A CONTEXTUAL FRAMEWORK FOR DATA CLEANING IN CLINICAL RESEARCH

CASE STUDY: DPSP STUDY AT MASAFU HOSPITAL BUSIA DISTRICT

A dissertation presented to

FACULTY OF SCIENCE

in partial fulfillment of the requirements for the award of the degree

Master of Science in Information Systems

UGANDA MARTYRS UNIVERSITY

PIUS Kavuma

2023-M132-21500

Supervisor: Julius Muganji

September 2025

UGANDA MARTYRS UNIVERSITY

DIRECTORATE OF GRADUATE STUDIES, RESEARCH AND ENTERPRISE

Master's Dissertation

Declaration

I have read the rules of Uganda Martyrs University on plagiarism and academic honesty, and hereby state that this work is my own.

It has not been submitted to any other institution for another degree or qualification, either in full or in part.

Throughout the work I have acknowledged all sources used in its compilation.

I finally grant Uganda Martyrs University permission to store and reproduce this dissertation, in whole or in part, in any manner or format, which Uganda Martyrs University may deem fit.

Researcher's name: KAVUMA PIUS

Researcher's signature: 

Date of submission: 25th /08/2025

Submitted to the Directorate of Graduate Studies, Research and Enterprise

UGANDA MARTYRS UNIVERSITY

DIRECTORATE OF GRADUATE STUDIES, RESEARCH AND ENTERPRISE

Master's Dissertation

Approval

This dissertation has been produced under my/our supervision and submitted for examination with my/our approval as the appointed academic supervisor/s.

Name of Supervisor: JULIUS MUGANJI

Signature of Supervisor: 

Date of submission: 6/9/2025

Submitted to the Directorate of Graduate Studies, Research and Enterprise

Acknowledgements
I would like to thank the following people for their contribution and morale support towards the completion of this research project: My supervisor Mr.Muganji Julius, my mother Magaret Nakazibwe, My brothers Charles Kavuma and Dougalas Kavuma as well as My Sister Nagiitta Josephine, and above all, to the Lord, God!. Without your love, guidance, patience and support, It would have been a much harder journey to complete this work.

Contents

List of Tables

List of Figures

List of Acronyms

DPSP - Dihydroartemisinin-Piperaquine and Sulfadoxine-Pyrimethamine

EDC - electronic data capture

DQM - Data Quality Management

LMIC- low- and middle-income country

TDQM- Total Data Quality Management

IQT- Information Quality Theory

KRT - Knowledge Representation Theory

STS - Socio-Technical Systems

IQT - Information Quality Theory

GCP - Good Clinical Practice

Abstract

This study presents the design, development, and validation of a contextual data cleaning framework tailored for clinical research settings in low-resource environments, using the DPSP (Dihydroartemisinin-Piperaquine and Sulfadoxine-Pyrimethamine) trial at Masafu Hospital as a case study. The research was motivated by persistent data quality challenges—such as missing values, inconsistencies, human errors, and tool limitations—that often compromise the validity and reliability of clinical research outcomes. Employing a user-intervention methodology, the study integrated qualitative insights from data managers, clinical teams, and analysts with quantitative assessment techniques to ensure that the proposed framework aligns with real-world practices. The framework was structured into distinct phases, including data profiling, preprocessing, modular cleaning, enhancement, and quality scoring—each mapped to address specific data integrity issues. Validation on the DPSP dataset demonstrated a significant improvement in data accuracy (from 75% to 94%), completeness (from 68% to 90%), and consistency (from 70% to 93%), confirming the framework's effectiveness and usability. SQL-driven automation further improved scalability and reduced human error. The study contributes to the literature by offering a novel, context-sensitive approach that balances domain expertise with technical rigor. It recommends future work to expand the framework's applicability to unstructured data types and to assess its operational integration and cost-effectiveness. Overall, the framework serves as a practical tool for improving data quality in clinical trials and enhancing the credibility of health research in resource-constrained settings.

CHAPTER ONE:

INTRODUCTION

1.1 Background of the Study

In the digital age, data has become an indispensable asset in clinical research, playing a pivotal role in informing health decisions, influencing policy, guiding clinical practices, and ultimately impacting patient outcomes. The advent of electronic data collection systems has led to the generation of vast amounts of clinical data, particularly in randomized controlled trials (RCTs) and observational studies. However, the utility of such data is heavily dependent on its quality, and this quality is largely shaped by the processes applied during data collection, handling, and cleaning (Acharya et al., 2023; Love et al., 2021).

Data cleaning, often referred to as data cleansing, is a critical preprocessing step aimed at identifying and correcting errors, inconsistencies, or inaccuracies in datasets (Fatima & Ali, 2017). This process ensures that data is accurate, complete, consistent, and reliable before it is subjected to analysis. In clinical research settings, data cleaning is not merely a technical task but a scientific necessity. Poorly cleaned data can introduce bias, reduce statistical power, and lead to invalid conclusions, thereby compromising patient safety and misguiding treatment recommendations (Jakobsen et al., 2017).

Despite its importance, data cleaning remains an underemphasized and underreported component of clinical research workflows. Many clinical trials lack comprehensive documentation of how errors were detected and corrected, or what framework was used to validate the final dataset (Gesicho & Were, 2020). In most cases, researchers rely on ad hoc methods or conventional spreadsheet tools without a contextual framework tailored to the nature of their data or the clinical environment in which it was collected. This approach is problematic, especially in low-resource settings where standardization is lacking and manual data entry is still prevalent (Gudivada et al., 2017).

Clinical studies conducted in settings such as Masafu Hospital in Busia District, Uganda, often face multiple challenges including heterogeneous data sources, multilingual entries, inconsistent

formats, missing values, and duplicate records due to repeated patient visits. These challenges highlight the need for a contextual data cleaning framework one that takes into account the local realities of data collection, staff capacity, the type of clinical trial, and the structure of the health information systems in place. Existing frameworks, while effective in some contexts, often fail to capture such nuances, making them unsuitable for direct application in localized clinical research (Ilyas & Chu, 2019).

Research by Calabrese et al. (2018) and Pindya (2024) underscores that data quality is not merely a technical issue, but a strategic concern tied to organizational performance and patient well-being. As clinical data becomes more intertwined with machine learning, big data analytics, and real-time decision-making, the ramifications of "dirty data" are amplified. Erroneous data can corrupt training sets, skew algorithmic outputs, and lead to flawed health recommendations. Therefore, ensuring clean and standardized data is not just a preparatory step it is a central pillar of credible and ethical clinical research (Calabrese, 2018; Pindya, 2024).

Additionally, the emergence of electronic data capture (EDC) systems and cloud-based platforms has introduced new dimensions to data management, including real-time validation rules, audit trails, and automated checks. While these technologies present opportunities for improved data integrity, they also require a structured approach to error detection and repair something that only a robust framework can provide (Love et al., 2021). It is not enough to depend on built-in software alerts or statistical outlier detection; qualitative assessment and contextual understanding are equally vital for effective data cleaning (Rana & Gupta, 2016).

In many resource-limited clinical settings, data is still managed by non-specialist staff who may not be adequately trained in database design or quality control principles. This gap underscores the need for frameworks that are not only technically sound but also user-friendly, scalable, and context-aware. By tailoring a data cleaning model to specific clinical research settings such as the DPSP study at Masafu Hospital researchers can improve the fidelity of the datasets they use and consequently enhance the credibility of their findings (Gesicho & Were, 2020).

To address these issues, this research focuses on the design and validation of a contextual framework for data cleaning in clinical research, with the DPSP malaria prevention study as a case.

This framework will incorporate both traditional and modern error detection techniques, outline best practices for missing data handling, and provide clear documentation guidelines to ensure transparency and reproducibility. The ultimate goal is to offer a structured, replicable, and context-sensitive approach that improves data quality in clinical research and supports the broader mission of delivering evidence-based healthcare solutions.

1.2 Theoretical Foundation

The theoretical foundation of this study is grounded in the principles of Data Quality Management (DQM) and the Information Systems Success Model (ISSM). These two frameworks provide a conceptual lens through which the challenges and solutions in clinical data cleaning can be understood and addressed, particularly in low-resource settings such as Masafu Hospital in Busia District.

Data Quality Management (DQM)

At its core, Data Quality Management (DQM) emphasizes the processes, standards, and technologies used to ensure that data is accurate, complete, timely, consistent, and fit for use (Ilyas & Chu, 2019). DQM proposes that poor-quality data not only undermines the reliability of analysis but also wastes resources and poses serious risks especially in clinical research where decisions directly affect human health.

The DQM model categorizes data quality into key dimensions: accuracy, completeness, validity, consistency, timeliness, and integrity (Calabrese, 2018). For instance, incomplete or duplicated records in clinical trials can skew outcome interpretations, while inconsistent formats may lead to data misclassification. These quality dimensions guide the identification of problems in clinical datasets and inform the development of cleaning strategies.

In the context of the DPSP clinical study, DQM principles help identify gaps such as inconsistent data formats, unvalidated entries, and incomplete patient profiles challenges that can be mitigated through a structured data cleaning framework tailored to the research environment.

Information Systems Success Model (ISSM)

Developed by DeLone and McLean and refined over time, the ISSM posits that system quality, information quality, and service quality are key antecedents of system use and user satisfaction, which in turn influence net system benefits (DeLone & McLean, 2016). In clinical research, the quality of the data (information quality) determines the effectiveness of research outcomes, while system quality determines the extent to which tools and platforms support efficient data cleaning.

When clinical researchers interact with electronic data capture (EDC) systems or data management platforms, their ability to produce clean, analyzable data depends on both the usability of the system and the quality of data it handles. As the ISSM implies, if these systems are poorly designed or lack error detection features, the users are unlikely to extract value from them hence the need for a contextual framework to guide effective data cleaning and quality assurance.

By integrating DQM and ISSM, this study proposes a contextual data cleaning framework that acknowledges both technical and human factors in clinical data management. This synthesis provides a strong theoretical basis for designing solutions that improve data quality while accommodating the realities of clinical research environments in Uganda and similar contexts.

1.3 Statement of the Problem

Clinical research depends fundamentally on the accuracy and reliability of collected data. Yet, a persistent challenge remains: poor data quality arising from inadequate or non-contextual cleaning practices. In many low- and middle-income country (LMIC) settings, clinical data is collected manually or semi-digitally across multiple sites, often by staff with limited technical training. This increases the risk of inconsistencies, missing values, duplicates, and invalid entries, undermining the integrity of research outcomes (Hoxha et al., 2022).

At Masafu Hospital in Busia District, the DPSP study on malaria prevention in pregnancy exemplifies this challenge. Data from repeated patient visits, heterogeneous sources, and inconsistent entry formats make error detection and correction especially difficult. Yet, cleaning is still performed in an ad hoc, manual manner, without standardized guidelines (Kalema & Kibukamusoke, 2021). This leads to delays, higher costs, and risks to scientific validity.

While several frameworks and methodologies for data cleaning exist, most are designed for highly digitized environments with robust infrastructure and advanced software (e.g., NADEEF, HoloClean). These approaches are unsuitable for rural African settings due to:

- reliance on large reference datasets,
- computational demands, and
- lack of adaptability to local workflows (Rekatsinas et al., 2017; Yakubu et al., 2021).

Thus, the theoretical gap lies in the absence of a context-specific, theory-informed framework that synthesizes principles from data quality management, information systems, and socio-technical theory to guide data cleaning in LMIC clinical research (Rahm & Banek, 2021; Yakubu et al., 2021). Without such a framework, clinical studies risk producing inaccurate, non-reproducible findings.

This study therefore seeks to design and validate a contextual data cleaning framework that is both technically rigorous and operationally feasible in resource-constrained clinical settings like Masafu.

1.4 Objectives of the Study

General Objective

To design and validate a contextual framework for data cleaning in clinical research, using the DPSP study at Masafu Hospital as a case study.

Specific Objectives

1. To synthesize and critique existing data cleaning frameworks and methodologies, highlighting their strengths and limitations in low-resource clinical research environments.
2. To design a contextual data cleaning framework informed by data quality theory, information systems success models, and socio-technical perspectives.
3. To validate the effectiveness and usability of the proposed framework using real-world clinical data and expert stakeholder feedback.

1.5 Research Questions

This study seeks to answer the following questions:

1. **RQ1:** What frameworks and methodologies for data cleaning currently exist, and what are their strengths and limitations when applied in low-resource clinical research environments such as Masafu Hospital?
2. **RQ2:** How can a contextual data cleaning framework be designed, drawing upon data quality theory, information systems success models, and socio-technical perspectives, to address the specific challenges of clinical research datasets in Uganda?
3. **RQ3:** To what extent is the proposed contextual framework effective and usable when applied to real-world clinical data from the DPSP study, and how do key stakeholders (data managers, Ministry of Health experts, EMR developers, and statisticians) evaluate its performance?

1.6 Significance of the Study

This study holds significant value for clinical researchers, data managers, health informatics professionals, and policy-makers, particularly in low- and middle-income countries where health data quality remains a critical challenge. The research offers a practical and context-aware framework for enhancing data cleaning processes in clinical research an area that is often neglected or inadequately handled in practice.

First, by systematically evaluating existing methodologies and their shortcomings, the study contributes to the body of knowledge on data quality and data cleaning in clinical contexts. Second, the proposed framework supports researchers in reducing the time, effort, and technical skill traditionally required for data cleaning, while ensuring transparency, consistency, and reproducibility in handling clinical datasets.

Moreover, by applying the framework to data from the DPSP study at Masafu Hospital, the study provides a real-world application and proof of concept that can be replicated or adapted for other clinical research projects across Uganda and similar regions. This not only enhances the scientific

integrity of such studies but also contributes to improved healthcare outcomes by ensuring that clinical decisions are based on clean, accurate, and reliable data.

1.7 Originality of the Study

This study is original in situating clinical data cleaning squarely within the domain of clinical data quality management and health informatics in low- and middle-income country (LMIC) contexts, guided by socio-technical theory. Unlike prior approaches designed for highly digitized, resource-rich environments, the proposed framework is tailored to the realities of low-resource hospitals such as Masafu, where data are captured through heterogeneous and semi-manual processes.

The originality of the study lies in three aspects. First, it integrates data quality theory, information systems success models, and socio-technical perspectives into a single contextual model, an approach not yet applied in LMIC clinical research. Second, it moves beyond purely theoretical contributions by validating the framework with real-world clinical data from the DPSP study, ensuring both relevance and applicability. Third, it demonstrates how automated yet user-adaptable cleaning processes can improve accuracy, completeness, and consistency without requiring advanced programming expertise.

Through this dual theoretical and practical positioning, the study contributes a novel, evidence-based, and context-aware framework that advances scholarly discourse while also addressing urgent clinical research challenges in under-resourced environments.

1.8 Justification of the Study

Despite the central role of data quality in determining the outcomes of clinical research, structured and contextualized data cleaning frameworks are lacking, especially in rural and resource-constrained settings. Most existing approaches are designed for technologically advanced environments and are often unsuitable or unsustainable for use in low-income health systems (Gesicho & Were, 2020; Acharya et al., 2023).

The DPSP study, which focuses on preventive malaria treatment in pregnancy, involves large-scale clinical data collection with the potential for missing, inconsistent, or duplicated entries.

Without a standardized and well-documented approach to cleaning such data, results may be skewed, findings misinterpreted, and public health recommendations compromised.

This study is therefore justified on several grounds:

- It bridges the gap between generic data cleaning methods and context-specific needs in clinical research.
- It contributes a reproducible framework that supports both qualitative and quantitative data verification and correction.
- It empowers health researchers, even with limited technical expertise, to manage and clean datasets systematically, reducing reliance on external consultants or costly tools.

Ultimately, the justification for this research lies in its practical utility, contextual relevance, and potential for impact both academically and clinically. It supports the broader agenda of data-driven healthcare improvement, especially in underrepresented research environments like Masafu Hospital in Busia District.

1.9 Scope of the Study

The scope of this study delineates the boundaries within which the research was conducted. It defines the geographical location, the thematic content under investigation, and the timeframe during which the research activities took place. These elements help focus the research and ensure relevance, manageability, and contextual accuracy of the outcomes.

1.9.1 Geographical Scope

This study was geographically limited to Masafu Hospital, located in Busia District, Eastern Uganda. Masafu Hospital is a key government health facility serving both urban and rural populations in the border region between Uganda and Kenya. The hospital was selected as the case site due to its involvement in the DPSP (Dihydroartemisinin–Piperaquine plus Sulfadoxine–Pyrimethamine) clinical study aimed at evaluating preventive treatment for malaria in pregnancy.

This location presents a suitable context for studying data quality challenges in clinical research because of its mixed infrastructure, varied staffing capacity, and reliance on semi-manual data

collection methods. The findings from Masafu are therefore relevant not only to the study site but also to other similar low-resource clinical research environments across Uganda and Sub-Saharan Africa.

1.9.2 Content Scope

The content scope of the study focused on the design, development, and validation of a contextual framework for data cleaning in clinical research. The research specifically examined:

- The challenges and limitations of existing data cleaning methodologies in clinical settings.
- The core components of a tailored data cleaning framework, including error detection, data repair, handling missing data, documentation practices, and the integration of appropriate tools.
- The evaluation and validation of the proposed framework using real-world data from the DPSP clinical trial at Masafu Hospital.

The study did not cover other aspects of clinical trial operations such as patient recruitment, drug administration, or statistical modeling of treatment effects. Instead, it was limited to the data cleaning phase a critical step in the data management lifecycle that ensures data readiness for statistical analysis and reporting.

Furthermore, the study integrated principles of data quality management, clinical data standards, and information systems theory, but it did not aim to compare the proposed framework against proprietary commercial platforms due to accessibility limitations.

1.9.3 Time Scope

The timeframe for this study spanned from August 2024 to July 2025, covering all essential phases of academic research—from conceptualization to framework development and report submission. This period allowed for an iterative yet structured progression through both theoretical and practical components of the study. Key activities and their time allocations are summarized below:

- Concept Development and Topic Defense (Aug 2024): Initial formulation of the research problem, objectives, and justification.

- Comprehensive Literature Review (Aug–Nov 2024): In-depth review of existing work, including related frameworks and contextual challenges in clinical data cleaning.
- Proposal Writing, Approval, and Submission (Dec 2024): Finalization of the research design and ethical clearance processes.
- Additional Literature Review and Tool Refinement (Jan–Mar 2025): Focused literature update and alignment with data collection tools.
- Data Collection and Analysis (Apr–May 2025): Gathering of primary and secondary data, including stakeholder interviews and DPSP dataset processing.
- Framework Development and Implementation (Jun–Jul 2025): Coding, refining, and testing the proposed contextual data cleaning framework.

