gov) (^[35] www.ncbi.nlm.nih.gov). Real-world examples (like the ASPREE trial with its improved data suite) demonstrate that thoughtful system design can yield dramatic error reductions and cost savings (^[2] trialsjournal.biomedcentral.com). Separate analyses confirm that a choice of data processing impacts error rates profoundly (^[1] [pmc.ncbi.nlm.nih.gov](https://pubmed.ncbi.nlm.nih.gov/)), guiding decisions around double data entry and direct EDC.

Looking forward, the core goal remains: *data that support the same conclusion as error-free data* (^[9] [pmc.ncbi.nlm.nih.gov](https://pubmed.ncbi.nlm.nih.gov/)). The 2025–2026 period has brought transformative developments: ICH E6(R3) now establishes quality-by-design as the standard for GCP, the FDA's CSA guidance modernizes software validation, and AI-powered tools are transitioning from pilots to production across major EDC platforms. Data managers must adapt by embracing new tools (ML, advanced EDC integration, NLP-based validation) while upholding the discipline of good CDM practices (audit trails, verification, documentation). Ultimately, rigorous data cleaning preserves the trial's validity, maintains patient safety, and protects the health of public policy. In the words of Sibson (cited in Shaheen *et al.*), “a dataset is indispensable to answer research questions.” Therefore **maintaining its accuracy** through diligent cleaning is not optional—it is a mandate of ethical and scientific rigor (^[72] [pmc.ncbi.nlm.nih.gov](https://pubmed.ncbi.nlm.nih.gov/)) (^[4] [pmc.ncbi.nlm.nih.gov](https://pubmed.ncbi.nlm.nih.gov/)).

- [24] <https://about.citiprogram.org/blog/ich-releases-final-version-of-e6r3-good-clinical-practice-guideline/>
- [25] <https://acrpnet.org/2026/02/17/ich-e6r3-unpacked-diving-deep-into-the-impacts-of-the-guideline-changes>
- [26] <https://ichgcp.net/monitoring#:~:Monit...>
- [27] <https://pmc.ncbi.nlm.nih.gov/articles/PMC10499417/#:~:clini...>
- [28] <https://www.federalregister.gov/documents/2025/09/24/2025-18468/computer-software-assurance-for-production-and-quality-system-software-guidance-for-industry-and>
- [29] <https://www.ncbi.nlm.nih.gov/books/NBK305509/#:~:Data%...>
- [30] <https://www.aihnet.com/blog/fda-ai-guidance-data-integrity-challenges/>
- [31] <https://pmc.ncbi.nlm.nih.gov/articles/PMC10775420/#:~:count...>
- [32] <https://scdm.org/gcdmp-chapter-review-for-clinical-trial-database-lock/>
- [33] <https://scdm.org/updates-from-the-board-q4-2025/>
- [34] <https://pmc.ncbi.nlm.nih.gov/articles/PMC12319021/#:~:Healt...>
- [35] <https://www.ncbi.nlm.nih.gov/books/NBK305509/#:~:Data%...>
- [36] <https://www.ncbi.nlm.nih.gov/books/NBK305509/#:~:not%2...>
- [37] <https://www.ncbi.nlm.nih.gov/books/NBK305509/#:~:One%2...>
- [38] <https://www.ncbi.nlm.nih.gov/books/NBK305509/#:~:Data%...>
- [39] <https://www.ncbi.nlm.nih.gov/books/NBK305509/#:~:Doubl...>
- [40] <https://pmc.ncbi.nlm.nih.gov/articles/PMC6378235/#:~:amoun...>
- [41] <https://www.ncbi.nlm.nih.gov/books/NBK305509/#:~:Some%...>
- [42] <https://www.ncbi.nlm.nih.gov/books/NBK305509/#:~:It%20...>
- [43] <https://pmc.ncbi.nlm.nih.gov/articles/PMC6378235/#:~:Queri...>
- [44] <https://pmc.ncbi.nlm.nih.gov/articles/PMC6378235/#:~:deliv...>
- [45] <https://www.ncbi.nlm.nih.gov/books/NBK305509/#:~:When%...>
- [46] <https://pmc.ncbi.nlm.nih.gov/articles/PMC6511206/#:~:At%20...>
- [47] <https://pmc.ncbi.nlm.nih.gov/articles/PMC6378235/#:~:inspe...>
- [48] <https://pmc.ncbi.nlm.nih.gov/articles/PMC10499417/#:~:CRFs%...>
- [49] <https://pmc.ncbi.nlm.nih.gov/articles/PMC6378235/#:~:Intro...>
- [50] <https://www.ncbi.nlm.nih.gov/books/NBK305509/#:~:When%...>
- [51] <https://www.medidata.com/en/clinical-trial-products/clinical-data-management/edc-systems/>
- [52] <https://www.castoredc.com/insight-briefs/decentralized-clinical-trial-platforms-in-2025-a-practical-guide-for-clinical-operations/>
- [53] <https://pmc.ncbi.nlm.nih.gov/articles/PMC6378235/#:~:A%20s...>
- [54] <https://www.mckinsey.com/industries/life-sciences/our-insights/unlocking-peak-operational-performance-in-clinical-development-with-artificial-intelligence>
- [55] <https://www.clinion.com/insight/ai-in-clinical-data-management/>
- [56] <https://pmc.ncbi.nlm.nih.gov/articles/PMC3045557/#:~:In%20...>
- [57] <https://pmc.ncbi.nlm.nih.gov/articles/PMC10775420/#:~:Data%...>

