IMPACT OF DATA CAPTURING METHODOLOGY ON DATA QUALITY: FOCUSING ON OPTICAL CHARACTER RECOGNITION (OCR) CASE OF OPTION B+ EVALUATION IN MALAWI

MASTER OF SCIENCE (INFORMATICS) THESIS

EFRIDA KUTENGULE GHOBEDE

UNIVERSITY OF MALAWI

MARCH 2023

Impact of data capturing methodology on data quality: Focusing on Optical Character Recognition (OCR)– Case of Option B+ Evaluation in Malawi

MSc (Informatics) Thesis

By

EFRIDA KUTENGULE GHOBEDE

BSc– University of Malawi

Submitted to the Department of Mathematical Sciences, Faculty of Science,
in partial fulfillment of the requirements for the degree of
Master of Science in Informatics

UNIVERSITY OF MALAWI

MARCH 2023

DECLARATION

I hereby declare that this thesis/dissertation is my original work, which has not been submitted to any other institution for similar purposes. Where other people's work has been used, acknowledgments have been made.

EFRIDA KUTENGULE GOBEDE

_____

Full Legal Name

_____

Signature

_____

Date:

CERTIFICATE OF APPROVAL

The undersigned certify that this thesis represents the student's own work and effort and has been submitted with our approval.

Signature: _____ Date: _____

Chipo Kanjo, PhD (Associate Professor)

Main Supervisor

Signature: _____ Date: _____

Patrick Albert Chikumba, Dr

Co-supervisor

DEDICATION

Dedicated to the memory of my late parents, Lyall McLeornard and Elizabeth Kutengule, they always wanted me to succeed in life, I will always make them proud, I promise.

ACKNOWLEDGEMENTS

I wish to acknowledge efforts of my supervisors, Dr. Chipo Kanjo and Patrick Chikumba, for their tireless efforts in providing advice, guidance, helpful suggestions and critical reviews during the thesis exercise.

I am deeply indebted to the Management Sciences for Health and the Ministry of Health for approving and supporting this research. Not forgetting lecturers and my fellow Master of Science in Informatics students for their valuable contributions to the success of my studies.

I sincerely thank all friends and relatives, who have contributed to my studies and more especially those who have encouraged me to complete this research.

ABSTRACT

This research investigated the impact of data collection methodology on data quality in the clinical research setting of Malawi's Prevention of Mother to Child Transmission Program (PMTCT). Employing questionnaires, interviews, and literature reviews, the study focused on Optical Character Recognition (OCR) technology for efficient data capture. Despite the methodology's benefits, including ease of use, the findings revealed data inconsistencies, gaps, and inaccuracies when compared to source documents. The study's scope was limited to PMTCT clinical evaluations in three health facilities, and budgetary constraints constrained the sample size. The researcher recommends further research comparing various data collection methodologies, especially in clinical evaluations, and suggests exploring the integration of AI or advanced technology to enhance character recognition for complex data types. While acknowledging the study's constraints, such as exclusive emphasis on PMTCT and a limited sample size, the research underscores the need for comprehensive investigations into the impact of different data collection methods on data quality in clinical settings, aiming to improve overall research practices and outcomes.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ABBREVIATIONS

ANC	Antenatal Care
ART	Anti-Retroviral Therapy
BHT	Baobab Health Trust
CDC	National Statistical Office
CMED	Central Monitoring and Evaluation Division
DHS	Demographic Health Survey
DQ	Data Quality
DQA	Data Quality Assessment
DQAF	Data Quality Assessment Framework
EDC	Electronic Data Capture
HTC	HIV Testing and Counseling
ICT	Information and Communication Technology
MCH	Maternal and Child Health
NEMAPP	National Evaluation of PMTCT Program
OCR	Optical Character Recognition
OMR	Optical Mark Recognition
PMTCT	Prevention of Mother to Child Transmission Program
RHIS	Routine Health Information System
SOP	Standard Operating Procedures

CHAPTER 1

INTRODUCTION

This chapter gives a general overview of the study by presenting a detailed background, the context in which this study domain focused on and the problem it is addressing. It further describes the objective and research questions. Finally, the chapter explains what motivated the researcher to carry out this study and the organization of the entire thesis.

1.1 Introduction

Data capturing is the method of putting a document into an electronic format. The known methods of capturing data are manual and automated. According to Jameson (2022), Manual Data Capture is a method that uses manual keying of required data from written forms into a computer for digitized access. Manual data capture depends on human labor making it susceptible to errors or data omissions, the very reason why automated data capture technology is becoming an ideal solution.

The automated method typically includes Optical Character Recognition (OCR), Optical Mark Recognition (OMR), Intelligent Character Recognition (ICR), bar codes, QR codes and magnetic stripes (Hamzah, 2018). Further, Jameson (2022) defines these different automated data capture methods as follows:

OCR: Optical Character Recognition technology identifies machine-generated characters and typefaces to extract text from scanned documents, PDF files, etc. for editing.

ICR: Intelligent Character Recognition is the next-generation technology of OCR. It is designed to read handwritten characters of any font from forms and convert them into meaningful data for further use. Banks and finance organizations adopt ICR technology

solutions for their businesses. One example of OCR technologies is Teleform which is a software solution designed to automate the capture of data from paper forms and electronic documents.

Barcodes and QR codes: Barcode technology contains encrypted information as 1D barcodes that are read using a barcode scanner. The technology is accurate and used on shop floors to track inventory or employee logs, check patient details in hospitals, print bank passbooks, and so on. Quick Response (QR) codes, also called 2D barcodes, are more complex. They are useful to capture documents, webpages, etc. for a variety of purposes. QR codes are popularly used in shop establishments, courier services, advertising, product packaging, etc.

OMR: The Optical Mark Reading technology is an electronic data capture method that identifies human-filled data such as darkened fields or checkboxes in a document. Its high accuracy makes it an ideal tool for use in survey forms, ballots, or objective-type examinations.

Magnetic Stripe Cards: These cards contain encoded data via magnetic stripes that are decoded using reader devices. They are quite safe and used in credit/ debit cards, ID cards, access cards in hotel rooms, and transport cards.

Automated data capturing is rapidly becoming an integral part of the businesses. The system not only saves the time but also increases the speed and accuracy over manually entered data. Nevertheless, data capturing using this methodology has also an impact on data quality.

Quality data is defined as data that is accurate, valid, consistent, reliable, legible, complete and available in a timely manner to decision makers. Generally, data is considered of high or good quality if they are fit for their intended use in operations, decision making and planning (Redman, 2001).

In clinical research, the use of high quality and effective data assists in the accurate evaluation of the impact of health interventions to the public. However, the use of poor data quality leads to imbalances in the health sector which includes; inaccurate health performance measurements, inappropriate allocation of health funding, and failure in public health surveillance (Chen, 2014). Data of poor quality therefore makes it hard for users to make valid decisions. Clinical research endeavors to discover answers to the research query through the collection of data to either validate or refute a hypothesis, aiming for a deeper understanding. The results have a very significant impact to the health sector specially to promote sustainability of health systems and support patient care at the point of service, therefore, the quality of data generated plays an important role, especially supporting informed decision making.

Data quality (DQ) is usually understood as a multi-dimensional concept. The dimensions represent the views, criteria, or measurement attributes for data quality problems that can be assessed, interpreted, and possibly improved individually (Sattler, 2009).

Striving for high-quality data, it is imperative to enhance data quality. Improving data quality is much more than "clearing" out inaccurate data, it is a dedicated process that involves considering both sociological and technological aspects. These considerations need to be applied in the whole data processing cycle, which includes tools and methods used to collect the data; the whole collection processes; handling of the data; storage and retrieval. When determining inaccurate data, it is important to understand all key dimensions of data quality which includes completeness, consistency, conformity, accuracy, integrity and timeliness.

These key dimensions of data quality enhance analysts to determine the scope of the underlying root causes and to plan on the ways that these tools can be used to address data quality issues. Henceforth, it is valuable to understand these common data quality dimensions defined by Sattler (2009).

Completeness: Is all the requisite information available? Are data values missing, or in an unusable state? In some cases, missing data is irrelevant, but when the information that is missing is critical to a specific business process, completeness becomes an issue.
Conformity: Are there expectations that data values conform to specified formats? If so, do all the values conform to those formats? Maintaining conformance to specific formats is important in data presentation, aggregate reporting, search, and establishing key relationships.
Consistency: Do distinct data instances provide conflicting information about the same underlying data object? Are values consistent across data sets? Do interdependent attributes always appropriately reflect their expected consistency? Inconsistency between data values plagues organizations attempting to reconcile between different systems and applications.

Accuracy: Do data objects accurately represent the "real-world" values they are expected to model? Incorrect spellings of product or person names, addresses, and even untimely or not current data can impact operational and analytical applications.
All in all, data quality in public health has different definitions from different perspectives. These include: "fit for use in the context of data users", "timely and reliable data essential for public health core functions at all levels of government", and "accurate, reliable, valid, and trusted data in integrated public health informatics networks"

1.2 Research Context (The Case)

1.2.1 Geographic profile

The research was conducted in Malawi. Malawi has 28 districts and research was conducted in two districts – Lilongwe and Dedza. Three health facilities were used for data collection – two (Kawale and Area 25 health centers) from Lilongwe and Dedza district hospital These were selected facilities under a subset of the National Evaluation of Malawi PMTCT Program (NEMAPP) clinical study sites.

1.2.2 PMTCT in Malawi

Malawi has one of the highest HIV prevalence in the world despite the impressive progress the country has made in controlling its HIV epidemic in recent years (MOH, 2018).