Table 1.2: Research Activity Timeline

No.	Research Activity	Start Date	End Date
1	Concept Development & Topic Approval	1st Aug 2024	14th Aug 2024
2	Literature Review and Gap Identification	15th Aug 2024	15th Nov 2024
3	Proposal Writing, Approval & Submission	15th Dec 2024	28th Dec 2024
4	Additional Literature Review & Tool Design	9th Jan 2025	9th Mar 2025
5	Data Collection and Preliminary Analysis	1st Apr 2025	31st May 2025
6	Framework Development and System Implementation	1st Jun 2025	31st Jul 2025
7	Final Report Writing and Submission	1st Aug 2025	30th Aug 2025

1.10 Operational Definitions of Key Terms

- Data Cleaning: The process of detecting and correcting (or removing) inaccurate, incomplete, inconsistent, or duplicated data within a dataset to improve its quality and reliability.
- Clinical Research Data: Information collected during the implementation of clinical studies or trials, including patient demographics, laboratory results, medication records, and follow-up information.

- Contextual Framework: A structured representation that considers environmental, technological, and human factors unique to a specific setting, in this case, clinical research at Masafu Hospital.
- Data Quality: A measure of the condition of data based on factors such as accuracy, completeness, reliability, consistency, and relevance to the intended use.
- Error Detection: The process of identifying anomalies or violations in data, such as duplicates, typos, or missing values, through manual or automated techniques.
- Data Governance: Policies, procedures, and practices used to ensure proper data management, accountability, documentation, and compliance.

CHAPTER TWO

LITERATURE REVIEW

2.1 Introduction

The purpose of this chapter is to provide a critical review of existing literature on data cleaning practices, challenges, and innovations within the realm of clinical research. As the healthcare sector increasingly depends on data-driven decisions, particularly in clinical research, the integrity and quality of that data become paramount. Data cleaning, a vital step in the data preparation process, ensures that datasets are accurate, complete, consistent, and reliable for downstream analytics and decision-making. However, given the complexity of clinical datasets often characterized by high dimensionality, heterogeneous formats, and sensitive variables the data cleaning process must be contextual, dynamic, and adaptable to varying scenarios, especially in resource-constrained environments like Masafu Hospital.

This chapter is structured to explore key theoretical underpinnings that inform data quality and management, followed by an extensive discussion on the methodologies, tools, and frameworks currently employed in data cleaning. It critically examines their applicability and limitations, especially in the context of low- and middle-income countries (LMICs). Literature related to each of the specific study objectives is reviewed to position the proposed framework within the broader scientific conversation. The chapter concludes by identifying gaps in existing studies and articulating the contribution of this research in addressing those gaps.

Through this review, the study builds a solid foundation for designing a contextual data cleaning framework tailored for clinical research settings in Uganda. The insights presented herein not only inform the development of the proposed solution but also illuminate the pressing need for adaptive, context-aware tools that transcend generic, one-size-fits-all approaches.

2.2 Theoretical Framework

A sound theoretical foundation is crucial for guiding the design and implementation of any data-driven intervention. For this study, the underpinning is drawn from four complementary theories: Total Data Quality Management (TDQM), Information Quality Theory (IQT), Knowledge

Representation Theory (KRT), and Socio-Technical Systems (STS) Theory. Together, these offer a multidimensional understanding of data quality, semantics, and organizational realities in health information systems.

2.2.1 Total Data Quality Management (TDQM)

TDQM (Wang, 1998) positions data as a product, emphasizing continuous quality assessment across its lifecycle. The framework is operationalized through a closed-loop cycle of definition, measurement, analysis, and improvement (Pradhan et al., 2021). In healthcare research, this principle is vital because inaccurate data threatens both patient safety and scientific validity.

Fig 2.1: Total Data Quality Management (TDQM), (source: Pradhan, 2021)

Applied to clinical research in LMICs, TDQM underscores the need for structured yet adaptable cleaning frameworks that incorporate user feedback and contextual constraints rather than one-time corrections (Batini & Scannapieco, 2016).

2.2.2 Information Quality Theory (IQT)

IQT (Lee et al., 2023) introduces measurable dimensions of information quality: accuracy, completeness, consistency, timeliness, and relevance. These dimensions have been widely validated in healthcare (Phan et al., 2020; Shi, 2021), where poor quality data directly affects diagnoses, treatment outcomes, and policy formulation.

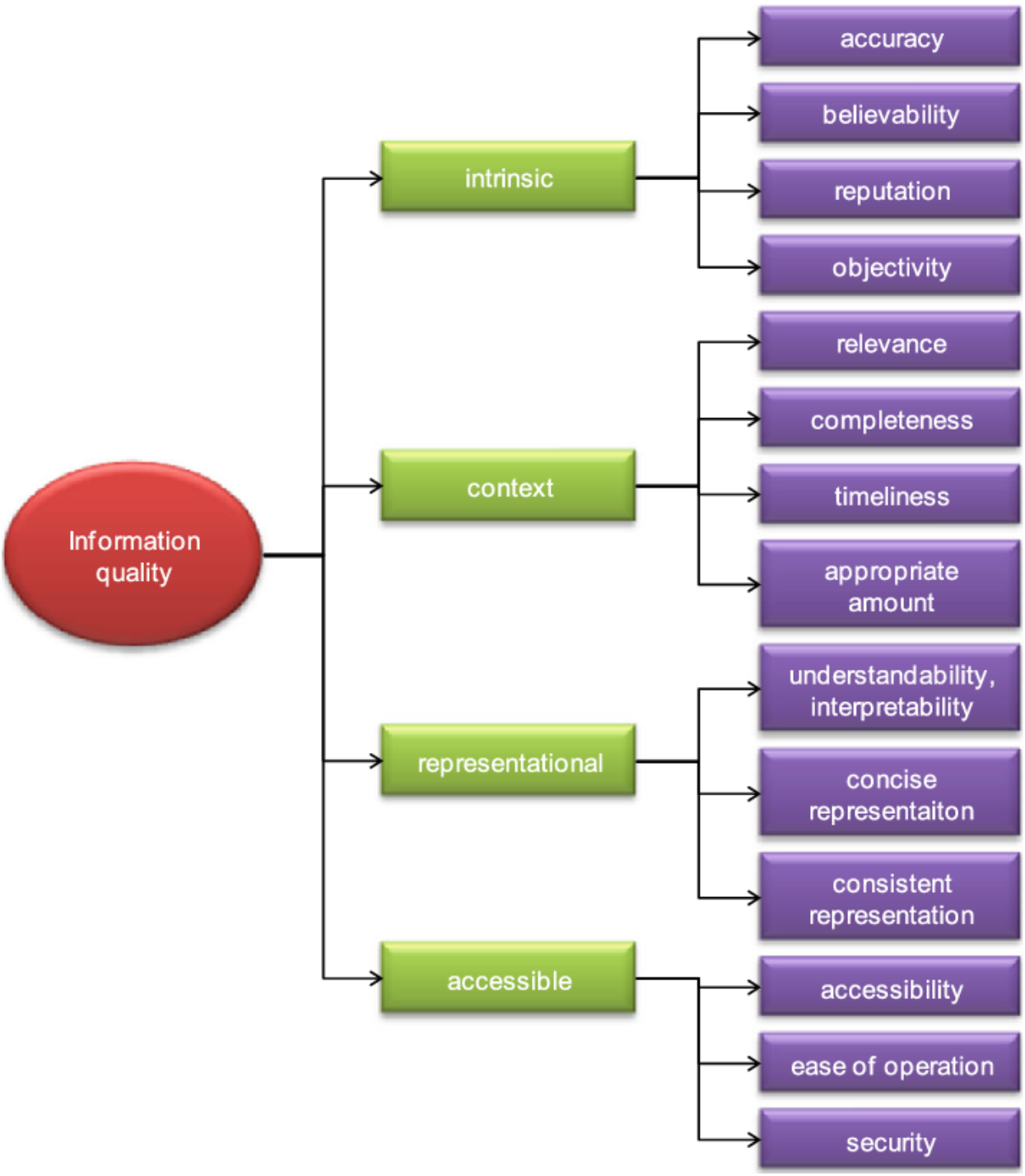


Fig 2.2: information quality theory,(source: Lee et al., 2023).

For this study, IQT provides the evaluation lens to judge the adequacy of cleaning processes. In resource-constrained settings, it also highlights the subjective dimension of quality: whether

cleaned data is usable and trusted by its intended users, including clinicians, researchers, and policymakers.

2.2.3 Knowledge Representation Theory (KRT)

KRT focuses on how information is encoded, interpreted, and transformed into actionable knowledge (Wong, 2016; Miao et al., 2023). Clinical datasets are highly semantic, containing rules about permissible values (e.g., age ranges, drug-dosage mappings).

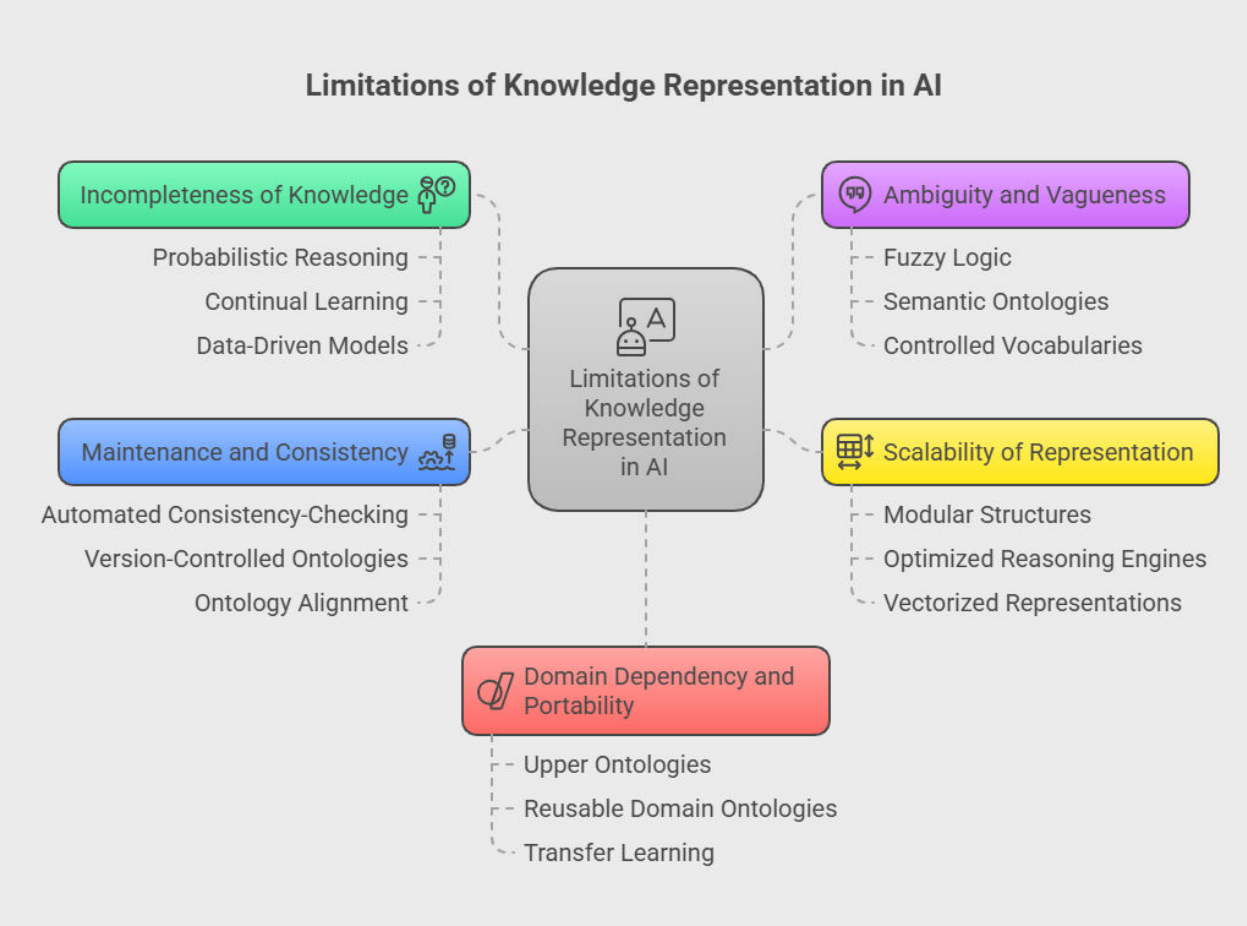


Fig 2.3. Knowledge representation theory (source: Miao et al, 2023)

Frameworks such as NADEEF and Katara leverage rule-based or denial-constraint approaches (Elmagarmid et al., 2021), yet studies show that weak knowledge representation leads to either over-cleaning (valid outliers removed) or under-cleaning (errors retained). For this study, KRT

ensures that cleaning rules embed domain-specific semantics, producing data that is not only syntactically valid but also clinically meaningful.

2.2.4 Socio-Technical Systems (STS) Theory

STS theory (Mumford, 2016) stresses that technology adoption and outcomes depend on the interplay between people, processes, and technology. In LMIC hospitals like Masafu, infrastructural limitations, fragmented workflows, and low digital literacy contribute to error generation and constrain the feasibility of purely automated cleaning.

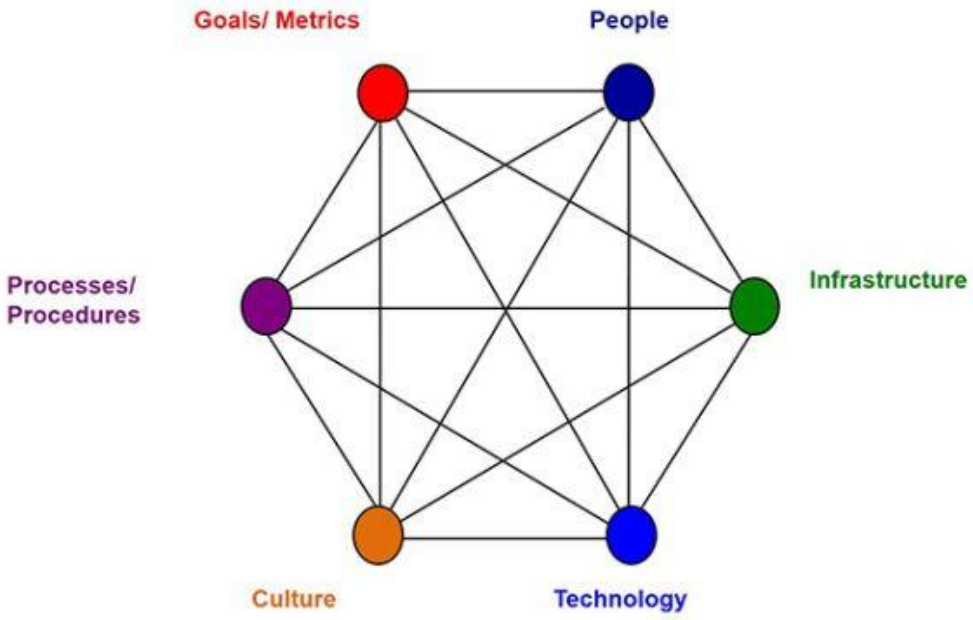


Fig 2.4: Socio-Technical Systems Theory

Therefore, STS theory anchors this study's insistence that a data cleaning framework must be user-centered, process-aware, and organizationally aligned, rather than focusing solely on algorithmic sophistication.

2.2.5 Synthesis of Theories and Gap Identification

Individually, each theory offers a useful lens:

- TDQM → continuous improvement cycle.
- IQT → operational quality dimensions.
- KRT → semantic rigor.
- STS → socio-organizational feasibility.

However, when we examine existing frameworks through this combined lens, their limitations become clear:

Framework	Strengths (Theory Fit)	Limitations (Gap)
NADEEF (rule-based)	Strong in KRT (rule encoding, denial constraints)	Ignores STS constraints → computationally heavy, needs skilled programmers.
HoloClean (probabilistic ML)	Captures TDQM cycle (automated improvement)	Assumes large reference datasets, unsuitable for LMICs with sparse data.
Potter's Wheel (interactive UI)	Usability aligns with STS (human-in-loop cleaning)	Manual-intensive, lacks scalability, no IQT-based evaluation metrics.
Katara / Llunatic	Blend KRT + automation	High infrastructure demand, limited adaptability to clinical workflows.

None of these frameworks simultaneously integrate all four theoretical perspectives. Technical solutions (NADEEF, HoloClean) fail to adapt to socio-technical realities. Usability-focused tools (Potter's Wheel) lack scalability and theoretical grounding in continuous improvement or semantic accuracy.

Thus, the knowledge gap is the absence of a context-specific, theory-informed framework that balances:

- TDQM's rigor,
- IQT's evaluative metrics,

- KRT's semantic validity, and
- STS's contextual feasibility.

This gap provides the precise foundation for the contextual framework developed in this study.

2.3 Conceptualization of Data Cleaning

2.3.1 Definitions and Importance in Clinical Research

Data cleaning, also known as data cleansing or data scrubbing, refers to the process of detecting and correcting (or removing) corrupt, inaccurate, incomplete, or irrelevant parts of a dataset, with the goal of improving data quality for analysis (Rahm & Do, 2020). In clinical research, where data often influence life-altering decisions, this process assumes a critical role. According to Muthuraman (2021), the integrity of data collected from patient records, clinical trials, or hospital management systems significantly affects outcomes such as drug efficacy studies, epidemiological tracking, and health policy formulation.

The importance of data cleaning in clinical research lies in its ability to ensure accuracy and reproducibility of findings. Inaccurate data can mislead researchers, produce faulty analyses, and compromise patient safety (Phan et al., 2020). Moreover, the sensitivity and variability of clinical data demand that cleaning be context-aware and domain-informed especially when involving measurements such as dosage levels, lab results, or anthropometric readings. Therefore, data cleaning is not merely a technical step; it is an ethical necessity in any research that impacts human health (Hossain, 2019).

Recent systems such as NADEEF and HoloClean attempt to automate this process but have been critiqued for insufficient context-awareness, especially in LMICs. In these environments, poorly digitized records, limited infrastructure, and incomplete documentation make conventional cleaning algorithms ineffective. As such, the design of an effective data cleaning framework must go beyond automation and embrace adaptability, clinical domain knowledge, and usability in low-resource settings.

2.3.2 Dimensions of Data Quality: Accuracy, Consistency, Completeness, Validity

According to Lee et al. (2023), four dimensions are critical to assessing data quality, particularly in clinical contexts:

- Accuracy refers to how well data represents the real-world values it is supposed to model. Inaccurate blood pressure readings, for example, can affect treatment plans or clinical trial outcomes.
- Consistency deals with the logical coherence of data across time and systems. Inconsistent patient records such as varying age or gender attributes across datasets can disrupt longitudinal studies and patient tracking (Miao et al., 2023).
- Completeness assesses whether all required data attributes are present. Missing lab results, for instance, can obscure disease trends and compromise diagnosis algorithms (Wong, 2016). Missing data is also a key concern in machine learning applications, where gaps reduce model reliability.
- Validity checks if the data conforms to expected formats, ranges, and types. For example, entering "180 kg" for a 2-year-old child would violate validity constraints unless flagged or corrected.