IntuitionLabs - Industry Leadership & Services

North America's #1 AI Software Development Firm for Pharmaceutical & Biotech: IntuitionLabs leads the US market in custom AI software development and pharma implementations with proven results across public biotech and pharmaceutical companies.

Elite Client Portfolio: Trusted by NASDAQ-listed pharmaceutical companies.

Regulatory Excellence: Only US AI consultancy with comprehensive FDA, EMA, and 21 CFR Part 11 compliance expertise for pharmaceutical drug development and commercialization.

Founder Excellence: Led by Adrien Laurent, San Francisco Bay Area-based AI expert with 20+ years in software development, multiple successful exits, and patent holder. Recognized as one of the top AI experts in the USA.

Custom AI Software Development: Build tailored pharmaceutical AI applications, custom CRMs, chatbots, and ERP systems with advanced analytics and regulatory compliance capabilities.

Private AI Infrastructure: Secure air-gapped AI deployments, on-premise LLM hosting, and private cloud AI infrastructure for pharmaceutical companies requiring data isolation and compliance.

Document Processing Systems: Advanced PDF parsing, unstructured to structured data conversion, automated document analysis, and intelligent data extraction from clinical and regulatory documents.

Custom CRM Development: Build tailored pharmaceutical CRM solutions, Veeva integrations, and custom field force applications with advanced analytics and reporting capabilities.

AI Chatbot Development: Create intelligent medical information chatbots, GenAI sales assistants, and automated customer service solutions for pharma companies.

Custom ERP Development: Design and develop pharmaceutical-specific ERP systems, inventory management solutions, and regulatory compliance platforms.

Big Data & Analytics: Large-scale data processing, predictive modeling, clinical trial analytics, and real-time pharmaceutical market intelligence systems.

Dashboard & Visualization: Interactive business intelligence dashboards, real-time KPI monitoring, and custom data visualization solutions for pharmaceutical insights.

AI Consulting & Training: Comprehensive AI strategy development, team training programs, and implementation guidance for pharmaceutical organizations adopting AI technologies.

Contact founder Adrien Laurent and team at <https://intuitionlabs.ai/contact> for a consultation.

DISCLAIMER

The information contained in this document is provided for educational and informational purposes only. We make no representations or warranties of any kind, express or implied, about the completeness, accuracy, reliability, suitability, or availability of the information contained herein.

Any reliance you place on such information is strictly at your own risk. In no event will IntuitionLabs.ai or its representatives be liable for any loss or damage including without limitation, indirect or consequential loss or damage, or any loss or damage whatsoever arising from the use of information presented in this document.

This document may contain content generated with the assistance of artificial intelligence technologies. AI-generated content may contain errors, omissions, or inaccuracies. Readers are advised to independently verify any critical information before acting upon it.

All product names, logos, brands, trademarks, and registered trademarks mentioned in this document are the property of their respective owners. All company, product, and service names used in this document are for identification purposes only. Use of these names, logos, trademarks, and brands does not imply endorsement by the respective trademark holders.

IntuitionLabs.ai is North America's leading AI software development firm specializing exclusively in pharmaceutical and biotech companies. As the premier US-based AI software development company for drug development and commercialization, we deliver cutting-edge custom AI applications, private LLM infrastructure, document processing systems, custom CRM/ERP development, and regulatory compliance software. Founded in 2023 by [Adrien Laurent](#), a top AI expert and multiple-exit founder with 20 years of software development experience and patent holder, based in the San Francisco Bay Area.

This document does not constitute professional or legal advice. For specific guidance related to your business needs, please consult with appropriate qualified professionals.

© 2025 IntuitionLabs.ai. All rights reserved.