The government, in 2013, revised the national policies for **Prevention of Mother to Child Transmission** (PMTCT), termed as 'Option B+' to rapidly increase coverage of Anti-Retroviral Therapy intake of pregnant and breastfeeding mothers. All HIV-infected pregnant or breastfeeding women are offered lifelong ART to reduce the risk of mother-to-child HIV transmission.

1.2.3 Evaluation of Option B+ in Malawi

In an effort to support Government initiatives in combating HIV and AIDS, the Ministry of Health through Management Sciences for Health (MSH) initiated a National Evaluation of Malawi PMTCT Project (NEMAPP) to assess effectiveness of Option B+. The primary objective of NEMAPP was to measure rates of mother-to-child HIV transmission over time as the mother breastfeeds, measure HIV-free survival after the mother has stopped lactating and also measure longer term survival and virological and clinical outcomes after the initiation of ART to all the (MoH, 2014) children which are infected. Option B+ has been adopted by several other countries in the region and was included in the World Health Organization's updated 2013 guidelines (WHO, 2003).

The evaluation of Option B+ used both predesigned manual questionnaires and the routine data collection tools that are utilized by MoH in ART during service delivery. Teleform technology which uses Optical Character Recognition (OCR), Intelligent Character Recognition (ICR), Optical Mark Recognition (OMR) and barcode recognition was also used to transcribe the predesigned manual questionnaires into the database. All screening, enrollment and questionnaire forms, additional consent and questionnaire forms were transformed into a scannable Tele-forms system to reduce human error during data entry and speed up data entry.

Tele-form solution provides fast and powerful data capture capabilities as recognition technologies interpret machine and hand-printed marks which later converts them into data hence making data capturing efficient and effective. However, the challenge of using character recognition software is mainly on identifying and removing errors produced by stray marks, respondent corrections, or improperly completed fields (Jenkins TM1, 2014).

The methodology used for collecting, cleaning, storing, monitoring, reviewing, and reporting on data determines the quality and utility of the data for meeting its goals. Quality assurance, on the other hand, aims to assure that the data were, in fact, collected in accordance with the procedures and that the data stored in the database meet the requisite standards of quality, which are generally defined based on the intended purposes (Gliklich, 2014).

1.3 Problem Statement

Although research efforts on data quality presented in the existing literature have addressed a significant advance in history, the existing literature on data quality assessment in healthcare, has focused primarily on assessing the quality of data on two main dimensions, accuracy and completeness (Arts, 2002).There has not been much research conducted to assess the challenges of data consistency as one of the less studied data quality dimensions especially in evaluation studies which uses Optical Character Recognition (OCR) to transcribe data as a collection methodology. This. motivated the researcher to assess data quality, with a deep dive into data inconsistencies, to understand if technological aspects impact data quality.

Errors in clinical research databases are common but relatively little is known about their characteristics and optimal detection and prevention strategies (Saveli I. Goldberg, 2008). Considering that such data errors originate from collecting huge amounts of data from facilities; processed and presented for decision making, the probability of drawing wrong conclusions on the findings is high if quality is compromised. Thus, this research sought to explore data quality gaps, causes and the impact on the final outcome.

1.4 Objectives of the Study and Research Questions

In relation to the background, the main objective of the study was to assess the impact of data collection methodology on data quality in the clinical evaluation domain.

1.4.1 Specific objectives

The specific objectives of this research were as follows: -

1. To identify opportunities and challenges associated with use of OCR technologies as a data capturing and transcription methodology.
2. To establish the causes of data errors.
3 To analyse the impact of the data errors on the final outcome.
4 To develop strategies to address the challenges and harness the opportunities identified in order to improve future use.

1.4.2 Research Questions

The research aimed at answering the following questions: -

1. What opportunities and challenges are associated with use of OCR technologies as a data capturing and transcription methodology?
2. What are the root causes of the errors?
3. What is the impact of data errors on the final outcome
4. What strategies can be adopted to avoid these data quality issues?
5. What are opportunities are identified to improve data quality in future evaluations?

1.5 Research Motivation

Generally, data quality in the health sector is a challenge, including data that is collected to support clinical researchers. Much as efforts are being implemented in training and mentoring data collectors and introduction of technology to promote data quality like the use of electronic data capture technologies like Tele-form is concerned, data quality remains a big problem in the health sector.

In the context of the study area, successful evaluation of OPTION B+ data, amongst other factors, largely depended on the quality of data collected to inform decision makers on effectiveness of the program in Malawi which in turn informed government, donors and clinicians on interventions to combating HIV especially in infants.

There may be many challenges that were encountered during data collection and data processing which could not explicitly establish its source, magnitude and associated impact

on the final outcome unless an assessment was to be carried out. The study aimed at establishing if the methodology used had an impact on data quality.

The researcher felt the study was to reveal opportunities and gaps affecting data quality in clinical research, focusing more on different methodologies used to collect and process data, common causes and establish some best practices which can be adopted or avoided in future interventions. This research potentially focused on reviewing some of the factors affecting data quality with the aim of providing best practices to be implemented in order to reduce bad data which affects valid clinical and administrative decisions.

There might be a requirement to conduct more research in the near future on real time collection versus retrospective data entry and processing using electronic data capture technologies like Tele-form.

1.6 Thesis Organization

The thesis has been organized as follows: Chapter 1 introduces the research by defining concepts discussed in this research. Then it describes the research context, giving insights of Malawi's health sector and initiatives to fight HIV.

Chapter 2 presents the literature review and theoretical framework, similar works on this research topic and also gaps. This discusses relevant literature on data quality with particular interest on how internal and external factors negatively affected data quality. In addition, it discusses the theoretical framework derived from evolution of data quality, describing four stages of data activities and how data evolves: collection, organization.

Chapter 3 presents the research methodology and basically discusses how the research was carried out to arrive at the results conclusion, including details of interviews carried out, observations and questionnaires used. The last chapters, thus 4 and 5 discuss in detail the findings, conclusions and recommendations drawn from the study with reference to the objectives and questions.

1.7 Conclusion of Chapter 1

This chapter provided orientation of the study leading to the problem statement. It introduced data collection methodology and its impact on data quality and several aspects of data quality which has an impact on evidence-based decision making and health interventions to the public. The chapter also discussed the context in which this study domain focused on, an intervention that Malawi's Ministry of Health carried on to evaluate Option B+ and then the problem it is addressing, mainly through the objectives and questions that guides this thesis. It further provided details which motivated the researcher to carry out this research topic. Further discussion on this is found in the literature review and all subsequent chapters.

CHAPTER 2

LITERATURE REVIEW

This chapter discussed a number of literatures the researcher studied which significantly provided in depth understanding of data collection methodology and its impact on data quality then specifically talks about data entry using OCR, its advantage and disadvantages, then it dives into understanding data quality and its several dimensions, with a deep review on internal and external data inconsistencies. The literature on effects of data quality especially on health data was also reviewed. In trying to understand data quality gaps and its impact, the researcher looked at data quality assessment framework in relation to data collection methodology. The research used elements of data assessment framework to address the research questions in Chapter 1.

2.1 How Data Entry Methodology Affects Data Quality in Research Settings

Data quality is of paramount importance in research settings as it directly impacts the reliability, validity, and overall integrity of research findings (NLM, 2009). The methodology used for data entry plays a crucial role in determining the quality of the data collected.

Several data entry methodologies are employed in research settings, ranging from manual entry to automated systems (Barchard, 2011). These methodologies can significantly influence data quality through various mechanisms such as human error, system integration, and data validation processes. Manual data entry involves human operators inputting data into a system. While this method allows for immediate verification of data, it is prone to errors stemming from typographical mistakes, misinterpretation of data, and fatigue. (NLM) highlighted the potential for data quality issues, particularly in larger datasets and studies with a high volume of variables. Automated data entry methods

leverage technology to input data, reducing the risk of human error. These methods include optical character recognition (OCR), barcode scanning, and electronic data capture (EDC) systems. Research by David *et al.* (1999)found that automated methods can improve data accuracy and consistency compared to manual methods. However, the initial setup and maintenance costs can be substantial.

Several factors impact data quality when using different data entry methodologies (Solomon M, 2021). These factors can be categorized into human-related, technological, and organizational aspects. Human factors include the skills, training, and attention to detail of data entry personnel. Research by Solomon *et al.* (2021) highlighted the importance of rigorous training programs to minimize errors caused by human factors. Additionally, the cognitive load of data entry tasks and the potential for multitasking can affect the accuracy of data input (Yen PY, 2017).

The technology used for data entry, such as data entry software and devices, significantly influences data quality (Ping *et al*, 2009). Compatibility, user interface design, and error handling mechanisms all impact the accuracy of data entry. Ping et al (2009) conducted a study to to explore the opportunity of filling this gap through developing and trial of a Personal Digital Assistant (PDA) based data collection/entry system. It evaluated whether such a system could increase efficiency and reduce data transcription errors for public surveillance data collection in developing countries represented by Fiji.

A generic PDA-based data collection software eSTEPS was developed. The software and the data collected using it directly interfaces with EpiData. A field trial was conducted to test the viability of public health surveillance data collection using eSTEPS. The design was a randomised, controlled trial with cross-over design. 120 participants recruited from the Fiji School of Medicine were randomly assigned to be interviewed by one of six interviewers in one of the two ways: (1) paper-based survey followed by PDA survey and (2) PDA survey followed by paper-based survey. Data quality was measured by error rates (logical range errors/inconsistencies, skip errors, missing values, date or time field errors

and incorrect data type). The study showed that user-friendly interfaces and real-time error notifications can enhance data accuracy.