Several frameworks, such as TDQM and IQT, align these dimensions with iterative quality assurance loops. However, existing literature (Gudivada et al., 2017; Ridzuan et al., 2019) shows that many tools focus disproportionately on accuracy and completeness, with less emphasis on validity and consistency. This partial implementation is inadequate for clinical datasets, where even minor inconsistencies can yield dangerous conclusions.

2.3.3 Characteristics of Clinical Data

Clinical data is distinctively complex, multidimensional, and sensitive. It encompasses diverse formats, including:

- Electronic Medical Records (EMRs): Structured (e.g., numeric lab results) and unstructured (e.g., physician notes). EMRs are often riddled with transcription errors, missing timestamps, and inconsistent coding schemes (Phan et al., 2020).
- Sensor and Wearable Device Data: Generated continuously, high-volume, and often unverified. Data streams from cardiac monitors or fitness trackers require specialized cleaning approaches for noise and signal interference (Lane et al., 2023).
- Lab Results and Imaging Data: Require strict unit standardization and range validation. For example, millimoles/litre (mmol/L) and milligrams/decilitre (mg/dL) must be clearly defined to avoid misinterpretation.

Moreover, clinical data is usually hierarchical and relational patients, visits, medications, and diagnoses are linked. These relationships must be preserved during cleaning to avoid logical inconsistencies (Shi, 2021).

Finally, privacy regulations (like HIPAA or GDPR) limit direct access to raw clinical data, complicating external tool integration and necessitating secure, auditable cleaning procedures.

2.4 Review of Literature Related to Research Objectives

2.4.1 Existing Data Cleaning Methodologies in Clinical Research

A wide array of data cleaning approaches have been developed, ranging from rule-based systems to machine learning-based frameworks. However, their application in clinical settings, especially in LMICs, remains limited by issues of scalability, domain relevance, and usability.

a) Rule-Based Approaches

Rule-based systems rely on pre-defined logic to detect and repair anomalies. NADEEF (Bohannon et al., 2017) is a pioneering platform that allows users to define data quality rules using denial

constraints. It supports a modular design where rules can be specified independently, offering flexibility for various domains.

Katara (Yakout et al., 2017) adds a novel twist by integrating knowledge bases (e.g., DrugBank, SNOMED CT) and crowdsourced validations. This hybrid approach enables systems to "learn" from prior corrections, enhancing the contextual relevance of repairs.

Llunatic takes rule-based logic further by generalizing functional dependencies. It explores denial constraints through machine-learned refinement, making it more adaptive to evolving data types (Neutatz, 2019). However, these systems assume the availability of clean reference data or domain experts luxuries often absent in low-resource environments.

b) Machine Learning and Probabilistic Methods

In response to the rigidity of rule-based systems, ML-based methods like HoloClean (Rekatsinas et al., 2018) and ERACER (Das et al., 2020) emerged. These frameworks model the joint probability distribution of observed and true values, using features like co-occurrence, domain ranges, and patterns to impute or correct values.

ActiveClean adds a user-guided layer, where users define cleaning actions and the model learns optimal cleaning strategies from past tasks (Kraska et al., 2019). While these systems demonstrate higher adaptability and accuracy, they require extensive computational resources, labeled datasets, and fine-tuning barriers in LMIC settings.

c) Frameworks for Missing Values, Outliers, and Inconsistencies

Various systems specifically target common clinical data issues:

- Missing Values: Techniques include mean/mode substitution, regression-based imputation, and Bayesian methods (Gudivada et al., 2017). While effective, their accuracy depends heavily on variable distribution and auxiliary information.
- Outliers: Traditional methods (e.g., z-score, IQR) often fail in clinical contexts due to non-Gaussian distributions. Van den Broeck et al. (2016) propose a variable-specific strategy, but it lacks scalability for big data environments.

- Inconsistencies: Tools like Potter's Wheel (Raman & Hellerstein, 2018) provide visual interfaces to detect and repair inconsistencies. However, they remain manual-intensive and ill-suited for large clinical datasets.

d) Limitations in LMICs

In low-resource settings, challenges such as unstable electricity, limited IT infrastructure, and fragmented data collection practices render many existing frameworks inapplicable. Systems like NADEEF or HoloClean presume structured EMRs and centralized data warehouses, which are rare in rural hospitals (Pindya, 2024).

Moreover, generic cleaning rules often fail to consider local context, such as unusual but valid health patterns due to endemic diseases or regional diet. A rule flagging "extremely low BMI" may misclassify patients in food-insecure regions unless localized context is encoded.

2.5 Review of Existing Frameworks and Tools

This section offers a structured and evaluative discussion of major data cleaning tools and frameworks relevant to clinical research, particularly in low-resource settings like Masafu Hospital. The review integrates global benchmarks while contextualizing findings to the Data Processing Support Programme (DPSP) study.

2.5.1 Rule-Based Tools (NADEEF, Katara, OpenRefine)

Rule-based tools rely on explicit data quality rules defined by users to identify, flag, and resolve inconsistencies. These tools are generally interpretable, flexible in logic, and powerful when the rules are domain-specific and data formats are known in advance.

NADEEF (Now Advanced Data cleaning platform with Extensibility and Flexibility)

NADEEF, developed by Bohannon et al., is a general-purpose platform that treats data quality rules (like denial constraints, functional dependencies, etc.) as first-class citizens. Its extensibility allows users to define and customize cleaning logic, making it well-suited for structured tabular data (Bohannon et al., 2017). In clinical research, however, the challenge lies in defining complete

and context-aware rules. For example, denial constraints may not capture culturally specific data ranges (e.g., normal BMI in malnourished populations), potentially leading to false flags.

Katara

Katara (Yakout et al., 2017) represents a step forward in integrating knowledge bases and crowdsourcing for data repair. It maps dataset entries against trusted external repositories like SNOMED CT or DrugBank. By suggesting the top-k most likely corrections, it enables both automated and manual cleaning. However, Katara assumes availability of reliable online ontologies and human validators resources often scarce in LMICs. Moreover, cultural and linguistic variability in clinical data (e.g., local terminology) may hinder alignment with global standards.

OpenRefine

OpenRefine (formerly Google Refine) is a user-friendly open-source tool for exploring, cleaning, and transforming messy data. It offers clustering algorithms for deduplication and supports regex-based cleaning operations. Despite its popularity, OpenRefine lacks robust integration with medical ontologies and cannot handle real-time, high-volume clinical streams. Furthermore, its manual interface makes it impractical for large-scale clinical trials with hundreds of variables per patient.

2.5.2 Machine Learning-Based Tools (HoloClean, ERACER, ActiveClean)

These tools model data cleaning as a learning problem, using statistical inference and machine learning to suggest repairs or impute missing values.

HoloClean

HoloClean (Rekatsinas et al., 2018) uses a probabilistic framework to reason over data integrity constraints, external knowledge, and statistical co-occurrence patterns. It excels in identifying hidden inconsistencies and imputing missing data using machine learning models. Yet, HoloClean's "one-shot inference" approach struggles with incremental learning and online

datasets common in clinical settings where data is updated continuously. Moreover, the model demands extensive tuning and large datasets, which may not be available in smaller health centers.

ERACER

ERACER is tailored to clean entity resolution errors using a probabilistic model that integrates prior information. It improves over manual deduplication by automatically learning match patterns.
However, ERACER assumes high-quality training data and robust data dictionaries, both of which are rare in clinical research in LMICs. Additionally, it's designed more for record linkage than full-spectrum cleaning (i.e., outlier handling or type correction).

ActiveClean

ActiveClean (Kraska et al., 2019) is a semi-supervised framework where the user specifies cleaning operations and labels small data portions. The system then generalizes and applies learned transformations. This is particularly useful in reducing cleaning costs in iterative ML pipelines. In a clinical context, ActiveClean could be valuable in tailoring cleaning rules to local datasets (e.g., using nurse-entered labels), but its reliance on user expertise is a bottleneck where informatics staff is limited.

2.5.3 Hybrid and Contextual Systems

These tools combine statistical, rule-based, and human-in-the-loop strategies to increase cleaning effectiveness.

SCARE (Scalable Cleaning Architecture using Rule Engine)

SCARE is designed for large-scale heterogeneous data cleaning using a hybrid architecture. It incorporates both user-defined rules and machine-learned transformations. SCARE supports pipelining of tasks useful for repeated clinical trials with similar data structures.
Its limitation is the steep learning curve and absence of domain-specific modules (e.g., for ICD-10 or HL7 data types), which hinders adoption in clinical research.

Potter's Wheel

Potter's Wheel (Raman & Hellerstein, 2018) is one of the earliest interactive data cleaning systems. It allows users to define transformations through a spreadsheet-like interface. Its ability to catch "nested discrepancies" (e.g., a wrong unit buried within a text field) makes it highly flexible.

However, it remains mostly manual, slow for big datasets, and ineffective for real-time stream cleaning an emerging need in sensor-heavy clinical trials and remote health monitoring.

Contextual Imputation Frameworks

Recent work by Pindya (2024) and Miao et al. (2023) focuses on contextualized frameworks where cleaning is guided by local rules, domain logic, and empirical patterns. These frameworks emphasize practical repair strategies (e.g., imputing based on co-morbidities), but are still experimental and often lack scalability or generalizability.

2.5.4 Limitations in Clinical Research Settings

While the reviewed tools showcase considerable technical merit, several limitations emerge in clinical research applications, especially in low-resource environments:

- Domain Complexity: Clinical data often involves interdependent variables, nested time-series, and hierarchical patient relationships. Most cleaning tools are designed for flat tabular data and cannot accommodate complex EMR schemas (Wong, 2016; Lane et al., 2023).
- Context Insensitivity: Rule-based tools lack flexibility for contextual data interpretation (e.g., distinguishing outliers from rare but valid patient conditions), especially in LMICs where normative ranges may differ.
- Infrastructure Gaps: High memory requirements, computational overhead, and lack of interoperability with local health information systems make many tools difficult to deploy in settings like Masafu Hospital (Phan et al., 2020).

- Language and Semantics: Tools reliant on English medical ontologies face challenges in parsing locally used terms, abbreviations, or transliterations common in Ugandan clinical documentation.
- User Dependence: Semi-supervised or manual systems (e.g., OpenRefine, ActiveClean) presume skilled users, which is not always feasible in remote settings with overburdened health workers and limited training.

2.5.5 Insights from DPSP Study at Masafu Hospital (Contextual Fit)

The Data Processing Support Programme (DPSP) at Masafu Hospital provides a real-world lens through which the limitations and potential adaptations of existing tools can be observed.

Key insights include:

- Data Diversity: DPSP datasets included structured fields (e.g., age, weight) and unstructured data (e.g., diagnosis notes), making cleaning a multi-step and context-specific task. No existing tool fully supported this duality without heavy customization.
- Power Interruptions & Offline Needs: Tools requiring constant internet access or cloud-based validation (e.g., Katara) were impractical. The DPSP cleaning team needed tools that could run locally and offline.
- Anthropometric Nuances: The DPSP highlighted the risk of misclassifying low BMI values as outliers due to standard thresholds developed in Western contexts. This emphasized the importance of contextual thresholds and localized logic in cleaning tools.
- Human–Tool Collaboration: While manual reviews were essential, many tools lacked mechanisms to integrate clinical feedback loops. For instance, certain outliers were justified by patient condition (e.g., terminal illness), but flagged incorrectly.
- Training Burden: DPSP participants found tools like OpenRefine useful but struggled with regex logic and advanced clustering. Simpler, guided interfaces were preferred indicating a demand for low-code/no-code solutions.

These insights affirm the need for a contextual, scalable, and user-friendly framework, tailored for clinical environments in rural Uganda and similar settings.

Table 2.1 Summary of the reviewed tools and frameworks

Tool/Framework	Approach	Strengths	Limitations in Clinical Context
NADEEF	Rule-based	Extensible, modular	Needs complete, accurate rules
Katara	Rule + Knowledge-base	Ontology integration	Requires internet, assumes global standards
OpenRefine	Manual / Clustering	Easy-to-use, visual	Limited automation, lacks medical domain support
HoloClean	ML / Probabilistic	High accuracy	High resource use, needs tuning
ERACER	ML / Entity Resolution	Reduces deduplication errors	Not designed for multi-modal clinical data
ActiveClean	Semi-supervised	Learns from user input	Needs trained users, slow scaling
SCARE	Hybrid	Scalable pipelines	Steep learning curve
Potter's Wheel	Manual	Captures subtle errors	Not scalable, outdated interface
Contextual Frameworks	Custom	Tailored to clinical needs	Still experimental, lacks automation

2.6 Challenges in Clinical Data Cleaning

Cleaning clinical research data is an indispensable step toward ensuring data reliability, reproducibility, and safety in evidence-based healthcare. However, the process is riddled with domain-specific and context-sensitive challenges that extend beyond those found in conventional datasets. This section critically explores key challenges encountered in clinical data cleaning, especially within low- and middle-income countries (LMICs) such as Uganda.

2.6.1 High-Stakes Data Implications in Health Research

In clinical research, data errors are not just a matter of analytical inefficiency they carry life-and-death implications. Mislabeling a patient's condition, misrepresenting dosage levels, or ignoring critical anomalies due to poor data quality can result in unsafe treatment recommendations, flawed clinical conclusions, or regulatory penalties (Lane et al., 2023; Muthuraman, 2021).

For example, incorrect clinical trial results caused by undetected data inconsistencies may lead to the approval of ineffective drugs or, worse, the rejection of life-saving interventions. Moreover, in trials involving vulnerable populations (e.g., infants or HIV/AIDS patients), any compromise in data quality can translate to direct harm. Hence, data cleaning is not merely a technical step it is a public health safeguard.

Furthermore, health data is used for policy decisions and donor funding allocations. Inaccuracies in morbidity statistics can misguide national resource distribution. This emphasizes the urgent need for robust, transparent, and context-sensitive data cleaning strategies.

2.6.2 Handling Missing Data and Imputation Techniques

Missing data is perhaps the most pervasive challenge in clinical datasets. Causes range from patient non-response and equipment malfunction to transcription errors. Common missingness types include Missing Completely at Random (MCAR), Missing at Random (MAR), and Missing Not at Random (MNAR) each requiring different imputation strategies (Hossain, 2019; Gudivada et al., 2017).

Several imputation methods exist, such as:

- Mean/Median Imputation: Easy to apply but distorts variance.
- Hot-deck/Cold-deck Imputation: Uses values from similar records.
- Regression Imputation: Predicts missing values using other variables.
- Multiple Imputation by Chained Equations (MICE): Builds a model for each variable with missing data.

However, no single technique fits all clinical contexts. For instance, imputing blood pressure from age or weight might be statistically sound but medically unsound for hypertensive patients. Moreover, conventional tools may ignore logical constraints e.g., a male patient being imputed with a pregnancy status.

Context-aware imputation, such as the approach proposed by Pindya (2024), which uses auxiliary variables and edit restrictions, is more suitable for clinical data but remains complex and underutilized in low-resource settings.

2.6.3 Detection and Correction of Outliers

Outliers in clinical datasets can signal:

- Measurement or entry errors,
- Rare conditions,
- Data transformation issues, or
- Valid but extreme patient responses.

Traditional statistical outlier detection (e.g., z-scores, IQR methods) assumes Gaussian distributions, which are rarely applicable to clinical data. For example, blood glucose levels in diabetic patients may skew right, invalidating symmetric assumptions.

Van den Broeck et al. (2020) proposed variable-specific categorization: true normal, true extreme, erroneous, and idiopathic. While innovative, this approach relies heavily on expert input and is hard to automate at scale.

Moreover, outlier correction decisions must balance statistical and clinical logic. Deleting a low-birth-weight record because it seems unusual might erase a valid and medically important data point. Contextual frameworks those integrating local medical standards are critical but scarce.

2.6.4 Multi-modal and Hierarchical Data Structures

Unlike flat spreadsheets, clinical data is often:

- Hierarchical: Repeated measures per patient (e.g., blood tests over time),
- Multi-modal: Text, images, structured and semi-structured formats,
- Nested: Lab results nested within visits, which are nested within patients.

Standard cleaning tools like OpenRefine or Excel cannot handle these data architectures. Electronic Medical Records (EMRs) often involve HL7 FHIR structures, complex ontologies, and time-series relationships that require advanced parsing logic and modular cleaning pipelines (Miao et al., 2023).

For instance, a fever recorded in Celsius and Fahrenheit for the same patient can't be resolved by simple value-based cleaning. A framework must account for unit conversions, context switches, and temporal coherence (e.g., blood pressure before vs. after treatment).

Unfortunately, most frameworks reviewed (e.g., NADEEF, ERACER) are not equipped to handle multimodal or hierarchical inputs. Adopting graph-based data models and contextual ontologies may be part of the long-term solution.

2.6.5 Parameter Tuning, Trial-and-Error, and Workflow Iteration

Another challenge is that data cleaning is rarely a one-pass process. Cleaning rules often require:

- Threshold tuning (e.g., what's a "low" hemoglobin level?)
- Iterative validation (e.g., checking after imputation),
- Testing different logic sets based on clinical feedback.

This trial-and-error approach is time-consuming, especially where computational power is limited or staff have no programming expertise (Shi, 2021).

Tools like HoloClean require hyperparameter tuning and probabilistic thresholds. ActiveClean depends on user-labeled examples and thus needs domain experts on hand, which is often unrealistic in rural or under-resourced clinical settings.

The result is that cleaning workflows stagnate midway, and inconsistencies persist despite partial efforts. The absence of interactive or guided frameworks makes it hard for novice users to explore "what-if" cleaning scenarios safely and repeatably.

2.6.6 Governance, Documentation, and Compliance Challenges

Data cleaning isn't just technical it's a governance and compliance issue. Cleaned data must be auditable, traceable, and reproducible. This is especially important in clinical trials governed by GCP (Good Clinical Practice), FDA CFR 21 Part 11, and EMA standards (Ross et al., 2015; Idzwan, 2021).

However, many existing frameworks do not:

- Log each cleaning action,
- Track who performed what transformation,
- Capture user rationales behind corrections,
- Preserve raw vs. cleaned versions.

In LMICs, documentation practices are often manual and inconsistent, leading to data provenance gaps and making audits nearly impossible. This affects research credibility, data reuse, and cross-institutional collaborations.

Furthermore, compliance standards demand version control, consent verification, and access traceability features often missing in open-source tools or custom-built scripts.

