CHAPTER 4

METHODS FOR ACQUISITION, CLEANING, AND VALIDATION OF CLINICAL DATA

Author

Mrs. Udaya Kumari Tula, Research Scientist,
DSK Biopharma Inc., Morrisville, North Carolina, USA

Abstract

The transition from protocol design to active data collection marks the operational heartbeat of a clinical trial, known as the Conduct Phase. This stage focuses on the mechanics of ingesting, interrogating, and refining vast quantities of medical information. The primary vehicle for this acquisition is the Electronic Data Capture (EDC) system, which has largely replaced paper-based methods, allowing for real-time visibility into site activities. However, raw data entered by clinical sites is rarely perfect. Ensuring its integrity requires a robust Discrepancy Management process, where Data Managers issue electronic queries to sites to resolve ambiguities, correct transcription errors, and explain protocol deviations without leading the investigator. This tactical cleaning is governed by a comprehensive Data Validation Plan (DVP), which translates the protocol's scientific criteria into executable logic checks and manual review strategies. Beyond the EDC, modern trials must also integrate complex streams of external data from third-party vendors, such as Central Laboratories and cardiac safety centers. Managing the technical transfer, normalization, and reconciliation of these external datasets is critical to ensuring that safety and efficacy endpoints are accurately captured. Ultimately, the rigorous application of these acquisition and cleaning methodologies transforms raw clinical observations into a high-fidelity dataset capable of supporting statistical analysis and regulatory scrutiny.

Keywords: *Discrepancy Management, Data Validation Plan (DVP), Query Management, External Data Reconciliation, Data Cleaning Strategies*

Learning Objectives

After completion of the chapter, the learners should be able to:

- Describe the workflow of data acquisition via Electronic Data Capture (EDC) systems and the transition from paper-based methods.
- Demonstrate the ability to write non-leading queries during discrepancy management to resolve data inconsistencies without biasing the investigator.
- Develop a Data Validation Plan (DVP) that translates protocol criteria into specific executable logic checks and manual review steps.
- Manage the technical transfer and reconciliation of external data sources, such as Central Laboratory results and ECG files.
- Apply data cleaning strategies to identify and resolve discrepancies between source documents and the clinical database.

DATA ENTRY PROCESSES AND ELECTRONIC DATA CAPTURE (EDC)

The transition from the Setup Phase to the Conduct Phase marks a pivotal shift in the lifecycle of a clinical trial. The planning is complete, the database is live, and the first patients have been enrolled. Now, the focus turns to the acquisition of data. This is the operational engine of the trial, where the theoretical design of the protocol meets the practical reality of clinical medicine. The primary mechanism for this acquisition in modern research is Electronic Data Capture (EDC), a technology that has fundamentally reshaped the speed and quality of drug development.

The Workflow of Data Acquisition

The journey of a data point begins long before it reaches the sponsor's database. It starts at the clinical site the hospital or clinic where the patient is being treated. The process adheres to a strict hierarchy of records to ensuring data integrity, moving from source documents to the EDC system.

Source Data and Source Documents

The first time a piece of data is recorded, it is classified as "Source Data." This could be a blood pressure reading written in a patient's medical chart, a lab result printed from an analyzer, or a diary entry made by the patient. The medium containing this original data is the "Source Document." Regulatory guidelines mandate that all data entered into the clinical trial database must be traceable back to these source documents. This principle, known as ALCOA+ (Attributable, Legible, Concurrent, Original, and Accurate), ensures that the data is not fabricated.

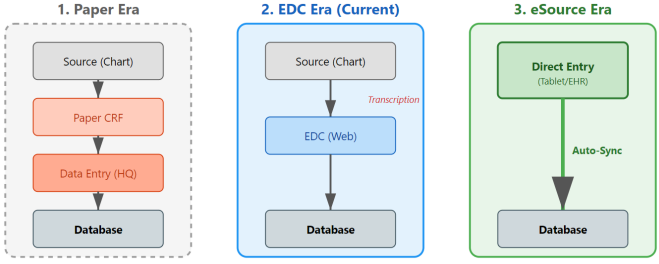

Figure 4.1: The Evolution of Data Capture

Transcription and the EDC Interface

In a traditional EDC workflow, the Clinical Research Coordinator (CRC) acts as the bridge between the source document and the database. After the patient visit, the CRC logs into the EDC system (a secure, web-based platform) and transcribes the data from the source documents into the electronic Case Report Forms (eCRFs). This transcription process is the most vulnerable point in the data chain, as human error can lead to discrepancies between the source and the database.

Electronic Data Capture (EDC) Systems

EDC systems replaced the carbon-copy paper CRFs of the 20th century. Modern platforms like Medidata Rave, Oracle InForm, and Veeva Vault CDMS are sophisticated software suites that do more than just store data; they actively manage the

trial.