Organizational factors, including data entry protocols, quality assurance procedures, and monitoring mechanisms, play a critical role in maintaining data quality (Elodia Cole, 2006). Elodia Cole, *et al.* (2006) emphasized the importance of standardized protocols and regular audits to identify and rectify data entry errors promptly.

To mitigate data quality issues in research settings, the following best practices are recommended:

- Implement comprehensive training programs for data entry personnel to enhance their skills and reduce human-related errors.
- Utilize automated data entry methods where appropriate to minimize the risk of human errors and enhance efficiency.
- Develop user-friendly data entry interfaces with real-time error detection and correction features.
- Establish standardized data entry protocols and conduct regular quality assurance checks to identify and address errors promptly.
- Foster a culture of data quality within the research team by emphasizing the importance of accurate data collection.

2.2 Optical Character Recognition (OCR)

Data Entry Using Optical Character Recognition (OCR) is a system that scans a sheet of data, which includes numbers and characters, and saves it into .csv files. According (IBM, 2022).

OCR software singles out letters on the image, puts them into words and then puts the words into sentences, thus enabling access to and editing of the original content. It also eliminates the need for manual data entry.

According to Brown (1950), the invention of OCR can be traced back to the mid-20th century, with considerable contributions from various researchers. The primary motivation

behind the invention of OCR was to streamline data entry and information retrieval processes. As businesses and organizations began dealing with an increasing volume of paper-based documents, the need for automated systems to extract and digitize text became apparent.

OCR operates through a multi-step process. Initially, the system scans the input document to create a digital image. Subsequently, the OCR software analyzes the shapes, patterns, and structures within the image to identify characters. Machine learning algorithms and pattern recognition techniques play a crucial role in enhancing the accuracy of character recognition. The recognized characters are then converted into machine-encoded text, making the content editable and searchable (IBM, 2022).

OCR technology holds paramount importance in various fields. It expedites data entry processes, reduces manual effort, and enhances accessibility by converting printed or handwritten text into digital formats. Additionally, OCR is pivotal in digitizing historical documents, archiving records, and facilitating text extraction for search engines and document management systems.

2.3 Data Entry Using OCR

Using the OCR data entry method, there could be advantages as well as disadvantages to it. Some of the advantages include; influencing high productivity within organizations, superior data security and aiding in cost reductions. OCR works best with limited documents and OCR systems are expensive, just to mention a few.

To begin with, the OCR helps with influencing high productivity within organizations. Its software helps businesses to achieve higher productivity by facilitating quicker data retrieval when required. The time and effort which the employees were required to put in for extracting relevant data be channelized to focus on core activities. Besides, employees do not have to make numerous trips to central records room to access the required documents, as they can access them without getting up from their desks (Flatworld, 2023).

Another advantage of the OCR is that it has high superior data security. Flatworld (2023) further indicates that documents are easily prone to loss or destruction. Papers can be misplaced, stolen, or destroyed by natural elements such as moisture, pests, and fire. However, this is not the case with data that is scanned, analyzed, and stored in digital formats. Furthermore, the access to these digital documents can also be minimized to prevent mishandling of the digitized data.

Cost reductions can also be considered as an advantage of the OCR. Using the OCR cuts down expenses needed to hire more professionals to carry out smooth data extraction tasks. This will prove to be a good addition to an organization or business as all documents can be handled without hiring a professional team. This remarkable tool also helps to lower other costs like shipment, printing and copying documents from credible sources. Therefore, OCR helps eliminate the overall cost of investing in the long process of recovering lost data and provides organizations with higher savings in the form of reclaimed office space which would have been otherwise used to store bulk paper files (Arsalan, 2021).

The OCR having its advantages also has its disadvantages. Firstly, data structuring in the OCR does not entirely depend on the OCR system and this can be one of its disadvantages. Suppose users take a picture of their ID document with their smartphone or webcam, multiple steps are required to extract and structure the information. The first step is to precisely recognize what kind of ID document is present. This enables the engine to properly structure the information read with the OCR, which means figuring out the first name, last name, date of birth and any other field of interest. *S*traight OCR without additional AI or technology specifically trained to recognize ID types will lack the requisite accuracy organizations need to fight fraud and deliver a good user experience (Nicolls, 2019).

Another disadvantage of the OCR could be that it works best with limited documents. According to Gilani (2015), the OCR works best with good quality typed documents. Handwritten documents cannot be easily read by OCR software. Likewise, typed fonts that

resemble handwriting as well as non-Latin fonts create many errors during the OCR process. If the document has poor contrast, is creased or dirty, or the text and the background are similar in darkness, then OCR may not work well. OCR has difficulty with documents that have both images and text. Spreadsheets will also produce more errors.
The OCR system itself is very expensive. This can be considered as another drawback to the OCR. (Caeiro, 2018)claims that the cost of acquiring the OCR software license is very expensive and the cost grows depending on the volume and capacity to process, scan and extract information from documents. Caeiro continues to state that the software itself does not do anything hence there is need to deploy the ORC software to the environment. Servers need to be allocated, there is a need for persons to manage them to guarantee operations and backups. All this requires large amounts of income and hence expensive.

2.4 Data Quality

Data are of high quality if they are fit for their intended uses in operations, decision making and planning (Thomas, 2001). Data that are fit for use are free of defects. High quality data can be described as being; accessible, accurate, timely, complete, and consistent with other sources. The observation is that users of data see defects as including inaccurate data, data that are out-of-date, data that are hard to interpret. The most noted defects are “erred data values”. The effects of this defective data produce enormous user dissatisfaction.

Health, just like all other sectors, experiences many challenges with quality data. Researchers and programmer implementers in Tanzania, drawing from other examples from East Africa, found out that PMTCT data are routinely collected in maternal and child health (MCH) clinics in East Africa using paper-based registers corresponding to distinct services within the PMTCT service continuum, this format has inherent limitations with respect to maintaining and accurately recording unique identifiers that can link patients across the different clinics (antenatal, delivery, child), and also poses challenges when compiling aggregate data (Gourlay, 2015).

In order to achieve data quality, most organizations employ approaches to data quality systems which are divided in three steps: -

- Correction of errors and other deficiencies that users find.
- Conducting periodic database clean-ups
- Conducting daily clean-ups as part of operations

This is illustrated in the **figure 1**.

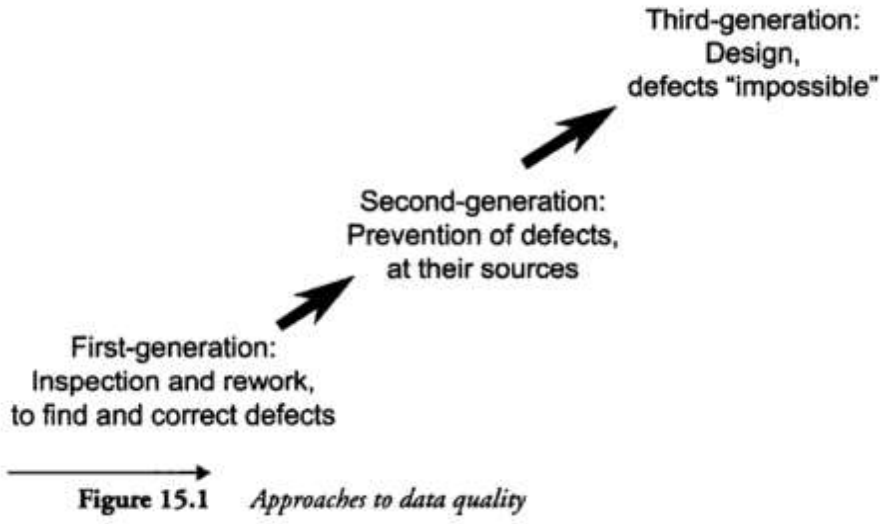


Figure 1: Models embedded in Organizations to respond to data quality problems

The diagram above also describes models or patterns that are being embedded in organizations to respond to data quality problems which aim at "getting better or faster at cleaning up data." However, there is usually not enough clean-up horsepower to get the job done because the hard part of preventing errors is not technical, but rather it is to overcome the emotional and intellectual investment in data clean-up exercise, the practical advice is to prevent errors in the first place which had proved to be a challenge.

Every scientist knows that research results are only as good as the data upon which the conclusions were formed (Nahm, 2012). However, most scientists receive no training in methods for achieving, assessing, or controlling the quality of research data. The **figure 2** indicates the definition, collection, processing and presentation processes and how they impact data and information quality and data use.