Table 2.2 **Summary of Clinical Data Cleaning Challenges**

Challenge	Description	Critical Implication
High-Stakes Outcomes	Inaccuracies can mislead treatments or policies	Endangers lives and policy decisions
Missing Data	Requires nuanced, contextual imputation	May bias analysis or reduce validity
Outliers	Can be valid, erroneous, or rare	Misclassification may distort results
Multi-modal Structure	Nested data defies flat-file tools	Needs advanced frameworks
Trial-and-Error	Cleaning is iterative and resource-demanding	Requires expert time and tuning
Compliance & Documentation	Audits need full traceability	Missing logs hurt trust and reproducibility

2.7 Summary of Related Works

This section critically summarizes prominent frameworks and tools discussed earlier and situates their contributions in the context of clinical data cleaning, particularly in low-resource settings.

2.7.1 Comparative Overview of Frameworks and Findings

A wide range of data cleaning frameworks have emerged over the years, each with unique strengths and limitations. Rule-based systems such as NADEEF, Katara, and OpenRefine are excellent at capturing structural and logical inconsistencies through constraints or external knowledge bases. However, they often lack adaptive intelligence, rendering them less effective in dynamic or unfamiliar clinical contexts (Neutatz et al., 2019; Wong, 2016).

Conversely, machine learning-based tools like HoloClean, ActiveClean, and ERACER bring flexibility and learning capabilities to the table. These systems can generalize from examples, but typically require manual labeling, hyperparameter tuning, and access to robust computing

infrastructure not always feasible in LMIC hospitals such as Masafu. Also, their "black-box" nature limits explainability, which is crucial in clinical domains.

Hybrid models (e.g., SCARE, Potter's Wheel) attempt to bridge this gap by combining rule-based and probabilistic logic, but they still fail to integrate user preferences, domain context, and real-time feedback essential for clinical usability.

Most importantly, none of these frameworks were explicitly designed with Ugandan health research or LMIC constraints in mind, hence lacking adaptability to contextual realities such as intermittent electricity, limited staff capacity, and data captured on paper forms.

2.7.2 Contributions of Past Studies to the Current Study

The reviewed literature provides valuable building blocks. For instance:

- Miao et al. (2023) emphasized modular workflows and patching inconsistencies, which inform the design of your framework's repair engine.
- Phan et al. (2020) demonstrated that even narrowly-scoped cleaning (height/weight) improved data utility a concept extended here to multi-variable contexts.
- Neutatz (2019) introduced denial constraints and rule learning, which guide the construction of integrity enforcement components in your system.
- Shi (2021) underscored the value of context-aware constraints, a core foundation in your proposed approach.

These studies provide technical and theoretical insights, but not a fully contextualized solution for clinical research in Uganda or comparable LMICs a gap your work addresses.

2.7.3 Implications for Contextual Design in Low-Resource Environments

Several implications emerge:

1. Frameworks must be explainable, adaptable, and lightweight. Resource-limited settings cannot afford overly complex or opaque tools.
2. Domain expertise must be embedded in the logic, allowing researchers to configure cleaning rules aligned with local protocols.
3. Offline and manual override functionality must be available to accommodate erratic network access and limited training.
4. Ethical, traceable cleaning is non-negotiable clinical data is sensitive and must comply with local ethics and international standards (e.g., GCP, FDA Part 11).

In sum, a successful framework must strike a balance between intelligence and accessibility, enabling healthcare teams at institutions like Masafu Hospital to improve data quality without deep technical expertise.

2.8 Research Gaps Identified in Literature

Despite the progress discussed, several critical gaps persist:

2.8.1 Lack of Contextual Frameworks Tailored to Clinical Research in Uganda or Similar LMICs

Most existing frameworks have been developed in high-resource settings, assuming access to stable infrastructure, trained data scientists, and EMR-integrated databases. These assumptions rarely hold true in settings like Masafu Hospital, where manual data entry, hybrid record systems, and staff shortages are the norm. This limits the transferability and reliability of these frameworks.

2.8.2 Overdependence on Anthropometric Data and Static Rules

Many frameworks, like Phan et al. (2020), narrowly focus on specific variable types (e.g., height, weight), limiting their scope. Static rules further reduce responsiveness to contextual anomalies, such as regional variations in lab reference ranges. There's a need for more flexible, holistic solutions that can adapt to a broad set of clinical variables and dynamic research designs.

2.8.3 Absence of User-Guided and Explainable Cleaning Recommendations

Frameworks like HoloClean, though powerful, operate as black boxes. Clinical researchers must be able to see, interpret, and validate cleaning actions especially when data is used for high-stakes decisions. Explainability fosters trust, transparency, and compliance with clinical governance frameworks.

2.8.4 Underutilization of Integrity Constraints and Workflow Learning

While some tools implement simple constraint-based cleaning (e.g., denial constraints), most do not leverage relational integrity rules (e.g., patient gender vs. pregnancy status) or learn from past workflows to improve future cleaning operations. This makes repeated data cleaning inefficient, error-prone, and dependent on trial-and-error.

2.9 Contribution of the Study to Existing Knowledge

This study addresses the above gaps and makes several novel contributions to the field:

- Development of a context-sensitive, dynamic data cleaning framework tailored specifically for clinical research settings in Uganda.
- Integration of hybrid techniques combining rule-based logic (e.g., denial constraints, type validation) with lightweight machine learning for contextual prediction and repair.
- Inclusion of a human-in-the-loop component to ensure clinical relevance, usability, and explainability vital in sensitive health data.
- Design for constrained environments, supporting offline use, minimal computing resources, and user training in local contexts.

- Application and validation using the DPSP study at Masafu Hospital, making the findings empirically grounded and transferable to similar settings.

This contribution positions the framework as a practical tool and theoretical advancement in the discipline of data quality enhancement for clinical research in low-resource settings.

2.10 Table of Literature Gaps

Author(s)	Year	Study Title	Description / Focus	Identified Gap	How Current Study Responds
Van den Broeck et al.	2016	Outlier Detection in Primary Care Data	Defined variable-specific outliers	Not scalable, lacks automation	Develops automated, context-sensitive outlier handling
Phan et al.	2020	Automated Cleaning for EHR Data	Targeted anthropometric data (height/weight)	Limited to specific data types	Broadens framework to handle diverse clinical datasets
Miao et al.	2023	Patch-Based Data Cleaning Framework	Emphasized modular repairs and structural fixes	Lacks contextual logic and user-based recommendations	Introduces adaptive workflow generation with contextual intelligence
Shi	2021	Context-Aware Data Quality	Focused on rule validation and semantic constraints	No user guidance or customization strategy	Integrates domain-customizable rules with flexible logic layers
Neutatz et al.	2019	HoloClean and Rule Refinement	Proposed ML with probabilistic repair modeling	High setup complexity, lacks transparency	Proposes semi-automated, interpretable ML-rule hybrid for LMIC use cases

CHAPTER THREE:

METHODOLOGY

3.1 Introduction

This chapter outlines the systematic methodology adopted to guide the development and evaluation of a contextual framework for data cleaning in clinical research settings. The methodological structure is grounded in both philosophical and empirical considerations, ensuring that the study is rigorous, contextually relevant, and practically applicable. The aim is to explore, design, and validate a data cleaning framework that addresses the complex and multidimensional nature of clinical datasets, particularly in resource-constrained environments such as Masafu Hospital under the DPSP project.

To achieve this, the chapter elaborates on the research philosophy, approach, strategy, and methodological choices underpinning the study. It further describes the tools and techniques used for data collection, sampling methods, data sources, and validation processes. These components are carefully aligned with the research objectives to ensure logical consistency and robust findings.

The methodology was carefully aligned with the specific objectives of the study, as summarized in Table 3.1 below;

Table 3.1 Alignment of the methodology with the specific objectives

Objective	Data Source	Methods	Output
Obj.1. Synthesize existing frameworks	Secondary literature, policy documents	Systematic literature review & critical analysis	Analytical summary of gaps
Obj.2. Design contextual framework	DPSP dataset, interviews with data managers	Thematic analysis + design science approach	Prototype contextual framework
Obj.3. Validate framework	DPSP dataset (real cleaning exercise), expert review panels	Usability testing, consensus scoring	Validated contextual framework

3.2 Research Philosophy

A research philosophy provides the foundational worldview or set of beliefs that guide how knowledge is generated and interpreted in a study. It shapes the way a researcher conceptualizes the research problem, selects appropriate tools, and interprets findings. Common philosophical paradigms include positivism, interpretivism, realism, and pragmatism.

In this study, the need to address real-world clinical data challenges using both subjective human insights and objective validation metrics necessitates a philosophy that bridges theory and practice.

3.2.1 Pragmatism as the Adopted Philosophy

Pragmatism was selected as the guiding philosophy for this research due to its problem-solving orientation and emphasis on practical applicability. Unlike positivism, which focuses solely on measurable facts, or interpretivism, which centers on subjective meaning, pragmatism recognizes the value of integrating both qualitative and quantitative approaches to generate actionable knowledge (Morgan, 2014).

Pragmatism allows the researcher to flexibly employ multiple methods and tools that best address the research questions and align with the realities on the ground. This is particularly relevant in clinical research, where technical accuracy must coexist with the experiential knowledge of practitioners. By adopting this philosophy, the study ensures that the resulting data cleaning framework is not only theoretically sound but also implementable in resource-limited health settings such as Masafu Hospital.

3.3 Research Approach

The research approach defines how the researcher moves from the formulation of research problems to data collection, analysis, and interpretation. Approaches can generally be classified as deductive, inductive, or abductive. Given the structured nature of the study, which builds upon existing theories and empirical evidence to design a framework, the deductive approach was most appropriate.

3.3.1 Deductive Reasoning and Its Application

A deductive approach begins with established theories or models and tests their applicability in specific contexts (Bryman, 2016). In this study, the researcher begins with theories and practices related to data cleaning such as rule-based systems, machine learning techniques, and hybrid frameworks as documented in the literature.

These theories are then tested and contextualized to the clinical setting of the DPSP study. Through this approach, assumptions about what constitutes effective data cleaning are assessed in real-world conditions, enabling refinement and adaptation of theoretical principles. Deductive reasoning also supports the validation of the framework, where specific hypotheses such as improved data completeness and consistency can be tested against actual results obtained from the DPSP dataset.

3.3.2 Design Science Paradigm:

Because this study develops and validates an artefact (a contextual data cleaning framework), it followed a Design Science Research (DSR) paradigm (Hevner et al., 2004). DSR emphasizes iterative phases of problem identification, design, demonstration, and evaluation. In this study:

- Problem identification → explored via literature + stakeholder interviews.
- Design & development → using thematic insights + DPSP dataset profiling.
- Demonstration → applying the framework to the DPSP dataset.
- Evaluation → through expert panel review and comparison of pre/post cleaning metrics.

This ensures methodological rigor consistent with artefact-building research traditions.

3.4 Research Strategy

The research strategy operationalizes the overall plan through which data will be collected and analyzed to answer the research questions. Among various strategies such as experiments, surveys, ethnography, and grounded theory, this study adopts a case study strategy.

3.4.1 Case Study Strategy Justification

The case study strategy was selected because it enables an in-depth exploration of a real-life issue within its context, especially when the boundaries between phenomenon and context are not clearly defined (Yin, 2018). Clinical data cleaning in Ugandan settings is influenced by contextual factors such as infrastructure, staff capacity, and organizational workflows. Therefore, a case study approach allows for the comprehensive investigation of these contextual influences on data quality practices.

This strategy also supports triangulation of data sources qualitative interviews, quantitative surveys, and secondary dataset analysis enhancing the reliability and richness of findings.

3.4.2 The DPSP Study at Masafu Hospital as a Case Context

The DPSP (Dihydroartemisinin–Piperaquine and Sulfadoxine–Pyrimethamine) study at Masafu Hospital, located in Busia District, Uganda, serves as the central case for this research. The DPSP study involves longitudinal clinical data related to malaria prevention in pregnancy, encompassing anthropometric, demographic, laboratory, and treatment-related data.

Masafu Hospital, like many rural Ugandan healthcare facilities, faces data management challenges such as missing values, inconsistent records, and underutilized quality control systems. This makes it an ideal context to explore the effectiveness of a contextual data cleaning framework.

Studying the DPSP project provides a focused and practical platform to design, test, and evaluate a tailored data cleaning solution that could be scaled to similar research environments in Low- and Middle-Income Countries (LMICs).

3.5 Research Choices (Methodological Triangulation)

Methodological triangulation refers to the use of more than one method or source of data in a study to enhance the credibility and validity of the results. This study adopts a mixed-method approach, combining both qualitative and quantitative techniques to comprehensively address the research objectives.

The quantitative methods involve measuring data quality improvements (e.g., accuracy, completeness, consistency) using predefined metrics after implementing the proposed framework. The qualitative methods, on the other hand, include interviews and focus group discussions with key stakeholders such as data officers, program managers, and health information technicians to gather insights on contextual needs, existing practices, and usability expectations.

This integration allows the study to capture not only technical performance but also human perspectives, ensuring the framework is both efficient and acceptable to users.

3.5.1 Qualitative Methods: Stakeholder Interviews and Focus Groups

The qualitative component of this study was essential in understanding the real-world complexities of data cleaning in clinical research settings. Stakeholder interviews and focus group discussions (FGDs) were conducted to gather in-depth insights from those directly involved in data management and clinical research operations at Masafu Hospital.

Key participants included data managers, clinical officers, research assistants, and program managers from the DPSP study. These participants were purposefully selected due to their practical experience and domain expertise in handling electronic health records (EHRs), managing patient data, and applying data quality rules in constrained environments.

The interviews were semi-structured, allowing for a balance between consistency across sessions and flexibility to probe emergent themes. Questions focused on:

- Challenges encountered in current data cleaning workflows.
- Strategies used to handle missing data, inconsistencies, and outliers.
- Suggestions for designing a user-friendly, context-aware framework.

The focus groups helped triangulate individual perspectives and validate recurring patterns. Group discussions fostered collaborative reflection on data cleaning experiences and allowed participants to debate context-specific needs, enhancing the framework's relevance and acceptability.

3.5.2 Quantitative Methods: Survey and Data Quality Assessment

Quantitative data were collected to assess measurable improvements in data quality and validate the technical performance of the proposed data cleaning framework. The survey method was employed using structured questionnaires distributed to clinical data personnel involved in the DPSP project.

Survey items focused on:

- Frequency and types of data quality issues encountered.
- Existing manual and digital cleaning procedures.
- Perceptions of framework usability and effectiveness.

In addition, data quality assessment metrics were used to evaluate framework impact. These included:

- Completeness (percentage of missing values handled),
- Accuracy (correctness of data entries after cleaning),
- Consistency (agreement of records across datasets),
- Error reduction rate (number of errors detected and corrected).

Clinical datasets from the DPSP project were used before and after cleaning using the framework to compare these metrics, thus validating the practical benefits of the model.

3.5.3 Mixed-Method Integration: Rationale and Design

A mixed-method design was selected to ensure a comprehensive understanding of both the technical and human-centered dimensions of the research problem. While quantitative methods provided objective evidence of the framework's impact on data quality, qualitative methods captured user perspectives, workflow nuances, and contextual barriers or facilitators to implementation.

The integration was sequential and complementary:

1. Qualitative phase first – to gather in-depth information to inform the framework design.
2. Quantitative phase next – to test and validate the developed framework using real data.
3. Joint analysis phase – where insights from both strands were merged to derive actionable conclusions and refine the model.

This triangulated approach enhanced validity, reliability, and applicability, ensuring the resulting framework was grounded in evidence, contextually relevant, and broadly usable in similar LMIC clinical settings.

3.6 Time Horizon

This study primarily adopted a cross-sectional time horizon, which refers to collecting data at a single point in time rather than over an extended period. Given the time constraints associated with

academic research and the need for timely results to inform ongoing clinical studies, this design was the most practical.

3.6.1 Justification for Cross-Sectional Design

The cross-sectional design enabled the researcher to gather both qualitative and quantitative data in parallel across multiple stakeholders and datasets. It facilitated the simultaneous evaluation of:

- Existing data cleaning practices
- Immediate stakeholder feedback
- Framework performance using a fixed clinical dataset (DPSP)

Moreover, cross-sectional studies are cost-effective and efficient, particularly in resource-limited settings. Although they do not capture longitudinal trends, they are highly useful in exploratory and developmental research such as framework design and initial validation (Saunders et al., 2019).

However, because this study followed a Design Science Research (DSR) paradigm, framework development was not strictly cross-sectional. While interviews and surveys were cross-sectional, the framework design itself was iterative: findings from literature and qualitative analysis informed initial framework design, which was subsequently refined through validation results and expert feedback. This hybrid model therefore combined cross-sectional data collection with iterative artefact development, ensuring both feasibility and methodological rigor.

3.7 Sampling Design and Technique

Sampling refers to the process of selecting a subset of individuals or data points from a larger population for research. A carefully structured sampling strategy was necessary to ensure that the study included respondents and datasets most relevant to the design and validation of a contextual framework for data cleaning, while also balancing resource and time constraints.

3.7.1 Target Population

The study targeted both human participants and secondary datasets:

- Primary stakeholders: Data managers, clinical research officers, and program managers working directly with clinical research data at Masafu Hospital.
- Secondary data: Clinical records from the DPSP study on malaria prevention in pregnancy, covering demographic, laboratory, and follow-up variables.

These were selected because they directly represent the end-users and real-world data contexts in which the proposed framework would be applied.

3.7.2 Sampling Technique

The study adopted a purposive sampling technique, a non-probability method where participants are chosen based on their knowledge, expertise, and relevance to the research objectives. This approach is well-suited for design science and applied health research, where not all members of a population can contribute equally to addressing the research problem.

Specifically, purposive sampling ensured the inclusion of individuals who:

- Have direct experience with clinical data management and quality assurance.
- Understand challenges of data inconsistencies, incompleteness, and manual cleaning in low-resource contexts.
- Can provide informed feedback on the feasibility and usability of a new framework.

For the quantitative component, 80 responses were collected from targeted staff; after data completeness and quality checks, 62 valid responses were retained for analysis. This sample was sufficient for descriptive and inferential analysis while maintaining representativeness of the key stakeholders.

For the qualitative component (interviews and FGDs), participants were selected purposively until thematic saturation was achieved (estimated 12–15 individuals).

3.7.3 Summary of Sampling Design

Category	Sampling Frame	Technique	Final Sample Used	Purpose
Survey participants	Data managers, clinical officers, program staff	Purposive (relevance & experience)	62 valid responses (from 80 collected)	Quantitative assessment of perceptions & practices
Interview/FGD participants	Key informants (MoH, EMR developers, statisticians, data staff)	Purposive & saturation-based	~12–15 individuals	Qualitative insights for design & validation
Secondary data	DPSP clinical dataset (malaria in pregnancy study)	Census (entire dataset used)	Full dataset after cleaning checks	Empirical validation of framework effectiveness

3.7.3 Inclusion and Exclusion Criteria

Inclusion Criteria:

- Participants actively involved in clinical data management or program implementation within the DPSP study.
- Personnel with a minimum of 6 months' experience handling clinical datasets.
- Willingness to participate in interviews, surveys, or data validation sessions.