Table 4.1: Comparison of Data Collection Methodologies

Methodology	Data Entry Workflow	Primary Advantage	Primary Challenge
Standard EDC	Source Document → Human Transcription → Database	Investigator reviews data before entry; established workflow.	Transcription errors and significant lag time between patient visit and data entry.
eSource (DDC)	Direct entry into Tablet/Device at Bedside	Eliminates transcription errors completely; real-time data visibility.	Requires robust internet connectivity and hardware management
ePRO / eCOA	Patient enters data on Smartphone/ Handheld (BYOD or Provisioned)	Eliminates recall bias (data entered when event happens); precise timestamp.	Patient compliance; device battery failure or loss.
EHR Integration	Automated pull from Hospital Electronic Health Records	Access to massive volume of rich clinical data without manual entry.	Interoperability issues (different hospitals use different data standards like HL7 vs FHIR).

Real-Time Data Visibility

The most transformative feature of EDC is real-time visibility. In the era of paper, sponsors might wait weeks or months for forms to be mailed and data entered. With EDC, a data point entered by a nurse in Tokyo is visible to a Data Manager in New York seconds later. This allows for immediate oversight of safety signals and enrollment trends.

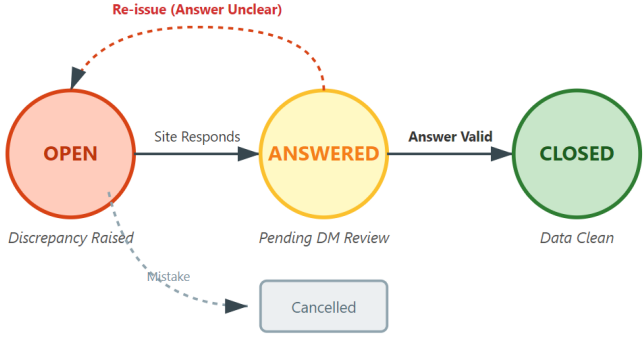

Figure 4.2: The Query Lifecycle

Integrated Functionality

EDC systems are rarely standalone tools anymore. They serve as the central hub of an integrated ecosystem. They often link directly with Interactive Response Technology (IRT) for randomizing patients and managing drug supply. They also integrate with medical coding dictionaries (MedDRA and WHO-DD) to standardize medical terms automatically. This integration reduces the need for manual data reconciliation between disparate systems.

The Rise of eSource and Direct Data Capture (DDC)

While EDC improved transcription, it did not eliminate it. The industry is currently witnessing a paradigm shift towards "eSource" or Direct Data Capture (DDC). In this model, the data is entered directly into an electronic device (like a tablet) at the point of care, eliminating the paper middleman entirely.

Eliminating Transcription Errors

eSource virtually eliminates transcription errors by removing the transcription step. The source data and the database data become identical by definition. This has profound implications for monitoring; since there is no transcription to verify, Clinical Research Associates (CRAs) can spend less time checking numbers (Source Data Verification) and more time ensuring the site is following the protocol and maintaining patient safety.

Electronic Health Record (EHR) Integration

The ultimate goal of eSource is full interoperability with Electronic Health Records (EHR). In this futuristic workflow, data collected during routine clinical care (e.g., a hospital's Epic or Cerner system) would automatically populate the clinical trial database. While technical and privacy challenges remain (due to regulations like HIPAA and GDPR), this integration represents the frontier of efficient data acquisition.

The Investigator's Responsibility

Regardless of the technology used whether paper, EDC, or eSource the ultimate responsibility for the integrity of the data lies with the Principal Investigator (PI). ICH-GCP guidelines are explicit: the investigator must ensure the accuracy, completeness, legibility, and timeliness of the data reported.

Electronic Signatures

In an EDC environment, the PI exercises this responsibility through electronic signatures. Once data for a patient visit is complete and cleaned, the PI must log in and "sign" the electronic casebook. This signature is the legal equivalent of a handwritten signature on a paper document. It signifies that the doctor has reviewed the data and attests to its validity. This oversight is crucial; it confirms that the trial data is not just an administrative output, but a verified medical record.

DISCREPANCY MANAGEMENT

RAISING AND RESOLVING QUERIES

In the ecosystem of a clinical trial, data entry is the ingestion of information, but **discrepancy management** acts as the immune system, identifying and neutralizing errors that could compromise the study's integrity. No matter how well a protocol is designed or how trained the site staff are, raw clinical data is inherently "dirty." Handwriting on source documents can be ambiguous, units of measurement can be transposed, and complex medical logic can be misinterpreted.

Discrepancy management is the systematic, documented process of identifying these inconsistencies, investigating their root causes, and resolving them to ensure the database accurately reflects the clinical reality. This process is driven by the **Query** a formal, audible electronic communication between the sponsor (or CRO) and the investigator site.

The Anatomy and Classification of Discrepancies

A discrepancy is defined as any data point that fails to meet the pre-defined criteria for quality, logic, completeness, or consistency. While the end result a query is often the same, the origin of discrepancies varies significantly. They are generally classified into two primary streams:

Automated Discrepancies (System-Generated)

These are the first line of defense, triggered instantaneously by the "Edit Checks" programmed into the Electronic Data Capture (EDC) system during the setup phase.

- **Range Checks:** Identifying values that are physiologically impossible or improbable (e.g., a Body Mass Index of 5 or 150).
- **Logic Checks:** Identifying logical impossibilities within the data structure (e.g., a "Stop Date" for a medication that occurs prior to the "Start Date").
- **Completeness Checks:** Flagging mandatory fields that have been left blank (e.g., missing adverse event severity).

Automated checks are highly efficient for catching "syntax"