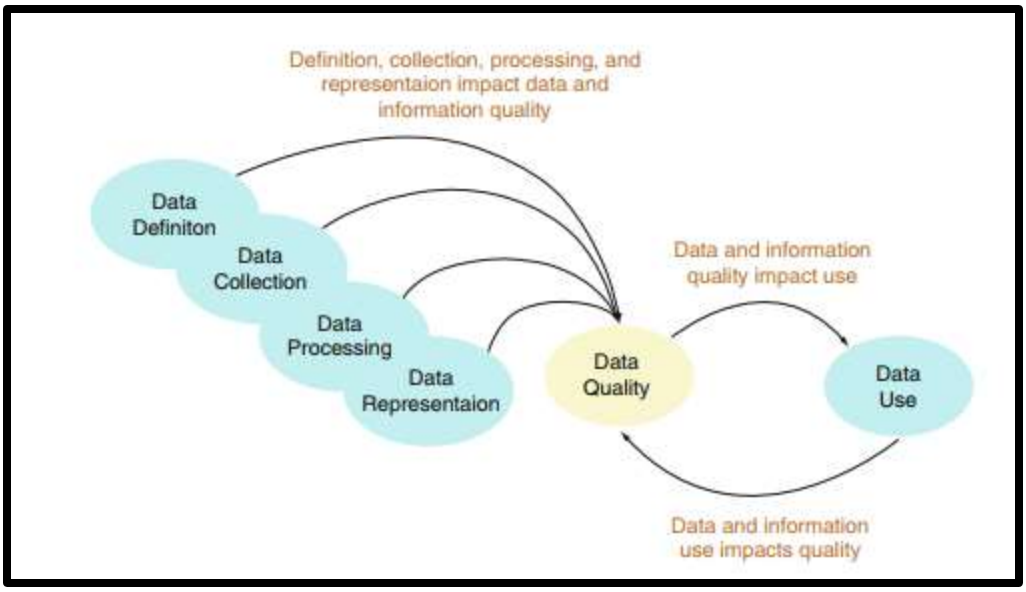


Figure 2: How definition, collection, processing, presentation and use of data impacts quality

2.5 Data Quality Dimensions

A Data Quality (DQ) Dimension is a recognized term used by data management professionals to describe a feature of data that can be measured or assessed against defined standards in order to determine the quality of data (DAMA, 2013). Organizations select the data quality dimensions and associated dimension thresholds based on their business context, requirements, levels of risk etc.

The six core dimensions of data quality are: completeness, uniqueness, timeliness, validity, accuracy and consistency. The dimensions according to DAMA, are presented in **figure 3**:

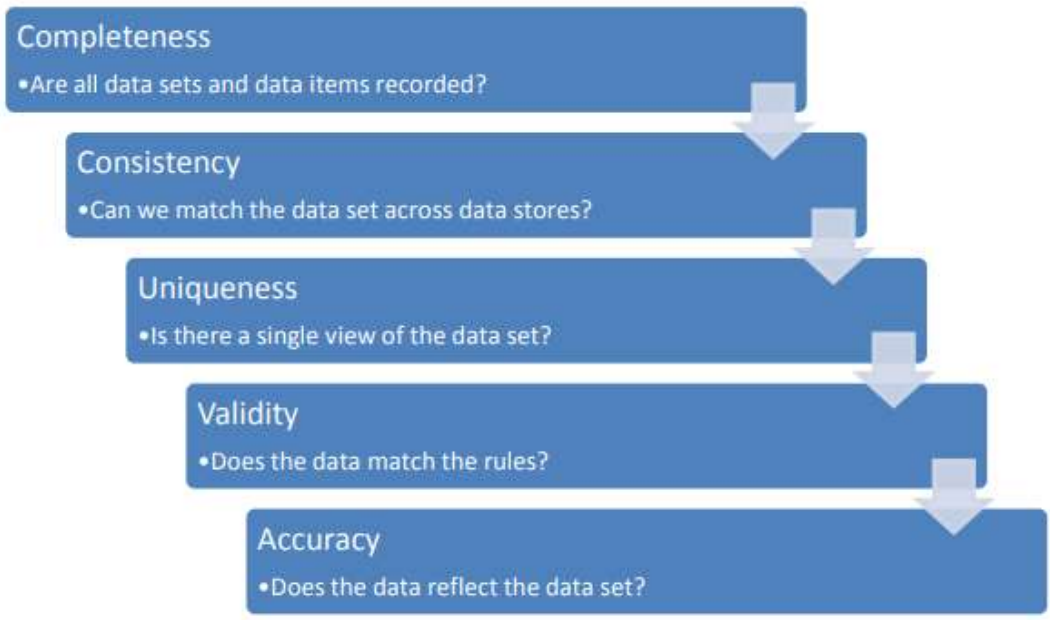


Figure 3: Example of the application of different data quality dimensions to a data set

2.6 Data Consistency and Data Inconsistency

Data Consistency simply means data is matching in all aspects. Data inconsistency exists when different and conflicting versions of the same data appear in different places. Data inconsistency creates unreliable information, because it will be difficult to determine which version of the information is correct. It is difficult to make correct and timely decisions if those decisions are based on conflicting information.

Data inconsistency is likely to occur when there is data redundancy. Data redundancy occurs when the data file/database file contains redundant, unnecessarily duplicated data. That is why one major goal of good database design is to eliminate data redundancy. (Mikkelsen, 2001) Observed that parallel use of electronic and paper-based patient records has resulted in inconsistencies between the record systems in hospital settings. Documentation is missing in both the electronic and paper-based records. When implementing electronic record systems intended to operate in parallel with paper-based systems, focus should be on securing the validity of all versions of the record. Furthermore, obtaining complete and accurate data from clinical sites to track progress presented a major

challenge in PMTCT data submitted to the district health information system (DHIS) in three districts of Kwazulu-Natal province, South Africa (Mphatswe, 2011).

On the other hand, data inconsistency can be described as the presence of information that is incompatible or inconsistent across various datasets is referred to as data inconsistency. This can occur inside a single dataset or across numerous datasets. These phenomena may occur for a variety of reasons, including the integration of data from numerous sources, the performance of updates by several users, or errors that occur during the data entry process. Finding and fixing inconsistencies in the data is absolutely necessary in order to keep the data quality high and to ensure that decisions are based on reliable information.

2.6.1 The Consequences of Inconsistent Data:

The process of decision-making: According to Redman (2001)inconsistent data might result in inaccurate analysis and decision-making. Inconsistent data can disrupt business operations and lead to inefficient processes according to Batini *et al.* (2009), inconsistent data can cause business operations to become disrupted. Trust from customers: According to Bose and Sugumaran (2003), inconsistent data causes customers to lose faith in systems that are dependent on correct information, such as customer relationship management (CRM) systems.

Inconsistent data continues to be a major problem in the field of data management, and it has a negative impact on decision-making, operational efficiency, and the trust of customers. In order to effectively address this difficulty, you will need a combination of reliable data integration processes, efficient cleaning methods, and sophisticated conflict resolution solutions. In order to keep the consistency and quality of data, new ways will be necessary as technology continues to advance and more data continues to be generated.

2.6.2 Reasons for Inconsistencies in Data

Data Integration: Inconsistencies might arise due to differences in schema, vocabulary, and representation when combining data from many sources (Fan *et al*., 2012). Concurrency Control: In situations with multiple users, simultaneous updates to the same data item can

lead to conflicts and inconsistencies (Bernstein *et al*., 1987). This problem can be avoided by using concurrency controls. Errors in Data Entry Errors in data entry can introduce inconsistencies that can spread throughout the dataset (Rahm & Do, 2000). Errors in data entry can be introduced manually.

2.6.3 Internal Inconsistency

This dimension examines the accuracy of reporting of selected indicators, by reviewing source documents to determine whether data are free of outliers (within bounds), by assessing whether specific reported values within the selected period (such as monthly) are extreme, relative to the other values reported. It also measures trends in reporting over time, to identify extreme or implausible values year-to-year. The program indicators are also compared to other indicators with which they have a predicable relationship, to determine whether the expected relationship exists between the two indicators.

2.6.4 External Inconsistency

This dimension examines the level of agreement between two sources of data measuring the same health indicator/ same data elements. The routinely collected and reported data from the health management information system (HMIS) or program-specific information system or any repository especially sources from clinical researches using data triangulation versus a periodic population-based survey, comparison with data surveys or routine data and consistency of population trends.

2.7 Factors Affecting Data Quality

One of the challenges of routine health information system (RHIS) in low- and middle-income countries revolves around medical personnel, who are faced with the dilemma of seeing patients and compiling monthly statistics. A major concern is that clinic personnel, such as Nurses, have multiple responsibilities, including clinical responsibilities, which may interfere with the time they allocate for data responsibilities and affect time they allocate for data collection. Clinic staff may value the care of patients over data collection; hence collection may be completed many days after the event has occurred, and this lag-time may impact on the quality of the statistics they produce (Edward Nicol, 2013).

Another notable concern is that at facility level, there are piles of registers and tally sheets that need to be collated, summarized and sent to the sub-district level. Training is not usually provided for clinic staff involved in data collection processes, who oftentimes, have very limited data quality checking skills and do not understand the value of the data being collected; as such data captured into the RHIS may be of low quality, thus are inaccurate and data collection methods are not complete.

In case of DHIS, the data are collected at the facility level in a paper format and captured into electronic format (Excel) at the sub-district level, which is then imported into the DHIS at the district level. Consequently, there are a number of opportunities for transcribing errors, particularly when these tasks are performed in unconducive environments.

Evaluations that have looked at the people aspect of the health information system in South Africa have only focused on the availability of human resources and not on competence or other behavioral factors. Using the Performance of Routine Information System Management (PRISM) tool that assumes relationships between technical, behavioral and organizational determinants of the routine information processes and performance, this paper highlights some behavioral factors affecting the quality of routinely collected data in South Africa. In the context of monitoring maternal and child health programmers, data were collected from 161 health information personnel in 58 health facilities and 2 district offices from 2 conveniently sampled health districts. A self-administered questionnaire was used to assess confidence and competence levels of routine health information system (RHIS) tasks, problem solving and data quality checking skills, and motivation. The findings suggest that 64% of the respondents have poor numerical skills and limited statistical and data quality checking skills. While the average confidence levels at performing RHIS tasks is 69%, only 22% actually displayed competence above 50%. Personnel appear to be reasonably motivated but there is considerable deficiency in their competency to interpret and use data. This may undermine the quality and utility of the RHIS.