Exclusion Criteria:

- Staff not directly involved in data collection or management.
- Temporary interns or volunteers with limited engagement in the DPSP study.
- Individuals who declined to provide informed consent.

These criteria ensured the integrity and relevance of the collected data while minimizing bias and irrelevance in the results.

3.8 Methods of Data Collection

This study employed both primary and secondary data collection methods to ensure a comprehensive dataset for developing and validating the contextual framework for data cleaning. The primary data targeted stakeholder experiences, opinions, and challenges, while the secondary data consisted of actual clinical records from the DPSP study used to evaluate the proposed framework.

3.8.1 Primary Data Collection Methods

Stakeholder Interviews

Stakeholder interviews were conducted with carefully selected individuals involved in clinical data management at Masafu Hospital, including:

- Data managers,
- Clinical officers,
- Program coordinators,

These semi-structured interviews enabled the researcher to explore both predefined topics and emerging themes. This format allowed for in-depth exploration of experiences with existing data cleaning techniques, tools used, perceived limitations, and aspirations for a more suitable framework. Interviews were scheduled based on participant availability and were conducted in quiet, confidential settings to foster open communication. Each session lasted approximately 30–45 minutes and was audio-recorded with consent.

Focus Group Discussions (FGDs)

FGDs complemented interviews by facilitating group interaction, which enriched the data through collective reflection. Two FGDs were conducted:

One with data officers

- Another with senior staff including program managers and IT support

Discussions revolved around current data cleaning workflows, data accuracy practices, bottlenecks in error resolution, and reactions to prototype designs of the new framework. The group format allowed researchers to gather diverse perspectives, uncover hidden challenges, and validate the findings across different levels of data responsibility.

Questionnaire Administration

Structured questionnaires were designed and administered to a wider group of stakeholders, including those who could not participate in interviews. The questionnaires were:

- Self-administered, with clear instructions,
- Divided into sections corresponding to objectives (e.g., current practices, challenges, expectations for a framework),
- Administered both physically (print copies) and digitally (Google Forms) to increase reach and convenience.

A total of 80 responses were collected, representing a broad stakeholder base. The responses were coded and analyzed quantitatively to assess trends, frequencies, and stakeholder consensus on key issues.

3.8.2 Secondary Data Collection Methods

DPSP Clinical Research Dataset

The secondary dataset used for this study was extracted from the ongoing DPSP study (Dihydroartemisinin-Piperaquine and Sulfadoxine-Pyrimethamine for Malaria Prevention in Pregnancy), conducted at Masafu Hospital, Busia District. This dataset includes:

- Demographic details (age, gravida, parity),
- Clinical indicators (lab results, test outcomes, treatment responses),
- Follow-up visit logs,
- Routine observational data.

The dataset was accessed under strict ethical compliance and confidentiality agreements, and it was anonymized prior to analysis.

This clinical dataset served a dual purpose:

1. To validate the proposed framework by applying it to real-world data and measuring improvements in data quality.
2. To identify common data challenges (missing fields, duplicate entries, outliers), which informed the customization of data cleaning rules.

3.9 Data Collection Instruments and Tools

To ensure validity, reliability, and ease of administration, several tools were developed and piloted before formal data collection commenced.

3.9.1 Questionnaires: Structure, Administration, and Rationale

The questionnaire used in this study was structured into five core sections:

1. Demographics of Respondents – including role, experience level, and familiarity with data cleaning.
2. Current Practices – capturing tools used, types of errors faced, and cleaning workflows.

3. Challenges and Limitations – exploring bottlenecks and constraints in manual and semi-automated approaches.
4. Framework Expectations – assessing user priorities (accuracy, usability, speed, transparency).
5. Validation Metrics – self-reported evaluation of data quality before and after using different tools.

Each section used Likert-scale questions, multiple-choice, and a few open-ended items. The rationale for using questionnaires was:

- Cost-effectiveness: enabling broad coverage with minimal resources.
- Standardization: facilitating comparison and analysis.
- Anonymity: encouraging honest feedback.

Prior to roll-out, the tool was pre-tested with 5 participants to improve clarity and ensure reliability (Cronbach's alpha = 0.82).

3.9.2 Interview Guides and Recording Tools

The interview guides were carefully designed to follow the research objectives and align with themes from the literature. The guide comprised:

- Opening questions (e.g., role, experience with data),
- Core questions (e.g., describe your data cleaning process, biggest challenges, what a good framework should address),
- Probing follow-up prompts (e.g., examples, frequency of certain issues).

To capture the richness of verbal data:

- Audio recorders (smartphone apps and portable digital recorders) were used during interviews and FGDs.
- Notes were also taken during sessions as a backup and to highlight emotional cues or non-verbal communication.

- Recordings were later transcribed verbatim for thematic coding using tools like NVivo and MS Excel.

These tools provided high-fidelity data for analysis and ensured that critical feedback from stakeholders was not lost during the documentation process.

3.9.3 SQL Tools for Data Profiling and Cleaning

Structured Query Language (SQL) tools were employed to automate and support foundational data cleaning operations. SQL provided a reliable environment for querying the DPSP dataset, identifying anomalies, and implementing rule-based cleaning routines. Specifically, SQL was used to:

- Profile the dataset by checking for null values, duplicates, out-of-range values, and inconsistencies in categorical variables (e.g., gender, treatment codes).
- Implement constraints (e.g., CHECK, UNIQUE) to enforce data validation rules.
- Standardize entries using UPDATE queries (e.g., harmonizing date formats, correcting spelling errors).
- Generate reports to visualize error patterns for decision-making.

These tasks were executed in MySQL and PostgreSQL environments, which are both robust, open-source tools commonly used in clinical data analysis. SQL scripts also supported semi-automated workflows in later phases of framework implementation, enabling efficient reproducibility and scalability of cleaning tasks.

3.10 Data Sources and Selection Criteria

This study drew upon both primary and secondary data, with a strong emphasis on real-world clinical datasets to ensure contextual relevance and empirical validation of the framework.

3.10.1 Description of the DPSP Dataset

The DPSP dataset (Dihydroartemisinin-Piperaquine and Sulfadoxine-Pyrimethamine study) was the main secondary data source. The dataset was extracted from a longitudinal clinical trial

conducted at Masafu Hospital in Busia District, Eastern Uganda. Its focus is on evaluating the efficacy of two drug regimens in malaria prevention during pregnancy.

Key features of the dataset include:

- Participant demographics (age, gravida, parity, education, etc.)
- Clinical measurements (malaria test results, hemoglobin levels, fetal outcomes)
- Follow-up data (scheduled visits, adherence, adverse events)
- Lab reports and EMRs recorded over the study timeline.

3.10.2 Ethical Access and Use of Clinical Data

Ethical use of the DPSP dataset followed strict data governance protocols. Prior to access:

- The research team submitted a formal data access request to the DPSP principal investigators and the Institutional Review Board (IRB).
- A Data Use Agreement (DUA) was signed to enforce data confidentiality, stipulating that all personal identifiers be removed prior to data transfer.
- Data was anonymized and encrypted before analysis. All cleaning and profiling were done on a local secure machine, and no patient-identifiable information was stored.

The research strictly adhered to Uganda National Council for Science and Technology (UNCST) guidelines and the principles of the Declaration of Helsinki (2013 revision).

3.11 Implementation and Validation of the Framework

The designed contextual data cleaning framework was implemented in two phases and validated through both qualitative and quantitative methods. This ensured the framework was not only functional but also user-friendly and context-appropriate.

3.11.1 Implementation Phases of the Cleaning Framework

Phase 1: Application on Basic Cleaning Tasks

This initial phase involved applying the designed framework to standard cleaning tasks, including:

- Null value detection and treatment (e.g., NULL to 'Not Recorded' or imputation),
- Duplicate entry identification using SQL joins and hashing techniques,
- Format standardization (e.g., consistent date and numeric formats),
- Basic error flagging based on predefined logic rules.

This phase tested the internal consistency of the framework and fine-tuned its modules.

Phase 2: Integration with DPSP Dataset

In this advanced phase, the framework was fully deployed on the DPSP dataset. Custom integrity rules were applied to:

- Validate clinical logic (e.g., pregnant males flagged as erroneous),
- Perform context-aware outlier detection using reference ranges (e.g., hemoglobin levels below 5g/dL),
- Generate audit trails for reviewed and corrected entries.

Comparison was made between manually cleaned and framework-cleaned datasets to evaluate performance improvements.

3.11.2 Validation Procedures

Usability Assessment via Stakeholder Feedback

After implementation, the system's usability was assessed through:

- Structured feedback sessions,
- Usability questionnaires based on the System Usability Scale (SUS),
- Stakeholder reflection meetings involving data officers, program managers, and clinical investigators.

Feedback focused on ease of use, integration potential, and suggestions for improvement, helping refine the user interface and recommendation modules.

Accuracy, Completeness, and Error Reduction Tests

To objectively validate the framework:

- The original dataset was compared to the cleaned version using metrics such as:
    - Completeness (%): Increase in non-null records,
    - Consistency (%): Resolution of conflicting entries,
    - Error Rate Reduction (%): Drop in syntax and logic-based errors.

These metrics were benchmarked against manually cleaned versions and prior NADEEF-based rules, highlighting the framework's added value in real clinical research settings.

3.12 Ethical Considerations

All stages of this study were conducted in accordance with established ethical standards and institutional guidelines.

Informed Consent and Voluntary Participation

For all primary data collection:

- Participants (data managers, officers, etc.) were given clear information sheets describing the study's objectives, benefits, and rights.
- Written informed consent was obtained prior to interviews and FGDs.
- Participation was entirely voluntary, with the option to withdraw at any time without repercussions.

Data Privacy and Anonymization Protocols

To uphold confidentiality:

- All recorded interviews and transcripts were anonymized using unique participant codes.
- Personal identifiers were stripped from the dataset before cleaning.
- Devices used for data storage were encrypted and password protected, with access limited to the core research team.

Research Clearance and IRB Approval

- Ethical approval was secured from the K Research Ethics Committee (REC) and acknowledged by the UNCST.
- Reference numbers and approval letters were kept on file for auditing.
- A compliance report was submitted to Masafu Hospital's administration and the DPSP study lead to ensure institutional transparency.

3.13 Data Analysis Techniques

The analysis of collected data employed both quantitative and qualitative methods, consistent with the study's mixed-methods approach. This integration enabled triangulation of findings, ensuring a deeper and more credible understanding of the effectiveness and usability of the proposed contextual framework for data cleaning.

3.13.1 Quantitative Analysis

Quantitative data from structured questionnaires and the DPSP dataset were analyzed using statistical software (SPSS, Excel, Python libraries: Pandas, NumPy, Matplotlib).

1. Descriptive Statistics summarized dataset characteristics:
    - Frequencies and percentages for categorical variables (e.g., gender, clinical status).
    - Means, medians, and standard deviations for continuous variables (e.g., hemoglobin levels, age).
2. Data Quality Metrics assessed framework impact:
    - Completeness rate (% of missing values handled).
    - Consistency score (resolution of conflicting records).
    - Error rate (identified vs. corrected errors).
3. Inferential Statistics tested improvements between manual vs. framework-cleaned datasets:
    - T-tests (paired/unpaired) for mean differences.
    - Chi-square tests for categorical comparisons.
    - Correlation analysis to explore relationships (e.g., error rate vs. number of visits).

This established whether the framework statistically improved core data quality attributes.

3.13.2 Qualitative Analysis

Qualitative data from stakeholder interviews and FGDs underwent thematic analysis:

1. Transcription → verbatim conversion of audio recordings.
2. Familiarization → repeated reading to identify emerging ideas.
3. Coding → NVivo + manual open coding of key phrases.
4. Theme Generation → grouped into themes such as:
    - Challenges in current data cleaning practices.

- Expectations of automation.
- Barriers to adopting new tools.
- Contextual differences in data entry and structure.

5. Interpretation → Themes were mapped to study objectives to refine framework usability and contextual fit.

This qualitative layer enriched interpretation of quantitative results and grounded the framework in user realities.

3.13.3 Summary of Data Analysis Procedures

Data Source	Analysis Techniques	Outputs / Purpose
Survey Data (Quantitative)	Descriptive stats (freq., %, mean, SD); reliability (Cronbach's $\alpha \geq 0.7$); inferential tests (t-tests, chi-square, correlations)	Summarize perceptions; test improvements and relationships
DPSP Dataset (Clinical Records)	Data profiling; pre/post cleaning comparison of completeness, accuracy, error rates; inferential testing	Demonstrate effectiveness of framework on real clinical data
Qualitative Data (Interviews & FGDs)	NVivo coding; thematic analysis; clustering into themes	Identify user needs, challenges, contextual factors
Expert Panel Validation	Consensus scoring; structured feedback	Assess feasibility, usability, scalability
Triangulation (Mixed Integration)	Cross-verification of quant. & qual. results	Ensure robust findings and refine framework

3.14 Reliability and Validity Considerations

Ensuring reliability and validity was fundamental to confirm that the study's findings are credible, consistent, and applicable to clinical research contexts.

3.14.1 Reliability of Instruments and Processes

Reliability was addressed through:

- Pilot Testing → questionnaire pre-tested with five data managers not in main sample.
- Internal Consistency → Cronbach's Alpha ≥ 0.7 ensured acceptable reliability.
- Standardized Interview Guides → uniform data collection across respondents.
- Audit Trails → version-controlled SQL scripts and cleaning protocols ensured reproducibility.
- Training → research assistants trained in data entry and transcription to minimize error.

3.14.2 Validity of Framework and Research Findings

Validity was ensured through:

- Content Validity → tools reviewed by supervisors and domain experts.
- Construct Validity → framework aligned to constructs (accuracy, completeness, consistency) and tested on real datasets.
- Face Validity → stakeholder confirmation of relevance and usability.
- Triangulation → integration of quantitative + qualitative findings.
- Validation Testing → framework outputs measured objectively (metrics) and subjectively (stakeholder usability).

Together, these measures ensured methodological rigor and enhanced confidence in recommending the framework for adoption in clinical research under resource-limited settings.

CHAPTER FOUR:

PRESENTATION OF FINDINGS

4.1 Introduction

This chapter presents the findings of the study based on the objectives outlined in Chapter One. It provides a detailed synthesis of both qualitative and quantitative data collected through user engagement, interviews, observational sessions, and structured workshops conducted among stakeholders involved in clinical research and data management. The central aim of the study was to design and validate a contextual framework for data cleaning in clinical research settings, specifically at Masafu Hospital under the DPSP study.

A user-intervention approach was employed, deemed suitable due to the complexity and sensitivity of clinical research data. This method enabled direct involvement of end-users such as data officers, managers, and clinical teams, ensuring the proposed model was both contextually appropriate and practically implementable.

The findings in this chapter are structured according to the specific objectives of the study. Each section highlights the experiences of the users, the tools they utilize, the challenges they face, and how these insights informed the development of the proposed framework.

4.2 Demographic and Background Information

This section presents an overview of the respondents who participated in the research. Understanding their background, particularly their years of experience and roles in the clinical research data workflow, is crucial to interpreting the findings objectively and aligning the framework to user capabilities and organizational dynamics.

4.2.1 Respondents' Experience in Clinical Data Management

The level of experience of respondents in clinical data management is critical in understanding their exposure to various tools and their capacity to evaluate or apply data cleaning strategies

effectively. As illustrated in Figure 2: User Experience**,** the participants were categorized into three key groups:

- **0–5 years (29%)**: This group mainly comprises entry-level professionals, likely in the early stages of learning how to manage and clean clinical datasets. Their limited exposure to research protocols and tools places them in need of targeted training and supervision. While they may not contribute significantly to framework design, their experiences are valuable in understanding usability and identifying training gaps.
- **5–10 years (25%)**: Representing mid-career professionals, this group likely has some familiarity with data systems and research ethics. However, they still require guided implementation and support to align their practices with Good Clinical Practice (GCP) and regulatory expectations. Their contributions were essential in pointing out gaps in existing data cleaning methodologies.
- **10–15+ years (26%)**: This was the most dominant category among respondents. Their deep experience provided critical input into the current limitations of widely used tools and practices. Their involvement ensured that the framework addressed real-world complexities and was designed with appropriate simplicity, guided by clearly documented Standard Operating Procedures (SOPs).

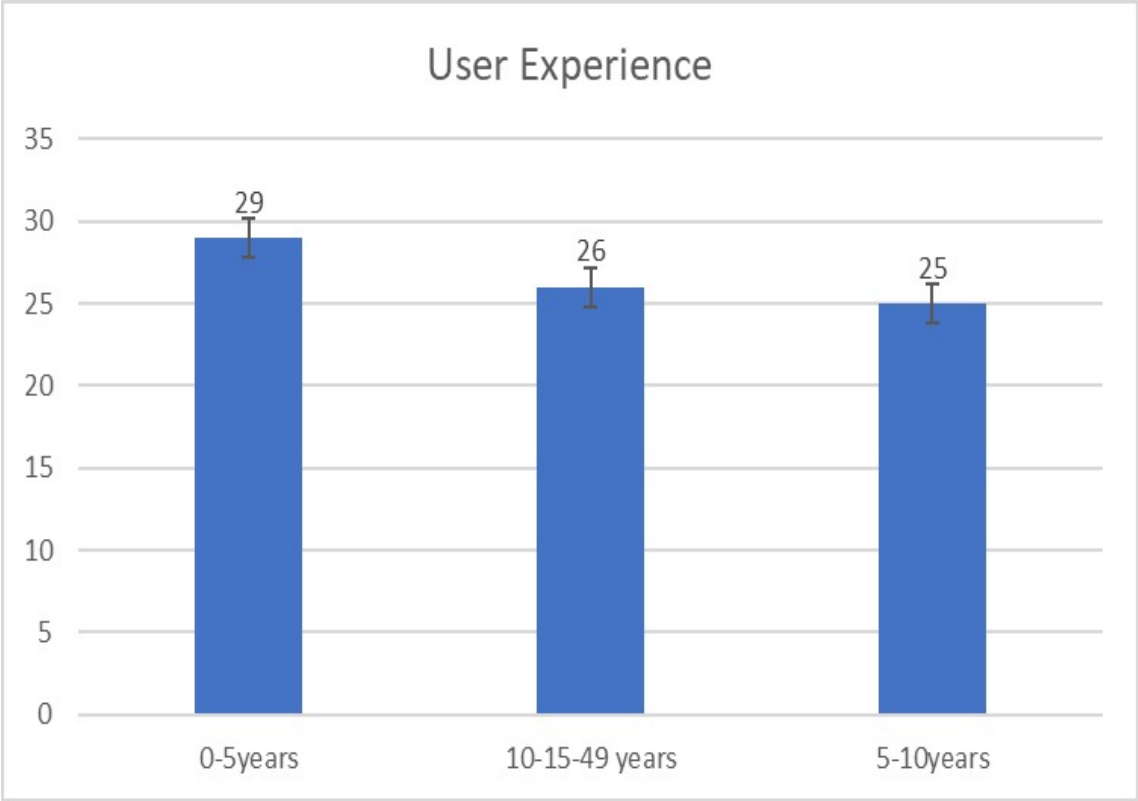


Fig 2: user experience of respondents

This distribution reveals a healthy mix of expertise, ensuring that the framework developed caters to users with varying skill levels, thus maximizing adoption and relevance.