2.8 Data Quality in Health Systems

Previous researchers established that public health is a data-intensive field which needs high-quality data to support public health assessment, decision-making and to assure the health of communities. Data quality assessment is important for public health. (Chen, 2014) conducted a study with the main aim of assessing the quality of data, examined the data quality assessment methods based on proposed conceptual framework which incorporates the three dimensions of data quality in the assessment methods for overall data quality, thus data, data use and data collection process.

Chen (2014) found out that the most-assessed three attributes' methods used in measuring data quality were completeness, accuracy, and timeliness. Quantitative data quality assessment primarily used descriptive surveys and data audits, while qualitative data quality assessment methods include primarily interview, documentation review and field observation. It was discovered that data-use and data-process have not been given adequate attention, although they were equally important factors which determine the quality of data, they observed limitations of the previous studies which had inconsistencies in the definition of the attributes of data quality, failure to address data users' concerns and a lack of triangulation of mixed methods for data quality assessment. The reliability and validity of the data quality assessment were rarely reported.

It was noted that in the future, data quality assessment for public health needs to consider the three dimensions of data quality, data, data use and data process and more work is needed to develop clear and consistent definitions of data quality and systematic methods and approaches for data quality assessment. The results of their review highlighted the need to develop data quality assessment methods and measuring the perceptions of end users or consumers towards data quality which will enrich understanding of data quality issues and clear conceptualization, scientific and systematic operationalization of assessment to ensure the reliability and validity of the measurement of data quality. New theories on data quality assessment for Public Health Information Systems (PHIS) may also need to be developed.

2.9 Steps in Assessing Data Quality to Measure Error Prevalence

The purpose of data quality assessment is to identify data errors and erroneous data elements and to measure the impact of various data driven processes (Maydanchik, 2008). To understand data quality gaps, the first step in any data quality management is to assess magnitude of the problem. The process involves steps indicated in **figure 4**. The framework begins by asking questions like what and why to understand data quality gaps, magnitude and what contributes to such problems.

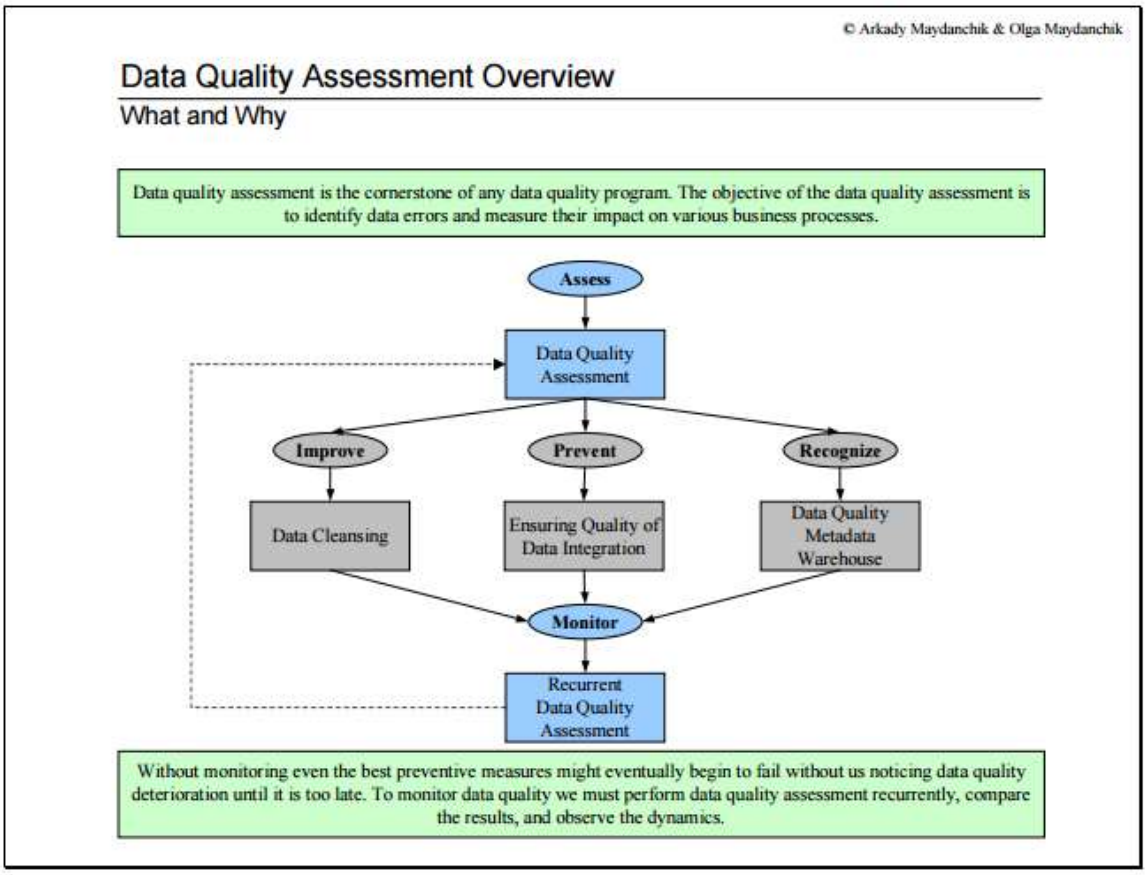


Figure 4: Steps in assessing data quality by Arkay and Olga Maydanchik

As noted, the result of data quality assessment can be used to correct existing data problems, improve data collection processes and prevent future data errors.

2.10Theoretical Framework

2.10.1Data Quality Assessment Framework

The Data Quality Assessment Framework is organized around a set of prerequisites and five dimensions of data quality assurances of integrity, methodological soundness, accuracy and reliability, serviceability, and accessibility. One of the main components in information quality assurance is an information quality measurement model design and operationalization. Quality needs to be measured meaningfully so as to establishing a causal connection between the sources of change, the problem types, the types of activities affected, and their implications. (Alverez, 2010).

The framework identifies five dimensions of quality - and for each dimension: -

- Elements (that can be used in assessing quality) - and for each element,
- Indicators (that are more concrete and detailed) - and for each indicator,
- Focal issues (that are tailored to the dataset) - and for each focal issue
- Key points (to be considered for the assessment)

2.11 Conceptual Framework

The study was guided by Data Quality Assessment Framework (DQAF) concepts. Relating to the nature of the research, the following concepts have been drawn from the two frameworks because they address the key factors under study. Understanding data quality: The study focused in understanding data quality in relation to the data collection methodology, in particular Teleform technology and how this has an impact on clinical evaluation outcomes. Guided DQAF, data was assessed in dimensions; which reviews elements, indicators, focal issues and key points on each dimension. This DQAF consists of comprehensive typologies of quality problems, related activities, and a taxonomy of quality dimensions organized in a systematic way based on sound theories and practices that are very useful and provided sound guidance in measuring data quality to the researcher.

The literature reviewed above indicates that data collection methodology has an impact on data quality.

Below are the concepts used:-

Concepts Used	Description
Consistencies	Data is matching in all aspects
Accuracy	Objects accurately represent the "real-world" values they are expected to model
Completeness	Availability of requisite information. Are data values missing, or in an unusable state?

2.12 Summary of Literature Review

Data quality is a cornerstone of rigorous research, and the methodology used for data entry significantly affects its accuracy and reliability. This literature review highlighted the impact of various data entry methodologies on data quality, identified key factors influencing data quality, and recommended best practices for ensuring accurate and reliable data collection in research settings. By adhering to these practices, researchers can enhance the integrity of their findings and contribute to the advancement of knowledge in their respective fields.

The Quality data contribute significantly in making good decisions concerning health systems. All three dimensions of data attribute namely completeness, accuracy, and timeliness need to be measured in order to achieve high quality data. The literature revealed that poor data quality has been observed previously where researchers confirmed internal and external factors negatively affected data quality.

It further explains how different data collection methodology impacted quality of data collected, including the quality of responses. To improve data quality, the literature indicates the importance of assessing data quality to determine its impact. This is very essential in clinical settings especially when there is a great amount of data collected.

CHAPTER 3

METHODOLOGY

3.1 Introduction

Chapter 3 describes the methods that were followed in conducting this study. The contents in this chapter details the research design used, population of the study area, sample size and sampling techniques, a description of data collection instruments used, as well as the techniques that were used to analyze the data. The data collection spanned 5 months to comprehensively explore the study's variables.

3.2 Research Design

The research design refers to the overall strategy that you choose to integrate the different components of the study in a coherent and logical way, thereby, ensuring you will effectively address the research problem; it constitutes the blueprint for the collection, measurement, and analysis of data (Vaus, 2001).

The design of this thesis was derived from a cross-sectional survey design in order to answer the research questions this thesis is addressing. Cross-sectional study was employed in this study because it emphasizes detailed contextual analysis of a limited number of events or conditions and their relationships. This study involved both quantitative and qualitative methods because of the use of a systematic process for obtaining quantifiable information to understand what impact data collection methodology had on data quality, what data quality gaps were observed during Option B+ evaluation, the causes and impact on outcomes. Primary data was collected through observations, a questionnaire, followed by interviews in order to understand the errors that were observed during the evaluation study.

In the study, secondary sources were reviewed in order to gain an understanding of the causes of the data gaps. This secondary data was collected through document reviews.
The information that was obtained from the administered questionnaire is presented in numerical form and analyzed through the use of statistics to describe and test relationships among variables and to what extent are the variables related.

3.3 Research Approach

The study used deductive approach where the researcher was testing theories of data quality mainly concentrating data collection methodologies and how they affect the quality of data being collected. Many studies revealed that methodology of collecting data has an impact on the data quality. There is wide agreement in the methodology community that the choice of data collection mode may affect the quality of response (Beverly & Alphonso, 2014). In addition, the method of choice may also influence respondent behavior and feelings which may also impact the quality of data. Each methodology has advantages and disadvantages which in turn has an impact on the data collected. In this study therefore respondents who were collecting data using Teleform Technology were interviewed in order to test and confirm these theories.

3.4 Research Strategy

This is applied research whose outcome was aimed at improving data quality in clinical studies. The study was exploratory research designed to achieve the objectives of this research, learning the types of errors encountered during data collection and their causes.
The study further sought new insights on the opportunities which could reduce data errors during data collection. There might be more underlying issues that affects data quality that this study might reveal. As earlier mentioned, health data is usually incomplete, inconsistent, inaccurate, lack integrity and outdated.

The study was conducted using case study as a research strategy. Data triangulation was used by targeting respondents using observations, questionnaire and reviewing documents

mainly the clinical research protocol, special questionnaires which were designed to be used for teleform and journals on PMTCT.

3.5 Study Population, Sample Size and Selection

A study population is the total collection of elements about which we wish to make some inferences (Cooper, 2003). To answer the research question, the researcher used judgmental sampling. Deliberate, critical, or judgmental sampling is the type of sampling where the researcher judges and develops his sample on the nature of the study and the understanding of his target audience. Only people who meet the research criteria and the final objective are selected.

A sample is simply a subset of the population. The population only included three categories of clinical staff who were administering the research, thus the Project Coordinators (District and Zonal), Data Processing team (Data Clerks and Data Manager) and Project Owners (Project Director). These categories of population were chosen because they were presumed to be the type of people who participated in data management and processing of data and therefore were in position to give accurate and reliable information about the study. This choice presented an opportunity for the examiner to collect highly accurate data which was economically feasible.

The

3.6 Data Collection Methods

Primary and secondary sources assisted the researcher to answer the research questions. Predominantly primary data was collected through a questionnaire while secondary data collection was conducted using interviews, observation, document/record review.

3.6.1 Measurement of Variables

Data quality dimension is described as feature of data that can be measured or assessed against defined standards in order to determine the quality of data (DAMA, 2013). Dimensions in this case are indicators helping us to measure and communicate the quality of data, as opposed to defining what the data itself means or represents. DAMA

(2013)further indicates that dimension is something (data item, record, data set or database) that can either be measured or assessed in order to understand the quality of data.

The researcher applied the data quality dimension by understanding the data quality rules of the evaluation against the data assessed. These rules were developed based upon the six data quality dimensions; the requirements for data and the impact of data not complying with these rules. The researcher adopted the following data quality assessment approach: -

1. Identify which data items need to be assessed for data quality, typically this will be data items deemed as critical to business operations and associated management reporting.
2. Assess which data quality dimensions to use and their associated weighting.
3. For each data quality dimension, define values or ranges representing good and bad quality data.
4. Apply the assessment criteria to the data items.
5. Review the results and determine if data quality is acceptable or not.
6. Where appropriate take corrective actions e.g. clean the data and improve data handling processes to prevent future recurrences.
7. Repeat the above on a periodic basis to monitor trends in Data Quality.

Using the 6-core dimension of data quality, the researcher was able to define the quality of data as presented by the Evaluation team during the fact-finding process. F**igure 5** shows the data quality dimension assessed.

Figure 5: Data Quality Dimensions

The types of data points or entities were numeric were presented in a table. A table is the simplest way of summarizing a set of observations.

3.7 Data Collection Tools

The data collection tools consisted of questionnaires targeted to respondents relevant to the study; interviews with respondents, records and document reviews, observations and data presentations. (Jenn, 2006)

3.7.1 Questionnaire

A questionnaire is a research tool used to conduct surveys. It includes specific questions with the goal to understand a topic from the respondents' point of view (Jenn, 2006). Questionnaires typically have closed-ended, open-ended, short-form, and long-form questions.

The researcher using a questionnaire in order to find answers to the research questions. The steps in getting the desired questionnaire and answers included the following: -

3.7.1.1 Questionnaires Design

Questionnaires were designed and administered by the researcher to the targeted respondents using a Kobo Toolbox. Kobo Toolbox is a free open-source tool for mobile

data collection, available to all. The researcher used this methodology because it is a web-based application enabling collection of data remotely without demanding face-to-face interviews, which is economical. Some respondents were based in different districts, the toolbox provided the opportunity to collect the required data remotely. The other advantage is that the toolbox allowed inclusion of data validation rules which helped in minimizing data errors, thereby collecting accurate data. The toolbox also provided data collection efficiency as the Data does not require transcription from paper before it was analyzed. Some analyses were applied within minutes of being collected. **Table 1** illustrates the teams that were interviewed.

Table 1: Teams Interviewed

Study Area	# Interviewed	Designation	Category	Responsibility
South, Central/North	3	Study Coordinator(Nurse)	Project Coordinators	Coordinate overall field work, scan the filled forms into Teleform
Central/Northern, South	3	Data Entry Clerk	Data Processing	Data entry and cleaning
Central (MSH Office)	1	Data Manager	Data Processing	Data cleaning, extraction and presentation
Central (MSH Office)	1	Deputy Project Director	Project Owner	Data analysis and use

3.7.1.2 Questionnaire Administration

The researcher took an automated questionnaire administration approach where links to assess the questionnaire were sent to the targeted respondents through emails and in other cases WhatsApp.

Figure 6 illustrates how respondents to the questionnaire were distributed. This to say there was 1 deputy project director, 1 data manager, 3 data entry clerks and 3 study coordinators (nurses) which participated in the study.

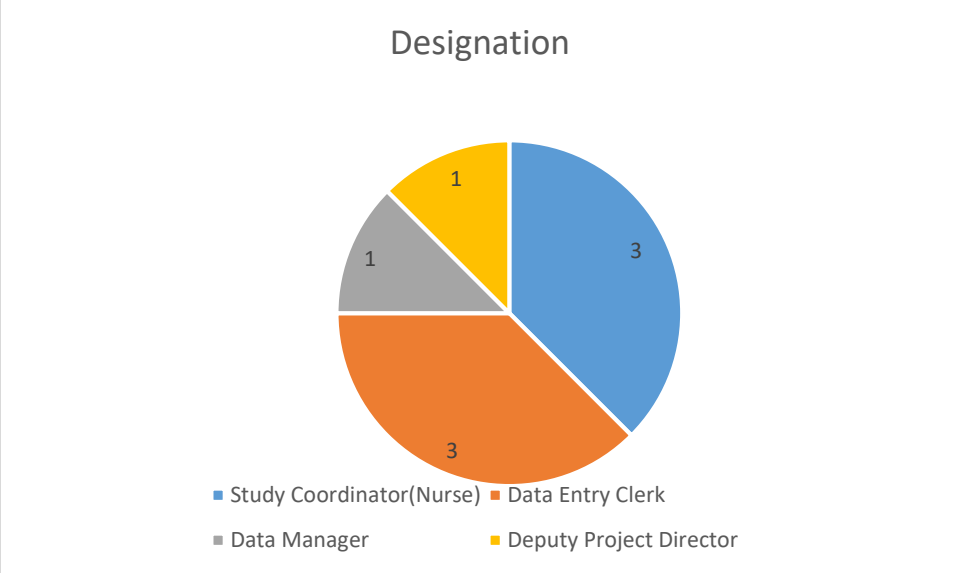


Figure 6: Questionnaire Distribution

a) The Study Coordinators

These coordinated the overall field work operations and training and quality control during study implementation to check screening procedures for infant HIV exposure, maternal and infant cards for data completeness and quality, and overall adherence to study protocol. The researcher interviewed this group in order to gain more insight on how data was captured at the source. This is the core team that handled all the data collection and transcription processes at the facilities.

b) Data Entry Clerks

Data Entry Clerks in the clinical research were responsible for reviewing data, identifying errors/gaps and data entry where necessary. They were also responsible for organizing files. The researcher interviewed this group to understand the type of errors encountered during data reviews that necessitated data cleaning and to also understand frequency of errors, causes and how the data errors impacted the results.

c) Data Manager

The Data Manager was responsible for designing the database, data cleaning protocols and entry processes including quality control and performance monitoring systems and data extraction. The researcher interviewed the Data Manager to understand the errors or any data management problems encountered and also learn on how data management specific issues were addressed.

d) Project Director

The Project Director was one of the project investigators (PI) whose core responsibility was to oversee execution of the clinical evaluation. The role entailed critically reviewing the data to make sure it responds to the objectives and also facilitation of findings and analyses which were periodically shared with MoH and all relevant stakeholders.

3.7.2 Interviews

The researcher conducted interview to the same population because the researcher wanted to deep dive and gain more insights and clarification on answers provided, derived from the questionnaire. The interviews were open-ended questions. This data was presented in accordance with the objectives of the study and helped to substantiate findings from quantitative data. Some themes and appropriate response from the interview were stated to support the quantitative findings in form of direct quotations from the respondents as noted by (Kothari, 1990).