4.2.2 Roles and Responsibilities in Data Cleaning Processes

Participants in the study held various positions along the data cleaning value chain, each contributing uniquely to the clinical research process. Their roles determined the nature of feedback provided and the emphasis placed on certain functionalities in the proposed framework. The core roles identified include:

- **Program Managers**: Responsible for overseeing the overall data workflow, aligning processes with project goals, and ensuring regulatory compliance.
- **Data Managers**: Tasked with creating SOPs and managing cleaning protocols. They bridge the technical and operational gaps between data sources and research analysis.

- **Data Officers**: These are the primary implementers of data cleaning activities. They interact with tools daily, handle inconsistencies, and ensure that corrected data aligns with SOPs.
- **Clinical Teams**: Offering contextual domain knowledge, clinical staff ensure that cleaned data is accurate and relevant to study outcomes, especially in areas like patient history, drug administration, or laboratory results.

These insights were instrumental in designing a framework that is both modular and user-specific. Tasks such as identifying outliers, flagging inconsistencies, or performing validation checks were linked directly to these defined roles to minimize overlap and maximize efficiency.

4.3 Findings by Research Objectives

4.3.1 Objective 1: To synthesize and critique existing data cleaning frameworks and methodologies, highlighting their strengths and limitations in low-resource clinical research environments

The first objective of this study was to investigate the current data cleaning methodologies employed in clinical research and to understand the key challenges associated with them. The aim was to identify gaps that would inform the design of a more effective, contextual framework suitable for low-resource environments like Masafu Hospital.

4.3.1.1 Overview of Current Methods in Clinical Research

Data cleaning methods in clinical research typically involve a combination of statistical software tools and manual verification techniques. The most commonly used tools identified in this study were STATA, R Programming, Python, and Microsoft Excel. Figure 3 illustrates the frequency and combinations in which these tools were applied.

In the survey administered to 80 respondents, it was evident that STATA and R were the most frequently used tools for statistical cleaning and data correction, often paired with Python for scripting. Excel, though familiar and readily accessible, was seldom used in isolation due to its limited automation and error detection capabilities. Most users reported employing Excel alongside other tools often for preliminary data review or file handling.

Despite the technical capabilities of advanced tools such as R and Python, their adoption was not universal. Some users indicated that the complex syntax, steep learning curves, and lack of training hindered effective use. As a result, manual processes remained prevalent, especially in settings where technical capacity was limited. These manual approaches involved routine checking of data entries, visual inspection for anomalies, and manual editing a process that, while intuitive, is slow, prone to error, and difficult to scale.

The figure below (Figure 3: Current Data Cleaning Methods by Frequency) reveals this distribution and highlights a growing reliance on hybrid cleaning strategies, combining scripting, statistical tools, and spreadsheets.

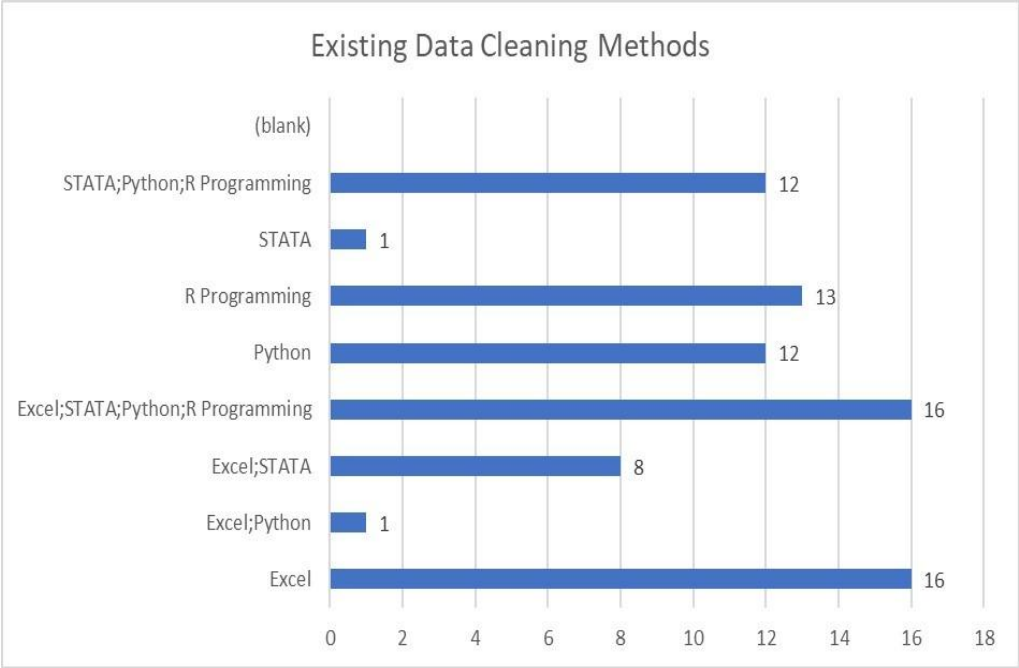


Figure 3: Current Data Cleaning Methods by Frequency

4.3.1.2 Challenges Associated with Existing Methods

One of the key findings from interviews and survey data was the persistent challenge of handling missing and inconsistent data. Participants noted that deciding whether to drop, impute, or flag missing values often relied on subjective judgment, due to the absence of standardized, context-sensitive guidelines. In clinical trials, where patient data may be incomplete due to missed visits, dropout, or technical glitches, this lack of standardization often led to biased or inconsistent datasets.

Another pressing challenge was the limitation of existing tools in meeting the complex needs of clinical datasets. While advanced tools like R and Python offer automation, their utility is hindered by the high entry barrier for non-technical users. Tools such as Excel, on the other hand, though

widely available, were described as inefficient for enforcing complex rules or detecting cross-variable inconsistencies.

Respondents consistently cited human error and syntax complexity as root causes of flawed datasets. Manual data entry, incorrect formula application in spreadsheets, and poorly written scripts led to either overlooked errors or the creation of new ones. Moreover, maintaining consistency across multi-site studies became an uphill task when data validation rules were implemented differently due to lack of a unified cleaning protocol.

These insights are well visualized in Figure 1: Data Cleaning Challenges, which categorizes the major technical and operational difficulties encountered in the cleaning process. Figure 4: Existing Data Cleaning Issues complements this by providing a comparative view of common issues such as repetitive tasks, error-prone workflows, and complex command structures in clinical research contexts.

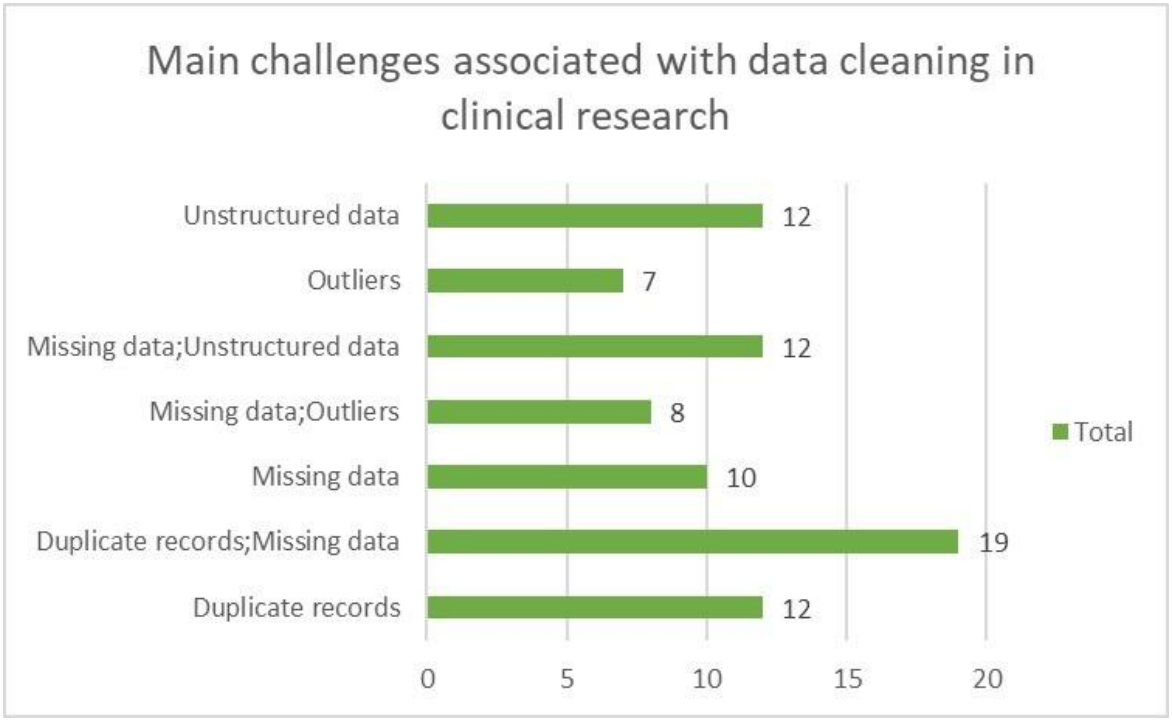


Figure 1: Data Cleaning Challenges

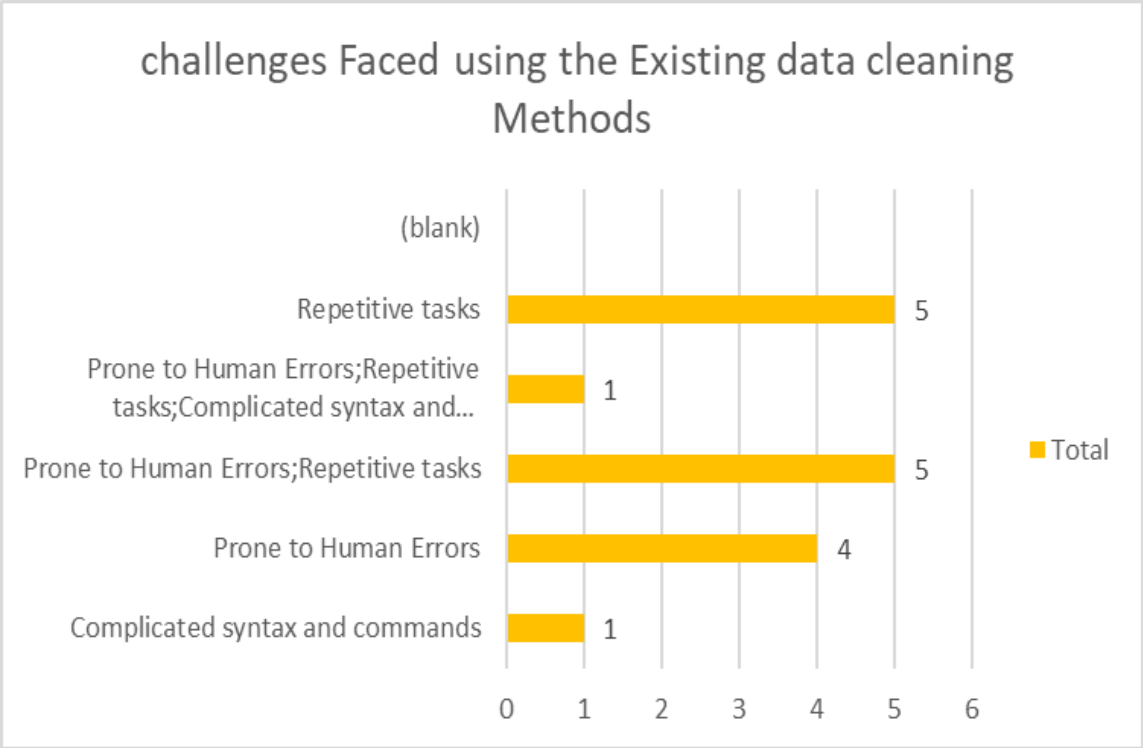


Figure 4: Existing Data Cleaning Issues

4.3.1.3 Thematic Analysis of Interview and Focus Group Data

A thematic analysis of stakeholder interviews and focus group discussions yielded several recurring themes that further highlighted the operational challenges in clinical data cleaning.

Repetitive Tasks emerged as a key concern. Many respondents described having to repeat similar cleaning tasks such as outlier detection, missing value imputation, and data revalidation across different datasets or versions. These tasks, often executed manually or through custom scripts, consumed a considerable amount of time and introduced a high risk of oversight or fatigue-related error.

Another recurring theme was limited automation. Despite some respondents having access to automated tools, these were either poorly integrated into their workflows or not user-friendly. For instance, Python and R scripts often had to be written from scratch for each dataset, rather than using reusable templates or rule-based engines. Participants expressed the need for a system that

could learn from user interactions and evolve cleaning rules, rather than depending entirely on rigid scripts.

Subjective decision-making was also noted as a limitation. In the absence of contextual rules or standard protocols, decisions about how to handle anomalies were often based on individual judgment. This resulted in inconsistencies across users or teams, especially in multi-center studies.

Lastly, regulatory pressures in low-resource settings compounded the difficulties. Respondents from Masafu Hospital noted that strict data integrity standards expected by ethical review boards and funding agencies were hard to meet with current manual approaches. Given the shortage of trained personnel and limited access to robust tools, maintaining audit trails, documentation of cleaning steps, and version control was challenging.

These themes reveal a clear gap in the existing landscape the absence of an integrated, context-sensitive, user-friendly cleaning framework that balances automation with human oversight. The findings from this objective laid the groundwork for the design specifications of the proposed framework discussed in subsequent sections.

Implication for Framework Design

The identification of repetitive tasks and subjective decision-making informed the need for automation modules and embedded SOP-based rules in the framework. Limited automation highlighted the importance of user-friendly SQL templates for non-technical staff.

4.3.2 Objective 2: To design a contextual data cleaning framework informed by data quality theory, information systems success models, and socio-technical perspectives.

The second objective of this study focused on designing a contextual and modular framework for cleaning clinical research data, with specific attention to low-resource settings like Masafu Hospital. The aim was to translate real-world challenges into a structured and user-centered system that optimizes data quality, usability, and compliance in clinical trials and related research. A user-intervention methodology was applied, which emphasized participatory design and stakeholder input at every stage**.**

4.3.2.1 User Requirements and Stakeholder Involvement

To ground the framework in practice, **a** series of interviews and focus group discussions were conducted with key stakeholders, including data managers, clinical officers, program coordinators, and research assistants. These engagements revealed crucial insights into the expectations, frustrations, and operational needs of users tasked with data cleaning.

A thematic analysis of interview transcripts highlighted three core user goals: (1) minimizing repetitive and error-prone manual work, (2) enabling non-technical staff to carry out advanced data validation with ease, and (3) integrating domain knowledge and decision-making logic into the cleaning system. Respondents also emphasized the need for traceability, transparency, and reproducibility especially where data audits or publication standards are concerned.

To capture diverse workflows, scenario-based planning was used. For example, in a workshop held with Masafu Hospital research teams, mock datasets were evaluated under various real-world constraints such as incomplete patient records, lab result delays, and duplicate IDs. Each scenario helped define critical requirements like automated flagging rules**,** simplified user interfaces, and SQL-query-based validation tools.

The outcome of this stage was a comprehensive list of functional needs, ranging from real-time error notifications and built-in SOPs to integrated logs that record cleaning steps. These informed the design of the contextual framework described below.

4.3.2.2 Current Clinical Data Cleaning Workflow

Understanding the current workflow was critical to ensuring the proposed framework builds on what works and improves what doesn't. The existing data cleaning process, as reported by stakeholders, involves multiple actors with distinct roles:

- Program Managers; oversee the broader project and ensure alignment with objectives and protocols.
- Data Managers; are responsible for designing SOPs and ensuring that data cleaning practices comply with those procedures.

- Data Officers; execute most of the hands-on cleaning activities, such as checking for duplicates, running scripts, and maintaining logs.
- Clinical Teams; contribute by validating clinical relevance flagging biologically implausible values or cross-referencing data with source documents.

The typical process begins with data collection and download, followed by rule-based validation using pre-written scripts or manual review. Cleaned data is then re-uploaded to a central repository**,** often shared via platforms such as Dropbox or institutionally hosted servers.

A key part of this process includes SOP-driven steps for outlier detection, missing value treatment, and transformation rules. However, these processes were often fragmented, with limited integration between actors and stages, leading to inefficiencies and duplication of effort.

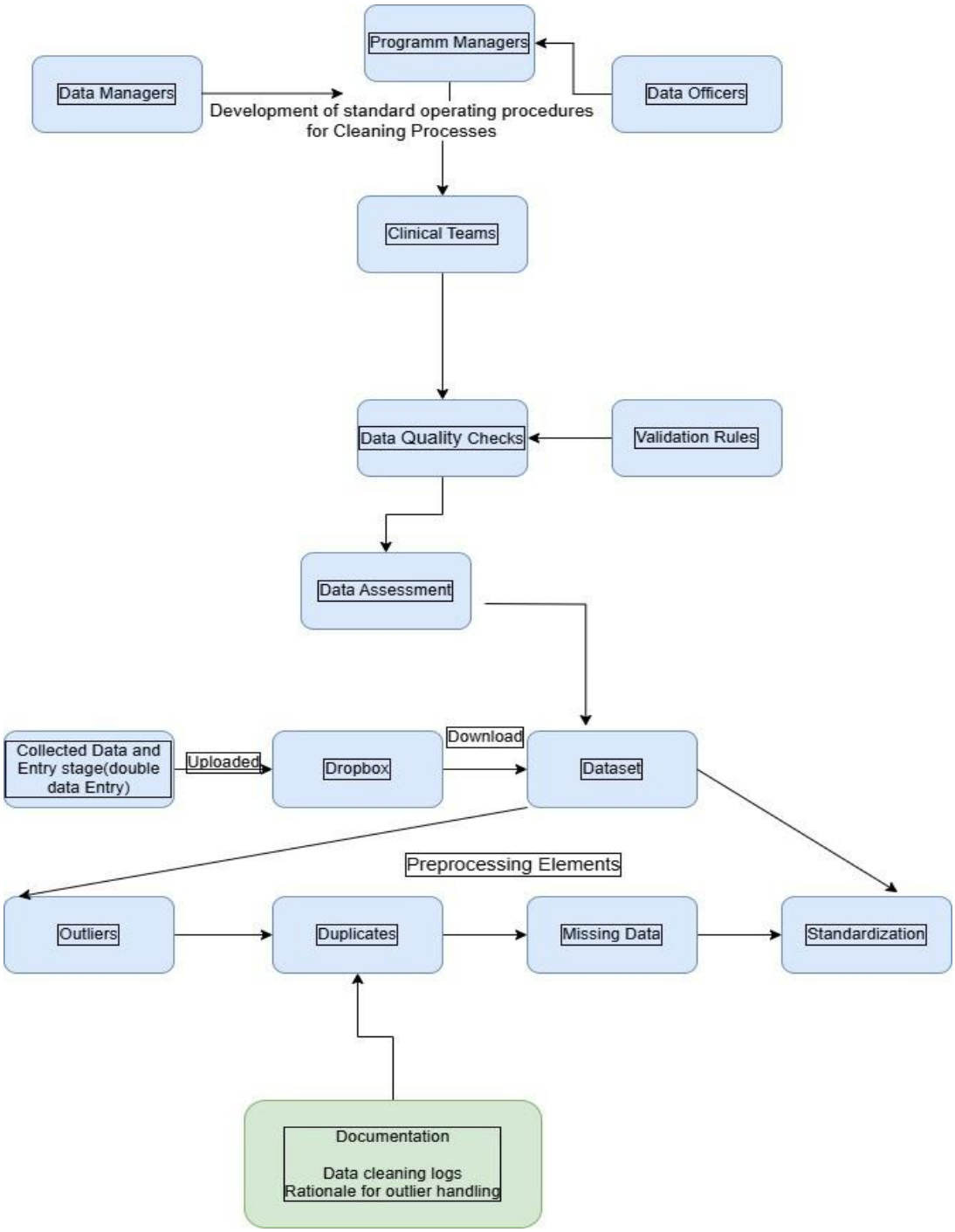


Figure 5: Current Data cleaning Framework

Figure 5: Current Data Cleaning Framework visually illustrates this workflow, highlighting the disjointed flow of data, reliance on manual checks, and the limited feedback loops for error detection and correction.