In order to understand how data collection was conducted, the researcher also took descriptions and findings back to the study participants to check if they adequately

represented their reality by asking them whether error rates and descriptions were complete and realistic and whether they reflected the true picture on the ground. In total 24 interviews were conducted, some of which were group interviews especially in facilities. The interview technique, especially the semi-structured interview, is an essential technique for many knowledge acquisition methodologies and semi-structured interviews combines a highly structured agenda with the flexibility to ask subsequent questions (Milton, 2003).

The questions were designed in an open-ended way to give flexibility to the facility staff while conducting their routine work, whereby they could speak freely, in others cases, they could give examples of different scenarios and challenges faced that impacted data quality and verified data gaps observed during record reviews. In open ended interviews you can ask key respondents about the facts of a matter as well as their opinion about events (Yin, 2002). The interview method was used on some few respondents in order to supplement the data obtained from the questionnaire, document review and observations. The reason why the interview method was preferred for these respondents was because the researcher intended to capture in-depth information that had direct impact to their data collection activities which could not be expressed using the questionnaire.

3.7.3 Records and Document Reviews

The researcher reviewed HIV protocol documents and previous related literature to gather background information and also understanding of the study domain (PMTCT in Malawi). Further understanding of data collection methodologies especially using Teleform, its advantages and disadvantages was also obtained through the sources from other studies, reports and journal articles as well. The main focus was on data quality dimension, different data collection methodologies and how they are applied in different researches and the impact has on the outcome.

Then researcher focused on establishing data errors and specifically to deep dive into understanding data quality gaps which was recorded in excel for easy analysis.

3.7.4 Observation

Observations in clinical research helps to unveil some habits that affect data quality negatively or positively. At facility level, both best and bad practices of data collection were observed. There were a number of issues observed, from collecting the data during NEMAPP questionnaire administration while attending to a participant, recording of results, completing study and routine registers up to storing of all these tools. The researcher recorded the observation in a journal and used this information to validate the causes of data errors especially during data collection process.

3.8 Data Presentations

A table is the simplest way of summarizing a set of observations. A table has rows and columns containing data, which can be in the form of absolute numbers or percentages or both. The researcher presented tabulated results in a table indicating data quality dimension and the assessment results. Data from questionnaires was later presented in form of tables, pie charts and bar graphs for ease of interpretation.

3.9 Data analysis

Data Analysis is the process of systematically applying statistical and/or logical techniques to describe and illustrate, condense and recap, and evaluate data. According to (Shamoo Adil, 2007) various analytic procedures “provide a way of drawing inductive inferences from data and distinguishing the signal (the phenomenon of interest) from the noise (statistical fluctuations) present in the data”.

Data from the questionnaires was cleaned for consistency and easiness and later analysed using the same Kobo Toolkit. This process entailed taking the data that you collect and looking at them in the context of the questions that you need to answer.

Regarding qualitative data, the different answers from the respective respondents were transcribed, categorized into common responses and analysed using thematic content analysis.

Data from observations and document review was recorded in.

3.9.1 Descriptive Statistics

In this research, descriptive analysis was used to analyze the basic relationships between items on the questionnaire.

3.10 Ethical Considerations

CMED is the authority vested with the management of the Health Information System (HIS) environment of MOH Malawi, including approval of the release of any health data by Data Custodians. In this respect CMED is the sole authority that approves the release of data by respective data custodians. In the execution of this authority, CMED takes into account approvals by overarching Institutions like the National Health Sciences Research Committee (NHSRC) or any other recognized bodies.

The researcher approached data custodian first, in this respect, CMED to get initial consent to access the data. After getting the consent from the data custodians, the research sought consent from NHSRC.

Strict confidentiality was observed. Pseudo-names of study participants were recorded on questionnaires and interview guides. Filled questionnaires were kept under lock and key and only the researcher had access to the keys.

3.11 Limitation of the Study

The challenge was not the absence of data, but rather the constrained time frame for data acquisition and the small sample population.

3.12 Summary

The researcher used a combination of data collection methods by way of methodological triangulation. This was done this way to enable the various methods to complement one another, thereby making up for the weakness in each method. As a result, the researcher

was able to capture a more comprehensive variety of information and more discrepancies in the data collected.

In facilities interviews were mainly focusing on understanding how the whole evaluation was conducted and whether the data issues were addressed as required taking into consideration their routine work that they carry daily. This was important to understand their perception towards the study and level of commitment employed to conduct parallel data correction. Individual interviews were targeted to data handlers; Project coordinators and data entry clerks who were directly involved in day-to-day data processing reporting and resolving of data. Emphasis was on the inconsistencies of data collected by the same facilities from different sources with variation concerning the same patient i.e. date of birth. The interview also targeted the understanding of the whole transcription process that data entry clerks were involved in, thus from scanning of the evaluation questionnaires, loading the data up to cleaning the errors. This complemented the findings that emanated from the questionnaire.

CHAPTER 4

FINDINGS

4.1 Introduction

This chapter outlines the study's findings, delving into the outcomes of data analysis. The findings center around the questions posed in the questionnaire, aligning with the objectives of the study. The chapter looks at the following issues: opportunities and challenges that were associated with use of OCR technologies as a data capturing and transcription methodology, the root causes of the errors, the impact of data errors on the final outcome and strategies developed to improve data capturing using OCR. Additionally, conclusions drawn from the findings take into account insights garnered from the literature review.

4.1.1 Opportunities in Using Teleform as Data Capturing methodology

In order to address the research question on what opportunities the methodology of data capturing presented, frequency tabulation was used by the researcher to present the results from the respondents.

There were a total of 8 respondents to the questions which were administered using the questionnaire. In terms of experience in using the methodology to collect data, to summarize it all, the respondents indicated that Teleform was efficient as far as data processing was concerned and easy to use. Others indicated that it was user friendly, as presented in **Table 2**:

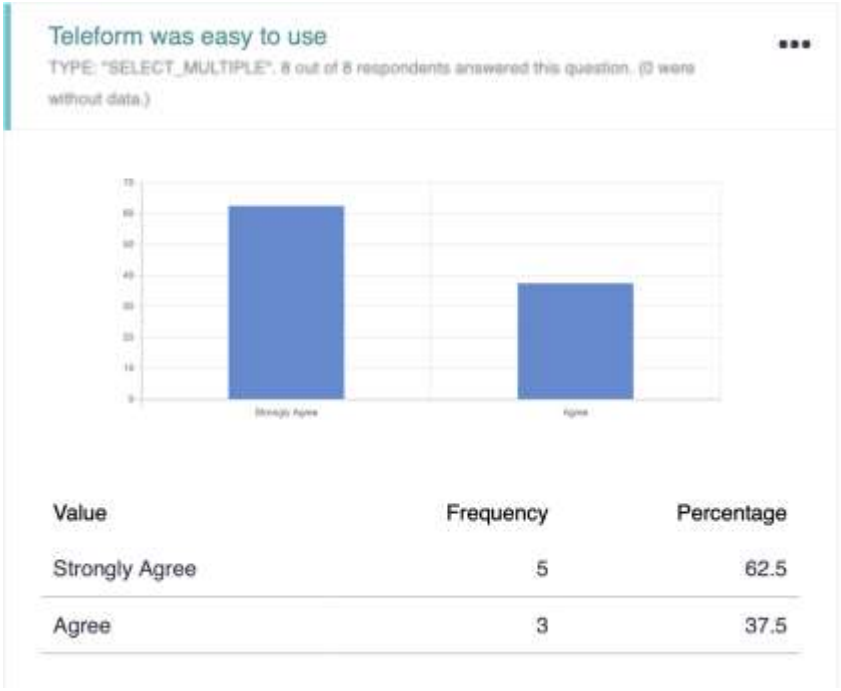
Table 2:Experience in Using Teleform

What was your experience with Teleform

TYPE: "TEXT". 8 out of 8 respondents answered this question. (0 were without data.)

Value	Frequency	Percentage
Proved to be effective and efficient	1	12.5
User friendly	1	12.5
My experience with Teleform was really great. It is one of the best and easiest database I've ever used so far. It is user friendly and able to capture data In a short period of time.	1	12.5
Very good and easy to use	1	12.5
It was a good experience	1	12.5
Easy to use and time saving.	1	12.5
It made data processing easier	1	12.5
Easy to use and efficient	1	12.5

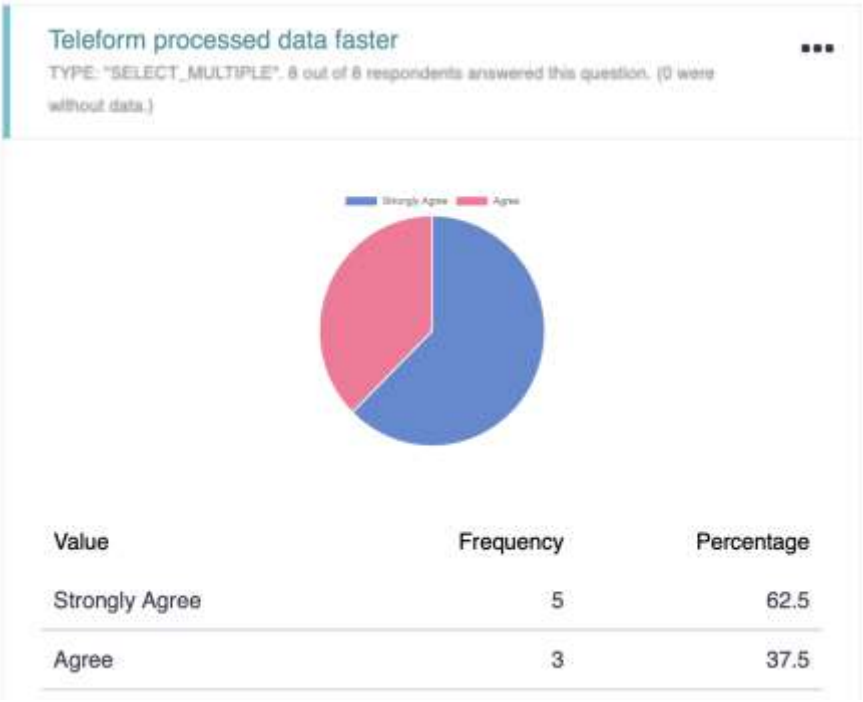
To qualify the answers on the opportunities of methodology, in terms of ease to use the methodology, 62.5% of the respondents strongly agreed that Teleform technology using OCR was easy to use while 37.5% simply agreed. **Figure 7** presents the feedback.