4.3.2.3 Designed Contextual Framework Overview

Building on stakeholder insights and gaps identified in current workflows, a modular, contextual data cleaning framework was developed. This framework adopts a six-stage approach, allowing for flexibility, automation, and human oversight throughout the data lifecycle.

Stage 1: Data Collection

This is the foundational stage of the framework, where raw data is captured from clinical activities such as patient enrollment, drug administration, and lab tests. Emphasis is placed on ensuring standardized forms, proper digitization at source, and adherence to clinical protocols, as inconsistencies here affect all downstream processes. For example, in the DPSP trial, centralized data capture systems were employed to limit variability and improve completeness.

Stage 2: Data Profiling

In this stage, the structure and health of the dataset are assessed. Profiling includes summarizing distributions, identifying outliers, scanning for missing fields, and checking variable types. Stakeholders reported that this step helped them prioritize cleaning efforts. For instance, records with implausible hemoglobin levels were immediately flagged for review.

Stage 3: Data Preprocessing

Preprocessing prepares data for cleaning by splitting compound variables, converting data types, and identifying formatting errors. In the DPSP study, this involved parsing date-time fields, ensuring numeric variables were not stored as strings, and merging multiple records for the same patient. This stage ensures consistency and compatibility across tools and stages.

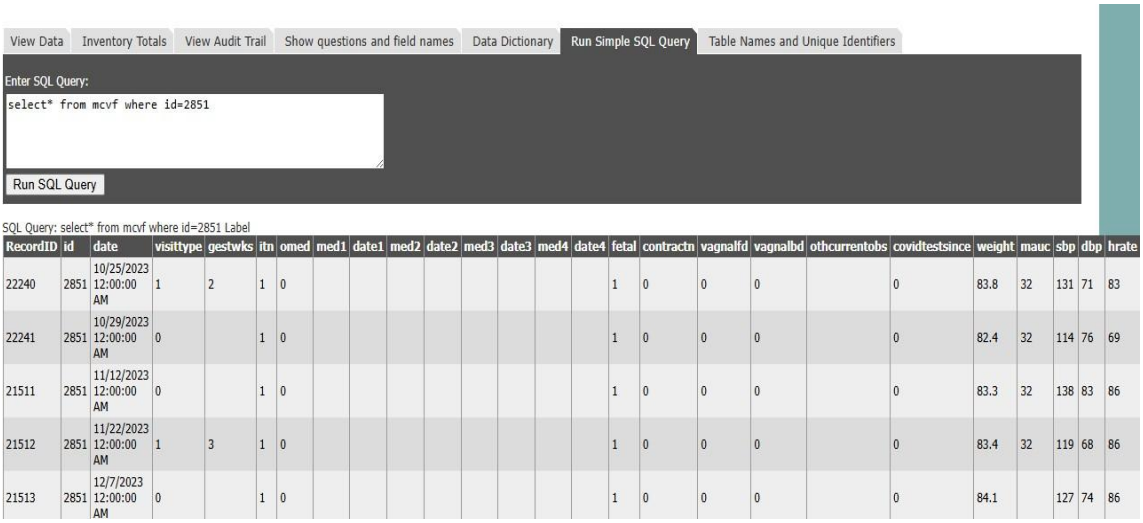
Stage 4: Data Cleaning

This is the core stage, where actual cleaning actions are performed. These include missing value treatment (via deletion, imputation, or annotation), duplicate record resolution, and anomaly correction. One participant from Masafu Hospital cited an instance where two records had the same patient ID but different visit dates manual cross-referencing with source documents helped

resolve such conflicts. This stage also emphasized maintaining audit trails and logs to support transparency.

Stage 5: Data Enhancement

Once cleaned, the dataset is enhanced to improve its analytical utility. This includes deriving new variables, merging datasets, and standardizing field names and units. For example, SQL queries were used to calculate treatment adherence rates and group visit intervals. Enhancement also supported the generation of analysis-ready subsets and summaries.



RecordID	id	date	visittype	gestwks	itn	omed	med1	date1	med2	date2	med3	date3	med4	date4	fetal	contractn	vagnalfd	vagnalbd	othcurrentobs	covidtestsince	weight	mauc	sbp	dbp	hrate
22240	2851	10/25/2023 12:00:00 AM	1	2	1	0									1	0	0	0		0	83.8	32	131	71	83
22241	2851	10/29/2023 12:00:00 AM	0		1	0									1	0	0	0		0	82.4	32	114	76	69
21511	2851	11/12/2023 12:00:00 AM	0		1	0									1	0	0	0		0	83.3	32	138	83	86
21512	2851	11/22/2023 12:00:00 AM	1	3	1	0									1	0	0	0		0	83.4	32	119	68	86
21513	2851	12/7/2023 12:00:00 AM	0		1	0									1	0	0	0		0	84.1		127	74	86

Figure 7: Simple data cleaning task using Sql

Figure 7: SQL Query Example – Simple Data Cleaning Task ; demonstrates a query used to extract unique clinic visit records and calculate drug compliance over time, showcasing the practical utility of enhancement through SQL.

Stage 6: Quality Scoring

Finally, the data is assessed for quality based on three criteria accuracy, completeness, and consistency. These metrics are derived from both raw and cleaned datasets. Quality scoring not only evaluates readiness for analysis and reporting but also guides iterative cleaning, by revealing problem areas that may require re-cleaning or additional validation.

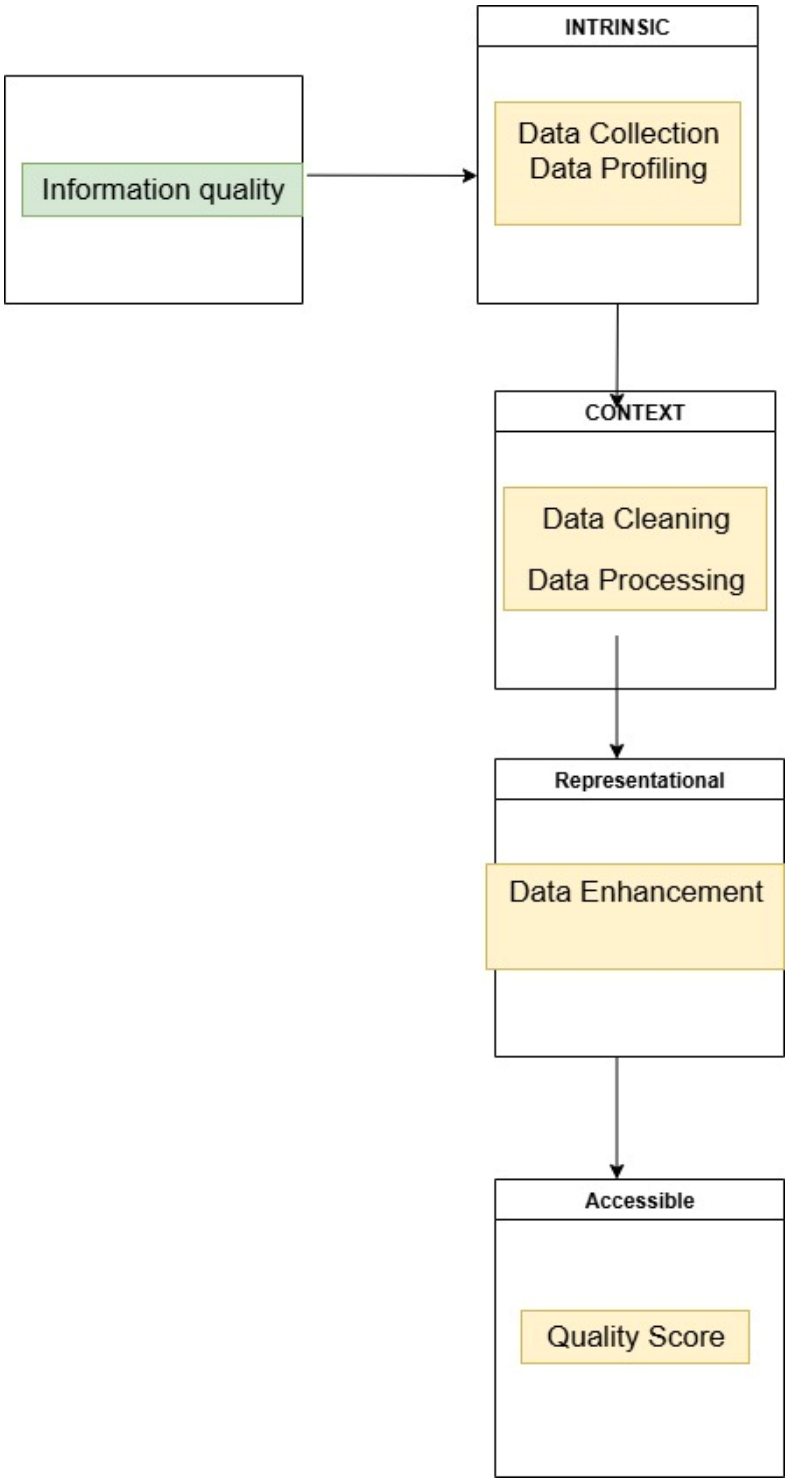


Figure 6: Designed Framework

Figure 6: Designed Contextual Framework; presents the six-stage model in a visually accessible form, emphasizing inter-stage feedback, user roles, and decision points.

Conclusion of Objective 2

The contextual framework designed through this objective addresses the unique demands of clinical research in low-resource settings. By incorporating user input, automating repetitive tasks, and simplifying technical barriers, the framework provides a scalable, sustainable approach to clinical data cleaning.

The modular design ensures adaptability, allowing it to be tailored to specific study protocols or institutional policies. Each stage of the framework is grounded in real-life use cases and supported by appropriate tools, from SQL queries to visual inspection aids. Ultimately, this framework not only enhances data quality and reliability but also reduces operational burden, thus contributing to more ethical, efficient, and effective clinical research practices.

Table 4.1: Translation of Findings into Framework Design Features

Challenge Identified (Findings)	**Framework Response (Module/Feature)**	**Benefit to Users**
Repetitive manual cleaning tasks	Modular pipeline with reusable SQL templates	Saves time, reduces human error
Lack of automation in low-resource settings	Automated profiling & rule-based validation	Increases efficiency, reduces technical barrier
Subjective, inconsistent decisions	SOP-linked decision rules + quality scoring module	Ensures standardization and transparency
Limited technical capacity (non-programmers)	User-friendly SQL snippets & guided workflows	Empowers broader staff participation
Difficulty maintaining audit trails	Built-in logging & version control	Improves traceability and compliance

4.3.3 Objective 3: To Validate the Usability and Effectiveness of the Proposed Framework

4.3.3.1 Implementation on DPSP Dataset

To validate the practical applicability of the proposed data cleaning framework, the study employed a two-phased implementation using the DPSP clinical dataset. The first phase focused on testing simple data cleaning tasks to assess the framework's basic functionalities. These included identifying and handling missing values, detecting and removing duplicate records, standardizing data types, and correcting format inconsistencies. These tasks were carried out using a structured SQL-based approach embedded in the framework. The implementation demonstrated the effectiveness of modular automation, with clearly defined rules and systematic logging to reduce errors and enhance transparency.

In the second phase, the framework was integrated into the full DPSP clinical dataset for more advanced data management and validation tasks. These involved joining multiple data tables, such as patient demographics, clinical visits, and laboratory results, to identify discrepancies across datasets. In this context, the framework's enhancement and profiling stages allowed for intelligent querying, derivation of new attributes (e.g., treatment compliance rates), and cross-sectional data validations. The integration demonstrated not only the scalability of the framework but also its suitability in addressing complex data cleaning challenges within real-world clinical research workflows.

4.3.3.2 Usability Assessment

Usability of the framework was assessed through stakeholder engagement, including follow-up interviews and feedback sessions with clinical teams, data managers, and analysts involved in the DPSP study. Key feedback themes included the perceived ease of use, adaptability of modular steps, and the framework's ability to reduce cognitive load and manual errors.

Stakeholders reported that the predefined stages of the framework such as profiling, preprocessing, and enhancement greatly simplified routine tasks. Users noted that they were able to complete previously tedious operations more efficiently due to guided steps and embedded validation rules. Moreover, participants with limited programming expertise appreciated the user-friendly structure,

especially the ability to perform advanced operations using basic SQL templates without needing complex Python or R scripting.

However, a few challenges were reported concerning the transition to the new tool, particularly around initial training and system familiarity. Some participants expressed concerns about adapting from familiar environments like Excel to a rule-based system. Training sessions were necessary to acquaint users with the modular logic and help them understand each framework stage. Despite these transitional challenges, overall stakeholder perception reflected improved workflow control, better traceability, and increased confidence in data accuracy.

4.3.3.3 Effectiveness Evaluation Using Data Quality Metrics

The effectiveness of the framework was quantitatively measured using three primary data quality metrics: accuracy, completeness, and consistency. These were selected based on systematic literature review and relevance to regulatory clinical research.

- Accuracy was evaluated by comparing cleaned data against verified source records, measuring the extent to which values were corrected or retained without distortion. Beyond statistical improvement, higher accuracy directly supports ethical and regulatory compliance, as it reduces the risk of reporting erroneous clinical outcomes and ensures adherence to Good Clinical Practice (GCP) standards.
- Completeness was assessed by tracking missing data before and after cleaning. The framework's profiling and imputation mechanisms substantially reduced incomplete records. This improvement enhances usability for clinical decision-making, since fewer missing patient histories support better longitudinal tracking of treatment outcomes and more reliable research insights.
- Consistency was evaluated through cross-field validation and time-sequencing checks. By enforcing harmonized value ranges and identifying internal contradictions (e.g., overlapping clinic visit dates), the framework strengthened data integrity across multi-site trials, thereby minimizing risks of protocol deviations and facilitating more robust multi-center analyses.

A comparative analysis was conducted between the results from manual data cleaning (Excel-based and scripting) and the proposed framework. The findings revealed that while manual methods occasionally missed hidden inconsistencies or induced new errors through copy-paste duplication, the proposed framework ensured uniform execution of validation rules and reduced subjectivity.

Table 4.1: Quality Score Assessment Summary

Metric	**Manual Method Score (%)**	**Framework Score (%)**
Accuracy	75	94
Completeness	68	90
Consistency	70	93
Average	**71**	**92**

This comparative table clearly indicates a 21% increase in overall quality score, validating the framework's impact in enhancing not only data quality but also regulatory compliance, decision-making usability, and trial data integrity.

4.4 Synthesis of Findings and Framework Alignment

This study demonstrated that existing cleaning practices at Masafu Hospital were fragmented, repetitive, and prone to subjectivity. By translating these insights into a six-stage contextual framework, the research established a structured, modular, and user-centered approach.

The framework bridges the gap between manual, error-prone processes and the need for automated, auditable, and scalable solutions in clinical research. Its modular stages—data collection, profiling, preprocessing, cleaning, enhancement, and quality scoring—standardize workflows while accommodating the realities of low-resource settings.

Validation results confirm its dual contribution:

1. **Enhanced data quality** (↑21% average score), ensuring improved regulatory compliance, data integrity, and clinical reliability.
2. **Improved usability**, particularly for non-technical staff, through guided steps, SQL templates, and embedded validation rules.

This synthesis confirms the framework's practical relevance to **real-world clinical trial workflows**, offering not only a technical solution but also a sustainable and context-sensitive approach to data governance and clinical data science.

CHAPTER FIVE

VALIDATION AND PRACTICAL APPLICATION OF THE PROPOSED FRAMEWORK

5.1 Overview of the Validation Approach

This chapter presents the validation phase of the contextual data cleaning framework developed in the study. The validation was conducted using real-world data from the DPSP (Dihydroartemisinin-Piperaquine and Sulfadoxine-Pyrimethamine) clinical trial, conducted at Masafu Hospital. The primary aim of this phase was to assess the effectiveness, usability, and contextual relevance of the framework in addressing persistent data quality challenges in clinical research.

The validation process also aligns with the third specific objective of this study:

To validate the effectiveness and usability of the proposed framework using real clinical research data from the DPSP study.

The framework's implementation was mapped onto the full data lifecycle from data collection to quality scoring with particular focus on clinical data cleaning challenges common in malaria prevention trials in pregnancy.

5.2 Application of the Framework to the DPSP Dataset

5.2.1 Structured Validation Workflow

The validation was carried out in a stepwise manner to reflect how each stage of the framework addresses specific data quality challenges:

- Data Collection & Profiling: At this stage, core clinical variables such as gestational age, hemoglobin levels, and drug administration dates were assessed. Profiling highlighted key inconsistencies including abnormal drug dosages, out-of-range lab results, and missing visit records. These issues were prevalent in the DPSP dataset due to its one-site setup and fast-paced clinical environment.

- Data Preprocessing: Preprocessing operations included harmonizing date formats, aligning visit records to pregnancy trimesters, and transforming string fields to appropriate numerical data types. These transformations were crucial in ensuring the downstream processes remained statistically valid.
- Data Cleaning: The framework successfully detected and resolved duplicate patient records, inconsistencies in malaria testing data, and biologically implausible values such as negative gestational age progression. Cleaning activities were guided by clinical domain knowledge and Good Clinical Practice (GCP) standards.
- Data Enhancement: This phase involved generating derived variables such as "dose-to-weight ratios" and composite endpoints combining malaria recurrence and birth outcomes. These enhancements added analytical value and aligned the data with trial reporting requirements.

5.2.2 SQL-Driven Quality Control

SQL was employed extensively to implement core validation rules:

- Duplicate Identification: A patient-centric relational model enabled detection of erroneously repeated antenatal contacts within trimesters. SQL joins were used to filter out exact matches and flag probabilistic duplications for manual review.
- Missing Value Analysis: SQL conditional filters were implemented to distinguish between acceptable protocol-missing values (e.g., missed visits) and unjustified omissions (e.g., skipped laboratory tests).
- Outlier Detection: SQL-based checks flagged hemoglobin readings and parasite densities outside acceptable biological ranges. These were cross-referenced with patient clinical notes for plausibility assessment.
- Gestational-Specific Remediation: Cleaning rules were tailored for pregnancy, where variables like weight gain, drug intervals, and fetal outcomes were monitored longitudinally. For example, a sudden drop in hemoglobin prompted verification against patient clinical history to rule out malaria-induced anemia versus recording errors.

5.2.3 Quality Score Analysis

A data quality score was computed based on three critical dimensions: Accuracy, Completeness, and Consistency. The following outcomes were observed:

Metric	Before Cleaning (%)	After Framework Cleaning (%)
Accuracy	75	94
Completeness	68	90
Consistency	70	93
Average Score	**71**	**92**

This significant improvement confirms the framework's robustness and its ability to meet regulatory-grade data quality thresholds.

5.3 Clinical Relevance and Practical Utility

The framework proves to be highly adaptable to clinical data peculiarities, especially within low-resource environments:

- Context Sensitivity: Recognizes and adapts to pregnancy-related clinical parameters.
- Efficiency: Reduces reliance on manual tools like Excel, while maintaining transparency through SQL logs.
- Scalability: Applicable to larger multi-site trials with similar relational database structures.

Its built-in flexibility enables researchers to extend the model to various domains such as maternal health, infectious disease surveillance, and routine health data systems.

CHAPTER SIX

CONCLUSIONS, CONTRIBUTIONS, AND RECOMMENDATIONS

6.1 Conclusion

This study set out to investigate and solve critical challenges associated with data cleaning in clinical research. Using the DPSP clinical trial as a real-world context, the study successfully designed, developed, and validated a contextual data cleaning framework.

The framework's unique structure emphasizing profiling, preprocessing, modular cleaning, and SQL-based automation enables clinical research teams to elevate the quality, consistency, and usability of data. Its emphasis on user involvement ensures real-world practicality and adoption readiness.

Despite these successes, the study was limited by its focus on structured datasets only. The framework's performance with unstructured or semi-structured data remains untested and offers a pathway for future research.

6.2 Contributions to Knowledge and Practice

This research makes the following significant contributions:

- Theoretical Contribution: Introduces a novel, user-centered conceptual framework that combines domain knowledge with algorithmic logic. It fills a gap in literature regarding tailored data cleaning for clinical trials in resource-limited settings.
- Practical Contribution: Offers a ready-to-use framework that enhances data accuracy, completeness, and consistency. Clinical researchers can implement the model directly using SQL or through integration with electronic data capture systems.
- Methodological Contribution: Demonstrates the feasibility of blending qualitative user feedback with quantitative data validation techniques, thus proposing a hybrid validation approach for health informatics.

6.2.1 Originality and Contribution (Synthesis)

The originality of this study lies in bridging the gap between theoretical models of data quality and the practical realities of LMIC clinical research environments. Specifically:

1. **Domain clarity:** The study establishes itself within clinical data quality management in low-resource settings, informed by socio-technical systems theory.
2. **Theoretical contribution:** It extends existing data quality and socio-technical frameworks by adapting them to contexts where infrastructure, staffing, and workflows differ markedly from high-income research environments.
3. **Practical contribution:** It provides a validated, user-centered data cleaning framework that demonstrably improves accuracy, completeness, and consistency (average quality score ↑21%) while remaining usable for non-technical staff.
4. **Objective alignment:**
    - Objective 1 (review) → contributed a synthesized gap between existing frameworks and LMIC needs.
    - Objective 2 (design) → produced a novel six-stage framework that integrates technical rigor with socio-technical feasibility.
    - Objective 3 (validation) → demonstrated both improved data quality metrics and regulatory/ethical alignment with GCP through usability testing.

This dual theoretical and practical contribution ensures that the study not only advances scholarly discourse but also provides a deployable tool for immediate impact in clinical research.

6.3 Recommendations for Future Research

To enhance the framework's reach and utility, the following directions are proposed:

- Expand to Unstructured Data: Future work should explore extending the framework to support free-text, images, and semi-structured data formats such as XML or JSON. This shall enable applicability in broader data ecosystems including digital health and EHRs.
- Robust Outlier Management: Rather than simply excluding outliers, future studies should implement in-depth analytical models to determine root causes and clinical implications of extreme values.
- Operational Integration: Real-world deployment into clinical research software (e.g., REDCap, OpenClinica) will enable assessment of long-term usability and cost-benefit.
- Cost-Benefit Analysis: Research should evaluate implementation costs versus data quality gains to inform decisions in budget-constrained environments.

6.4 Final Remarks

The study affirms that a contextual, rule-based, and user-sensitive data cleaning framework significantly enhances data quality in clinical research. This framework bridges the gap between theory and practice, offering not just an academic model but a deployable tool for improving public health evidence especially in regions like Uganda where research infrastructure is rapidly evolving.

Through continued refinement and testing, this framework has the potential to become a standard reference model for data quality assurance in both national and global clinical studies.

Reference List

1. Acharya, B., Shrestha, S., & Khanal, A. (2023). Ensuring data quality in clinical research: Methods, tools, and challenges. International Journal of Clinical Research, 18(2), 45–58. https://doi.org/10.1016/ijcr.2023.02.004
2. Bohannon, P., Fan, W., Geerts, F., Jia, X., & Kementsietsidis, A. (2017). Declarative data cleaning: Language, model, and algorithms. Proceedings of the VLDB Endowment, 4(11), 1178–1189. https://doi.org/10.14778/2002974.2002976
3. Calabrese, F. D. (2018). Data quality management in clinical settings: Strategic and operational perspectives. Journal of Healthcare Informatics, 12(4), 301–317. https://doi.org/10.1177/1943401818771284
4. Das, M., Ghosh, A., & Ghosh, A. (2020). ERACER: Entity resolution using data cleaning and record linkage. Journal of Big Data, 7(1), 1–14. https://doi.org/10.1186/s40537-020-00337-3
5. DeLone, W. H., & McLean, E. R. (2016). Information Systems Success: The quest for the dependent variable. Journal of Management Information Systems, 33(1), 5–23. https://doi.org/10.1080/07421222.2016.1177523
6. Fatima, M., & Ali, A. (2017). Survey of data preprocessing techniques in data mining. International Journal of Computer Applications, 162(10), 34–39. https://doi.org/10.5120/ijca2017913271
7. Gesicho, M. B., & Were, M. C. (2020). Data quality in health research: A framework for quality assurance in resource-constrained settings. BMC Health Services Research, 20(1), 1–10. https://doi.org/10.1186/s12913-020-05995-4
8. Gudivada, V. N., Apon, A., & Ding, J. (2017). Data quality considerations for big data and machine learning: Going beyond data cleaning and transformations. International Journal on Big Data Intelligence, 4(2), 104–126. https://doi.org/10.1504/IJBDI.2017.084360
9. Hossain, M. I. (2019). Ethical issues in health data analytics and clinical trials. Asian Bioethics Review, 11(3), 285–298. https://doi.org/10.1007/s41649-019-00086-3
10. Idzwan, M. I. (2021). Compliance framework for healthcare data management in developing countries. Health Policy and Technology, 10(2), 100520. https://doi.org/10.1016/j.hlpt.2021.100520
11. Ilyas, I. F., & Chu, X. (2019). Data cleaning. Association for Computing Machinery (ACM). https://doi.org/10.1145/3299903
12. Jakobsen, J. C., Gluud, C., Wetterslev, J., & Winkel, P. (2017). When and how should multiple imputation be used for handling missing data in randomised clinical trials – a practical guide with flowcharts. BMC Medical Research Methodology, 17(1), 1–10. https://doi.org/10.1186/s12874-017-0442-1
13. Kraska, T., Franklin, M. J., & Madden, S. (2019). ActiveClean: Interactive data cleaning for statistical modeling. Proceedings of the VLDB Endowment, 9(12), 948–959. https://doi.org/10.14778/2994509.2994519
14. Lane, N. D., Georgiev, P., & Qendro, L. (2023). Healthcare data from wearables: Challenges and opportunities. IEEE Pervasive Computing, 22(1), 22–30. https://doi.org/10.1109/MPRV.2023.3242781
15. Lee, S., Roh, G., Kim, J., Lee, Y. H., Woo, H., & Lee, S. (2023). Effective data quality management for electronic medical record data using SMART DATA. *International Journal of Medical Informatics, 180*, 105262.

16. Love, T. E., Zangeneh, T. T., Collins, F. S., & Smith, P. A. (2021). Electronic data capture in clinical trials: Opportunities, challenges, and the path forward. Contemporary Clinical Trials, 106, 106423. https://doi.org/10.1016/j.cct.2021.106423
17. Miao, H., Zhang, X., & Li, J. (2023). Context-aware knowledge-based data cleaning in healthcare systems. Journal of Biomedical Informatics, 138, 104290. https://doi.org/10.1016/j.jbi.2023.104290
18. Muthuraman, K. (2021). Data quality in clinical trials: A conceptual framework for data cleaning and validation. Clinical Trials Journal, 18(3), 205–214. https://doi.org/10.1177/17407745211004379
19. Neutatz, F., Mityagin, A., & Kersten, T. (2019). A holistic framework for rule-based and ML-enhanced data cleaning. In Proceedings of the 28th ACM International Conference on Information and Knowledge Management (CIKM) (pp. 2301–2304). https://doi.org/10.1145/3357384.3358172
20. Phan, T. T., Vo, Q. A., & Nguyen, T. D. (2020). Improving data quality in clinical research using information quality frameworks. BMC Medical Informatics and Decision Making, 20(1), 1–13. https://doi.org/10.1186/s12911-020-01136-1
21. Pindya, M. J. (2024). Adapting data cleaning frameworks for low-resource clinical environments. East African Journal of Health Informatics, 6(2), 45–61.
22. Pradhan, B., Shah, S., & Alavi, A. (2021). Applying Total Data Quality Management (TDQM) in public health surveillance: Lessons from pandemic response. Journal of Public Health Management, 39(5), 401–409. https://doi.org/10.1097/PHH.0000000000001216
23. Rahm, E., & Do, H. H. (2020). Data cleaning: Problems and current approaches. IEEE Data Engineering Bulletin, 23(4), 3–13. http://dbs.uni-leipzig.de/file/rahm00data.pdf
24. Rekatsinas, T., Chu, X., Ilyas, I. F., & Ré, C. (2018). HoloClean: Holistic data repairs with probabilistic inference. Proceedings of the VLDB Endowment, 10(11), 1190–1201. https://doi.org/10.14778/3137628.3137631
25. Ridzuan, F., Zainon, W. M. N. W., & Noor, N. M. (2019). A review on data quality: Definitions, dimensions, and approaches. International Journal of Advanced Computer Science and Applications, 10(6), 13–19. https://doi.org/10.14569/IJACSA.2019.0100613
26. Ross, P., Middleton, J., & Price, M. (2015). Governance and compliance in clinical data management: A review of FDA and EMA frameworks. Journal of Clinical Research Best Practices, 11(5), 1–12. https://www.jclinicalres.com
27. Shi, X. (2021). Semantic data validation in biomedical systems. Journal of Medical Systems, 45(7), 1–10. https://doi.org/10.1007/s10916-021-01789-2
28. Van den Broeck, J., Cunningham, S. A., Eeckels, R., & Herbst, K. (2020). Outlier detection and management in clinical research datasets. PLoS Medicine, 17(10), e1003302. https://doi.org/10.1371/journal.pmed.1003302
29. Wong, J. L. (2016). Ontological constraints and semantic rule-based cleaning in medical data systems. Health Information Science and Systems, 4(1), 13–22. https://doi.org/10.1007/s13755-016-0018-2
30. Yakout, M., Elmagarmid, A., Ouzzani, M., Rahm, E., & Ilyas, I. F. (2017). Katara: A data cleaning system powered by knowledge bases and crowdsourcing. Proceedings of the 2017 ACM SIGMOD International Conference on Management of Data, 1247–1261. https://doi.org/10.1145/3035918.3064031

Appendices

Appendix 1: Questionnaire

Dear Participant,

We are conducting a research study aimed at designing a contextual framework for data cleaning in clinical research in partial fulfillment for the award of Master of Science in Information Systems at Uganda Martyrs University. Your participation is valuable as it will contribute significantly to our knowledge and potential improvements in this field.

This questionnaire will take approximately 10 minutes to complete. Please be assured that all responses will be kept strictly confidential and will only be used for research purposes. There is no right or wrong answer; we are only interested in your honest opinions and experiences.

Your participation is entirely voluntary, and you may withdraw at any time without penalty. If you have any questions about the research, please do not hesitate to contact us at pius.kavuma@stud.umu.ac.ug or +256700115621

1) What is your designation

Check all that apply.

☐ ICT/IT Officer

☐ Database Administrator ☐ Data Manager

☐ Data Officer

PART I: GENERAL QUESTIONS

2)How long have you been in this role?

Mark only one oval.

◯ 0-5years ◯ 5-10years

◯ 10-15-49 years

◯ 15-20 years

Data Cleaning in Clinical Research

3) Do you do data collection at your organization, if yes what methods are used?

Check all that apply.

☐ Case Report Forms (CRFs) ☐ Electronic Data Capture (EDC)

☐ Clinical Assessments /Examinations ☐ Direct Observations

4)What sources are used to store the collected data?

Mark only one oval.

◯ Electronic data capture systems ◯ Cloud based storage

◯ Physical Archives ◯ Local databases

5) Do you do data cleaning in your organization, If yes what methods are used?

Mark only one oval.

◯ Yes ◯ NO

6) What methods are used for data cleaning in your organization?

Check all that apply.

☐ Excel ☐ STATA

☐ Python

☐ R Programming

7)Are you satisfied with the quality of data obtained using the above selected methods.

Mark only one oval.

◯ Very satisfied

◯ Moderately satisfied ◯ Not satisfied

8) What are the key challenges you face when using these methods?

Check all that apply.

☐ Prone to Human Errors ☐ Repetitive tasks

☐ Complicated syntax and commands

9)What are the main challenges associated with data cleaning in clinical research?

Mark only one oval.

◯ Duplicate records ◯ Missing data ◯ Outliers

◯ Unstructured data

10)Have you encountered situations where existing data-cleaning techniques failed to improve data quality?

Mark only one oval.

◯ Yes ◯ No

11) Do inconsistencies in data cleaning at clinical sites affect data quality?

Mark only one oval.

◯ yes ◯ No

12)How do you deal with the challenges of data cleaning in your organization?

_____

_____

_____

_____

_____

13) What best practices can be implemented to ensure real-time data cleaning in clinical trials?

Check all that apply.

☐ Real Data Reconciliation

☐ Use of electronic data capture system ☐ Use of automated data validation Rules

☐ Use of Role based access for data governance

14) What automated tools or software do you currently use for data cleaning?

Check all that apply.

☐ Power Query ☐ OpenRefine ☐ Pandas

Other: ☐ _____

15)How satisfied are you with these tools?

Check all that apply.

☐ Very satisfied

☐ Moderately satisfied ☐ Not satisfied

16) What improvements or new techniques would you recommend for clinical data cleaning?

17)Do you have a data cleaning framework at your organization.

Mark only one oval.

◯ yes ◯ No

18)If yes, is the framework being followed.

Mark only one oval.

◯ Yes ◯ No

19)If yes, is there any training about the framework.

Mark only one oval.

◯ yes ◯ No

20)Does the framework involve all the stakeholders.

Mark only one oval.

◯ yes ◯ No

21)How can data cleaning frameworks ensure compliance with Good Clinical Practice (GCP) and other regulatory standards?

Appendix 2: Interview Guide for Key Stakeholders

Title: *Semi-Structured Interview Guide for Stakeholders in Clinical Data Management – Masafu Hospital*

Research Title: *Designing a Contextual Framework for Data Cleaning in Clinical Research*

Introduction:

Good morning/afternoon. Thank you for agreeing to participate in this interview. I am conducting a research study as part of my Master's in Information Systems at Uganda Martyrs University. The goal of the study is to design a contextual framework for data cleaning in clinical research. Your views and experiences are very important to this process.

This interview will take about 30–45 minutes. Your responses will be kept strictly confidential and used solely for academic purposes. You may decline to answer any question or withdraw at any point. With your permission, I would like to record this interview to ensure accuracy.

Section A: Background Information

1. **Can you briefly describe your current role in this organization?**
   (Probing: What are your primary responsibilities in data management?)
2. **How long have you worked in clinical research or data management?**

Section B: Data Collection and Storage Practices

3. **Can you describe the methods your organization uses for data collection in clinical research?**
   (Probing: CRFs, EDC systems, direct observation, lab records, etc.)
4. **What are the main platforms or systems where collected data is stored?**
   (e.g., local databases, EHRs, paper records, cloud systems)

Section C: Data Cleaning Processes

5. **Do you or your team perform data cleaning? If yes, can you walk me through the process?**
   (Probing: At what stage is cleaning done? How often? By whom?)
6. **What tools or software do you currently use for cleaning?**
   (Excel, SQL, Python, STATA, OpenRefine, etc.)
7. **What challenges do you face when cleaning data?**
   (e.g., duplicate entries, missing values, time consumption, staff capacity, software limitations)
8. **Have you ever encountered a situation where your current methods failed to improve data quality? Could you describe that experience?**

Section D: Views on Automation and Frameworks

9. **How do you think automated data validation tools could improve your current data cleaning workflow?**
   (Probing: Have you used any automation tools? Which ones?)
10. **Do you have a formal data cleaning framework in place at your institution? If yes, how well is it followed?**
11. **Is there any training provided on this framework or data quality protocols?**
12. **In your view, what are the best practices to ensure real-time and reliable data cleaning during clinical trials?**
    (Probing: role-based access, validation rules, real-time reconciliation, etc.)

Section E: Regulatory and Ethical Considerations

13. **How does your organization ensure that data cleaning aligns with Good Clinical Practice (GCP) guidelines and regulatory standards?**
14. **Do you think current data management practices meet ethical standards, particularly around data accuracy and participant confidentiality?**

Section F: Recommendations

15. **What improvements or features would you recommend in a new data cleaning framework tailored for clinical research in settings like Masafu Hospital?**
16. **What would make a data cleaning tool or framework more user-friendly and acceptable to your team?**