Value	Frequency	Percentage
Strongly Agree	5	62.5
Agree	3	37.5

Figure 7: Ease to Use Feedback

There was also a correlation in the results in terms of opportunities Teleform presented during the evaluation process in regards to its efficiency. The respondents strongly agreed to the efficiency this methodology presented in consideration to data processing. Just like the other advantage in terms of ease to use, the results varied between strongly agree and agree as presented in the F**igure 8.**

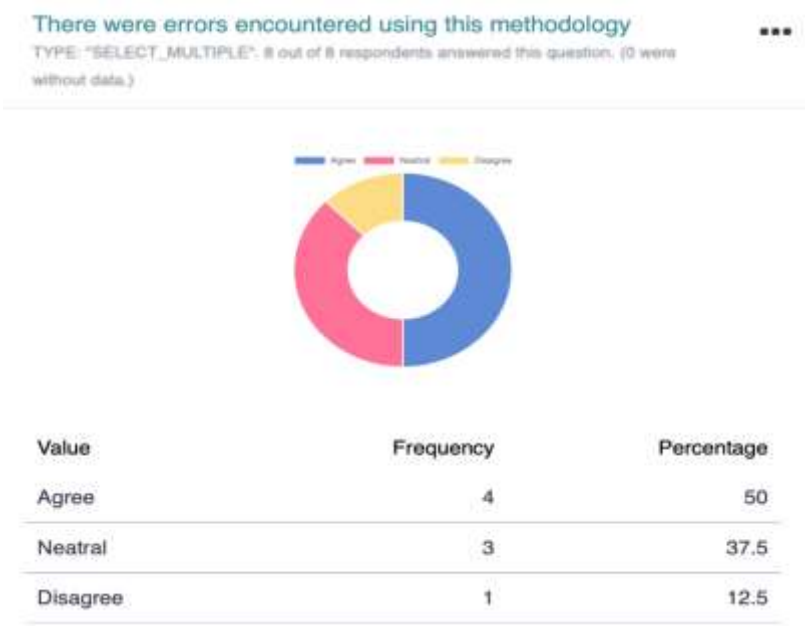


Value	Frequency	Percentage
Strongly Agree	5	62.5
Agree	3	37.5

Figure 8: Efficiency in data processing feedback

4.1.2 Challenges in Using Teleform as Data Capturing methodology

In terms of the challenges encountered using Teleform, 4 out 8 respondents agreed that they experienced errors, representing 50% (n=8), 3 remained neutral while 1 disagreed. The distribution responses were summarized in the pie chart presented **Figure 9**.

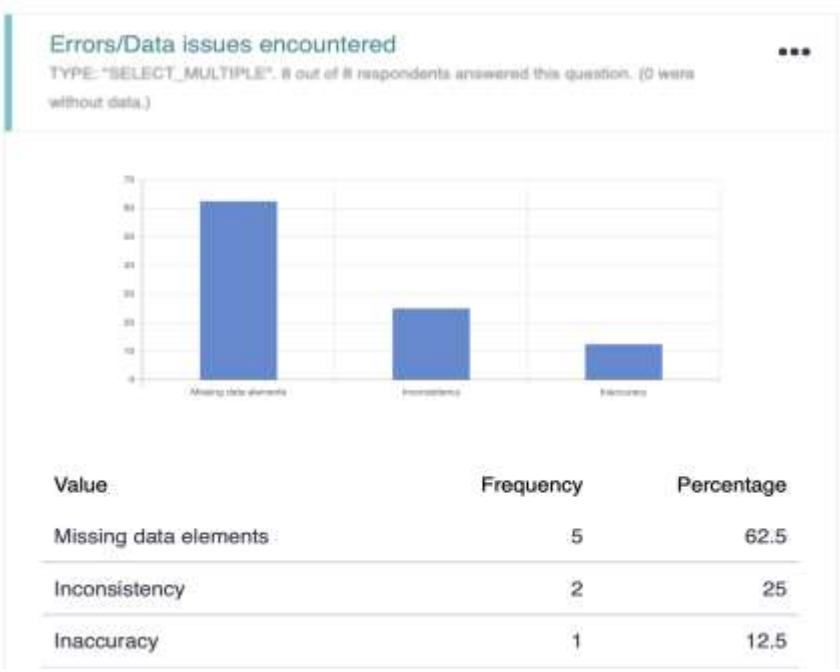


Value	Frequency	Percentage
Agree	4	50
Neatral	3	37.5
Disagree	1	12.5

Figure 9: Errors encounter feedback

4.1.3 Causes of Errors in Data Capturing

This was the study's objective 2, to establish causes of data quality gaps. The following findings emerged from the questionnaire, presented in **Figure 10**. 62.5% of the respondents indicated missing data elements as one of the causes of data errors encountered and 25% had indicated data inconsistency as another cause. Furthermore, 12.5% of the responses also indicated inaccuracy as another cause. Notably, one respondent indicated another cause of errors to be scanner failure.



Value	Frequency	Percentage
Missing data elements	5	62.5
Inconsistency	2	25
Inaccuracy	1	12.5

Figure 10: Causes of Errors

Another cause of data errors in data capturing was observed from the demanding schedules of facilitators that significantly impacted the data collection process. As observed during instances where staff had to attend to unforeseen emergencies alongside their study responsibilities. For instance, a nurse, while dealing with a participant, had to abruptly handle a clinic emergency, leading to an oversight in recording vital signs on the questionnaire. Despite having the information in hand, the nurse unconsciously washed her hands due to the urgency of the situation. This narrative exemplifies the challenges faced by facilitators in managing multiple tasks simultaneously, and a data collector's admission underscores the potential for errors and oversights when workload intensifies.

The researcher extracted response of one of the data collectors:

"When we have a lot of work, we sometimes forget to complete the questionnaires. There have been situations when we also transpose the stickers, sticking mothers' stickers on a child questionnaire and vice versa."

4.1.4 Impact of Data errors on Final Outcome of Clinical Evaluations

This was the study object 3: To determine effects of data quality gap on the final outcome. Each respondent had a different opinion on the impact the errors encountered during data

capturing had on the final outcome. Every respondent had a different answer to the question. Responding to the questionnaire, the findings are presented in the **Table 3**:

Table 3: Impact of Data Errors

What was the impact of data quality errors encountered on the study outcome

TYPE: "TEXT". 8 out of 8 respondents answered this question. (0 were without data.)

Value	Frequency	Percentage
Nothing serious just needed very good handwriting	1	12.5
Shrunk paper due to moisture could not be read by scanner	1	12.5
The errors encountered had no any big impact on the study outcome as most of the errors were taken care of. For instance, errors like DoB, visit dates etc where corrected using the backend which made the discrepancies not to have any impact	1	12.5
Not much impact as the results of the use of the teleform came out as expected	1	12.5
Extra time spent resolving queries	1	12.5
Not much errors were encountered	1	12.5
Delays as more time was spent in cleaning the data	1	12.5
Slight inaccuracy	1	12.5

During face to face interviews, the respondents also indicated that:

- This contributed to data inconsistencies when Teleform commits responses to the databases
- At times, default values (e.g. XXXX, 9999 or blanks) were inserted
- At times it reads wrong values. This is evidenced in the database.

4.1.5 Strategies to improve data quality using OCR

The respondents had provided their views on what strategies could be adopted to improve data quality. According to them, adequate training and mentorship of data collectors is vital. One respondent indicated that improving the scanner and paper versions that could function in any condition could be adopted to avoid the errors. The **table 4** illustrates respondent's contributions:

Table 4: Strategies to Improve Data Quality

What strategies could have been adopted to avoid these errors

TYPE: "TEXT". 8 out of 8 respondents answered this question. (0 were without data.)

Value	Frequency	Percentage
Adequate training of data collectors vital	1	12.5
Improved scanner and paper version that could easily function in any condition	1	12.5
Updating of the Teleform to a newer versions	1	12.5
on job mentorship during data collection by the data collectors had to be every two weeks at the beginning	1	12.5
Legible documentation at all times	1	12.5
Cross-checking all fields that they are completed properly	1	12.5
Tain users to mark the forms properly	1	12.5
Keeping data effectively	1	12.5

4.2 Summary of Findings

The study found that the methodology used had more benefits as it proved to be very efficient and easy to use. However, the findings revealed some data quality issues including inconsistencies between data transcribed through Teleform and source documents, data gaps and inaccuracies in some data elements.

CHAPTER 5

DISCUSSION AND ANALYSIS

5.1 Introduction

This chapter presents the analysis and discussion of the research based on the findings of the study. The chapter aligns the study objectives, questions and the findings and literature reviews in order to present the analysis. It draws its analysis upon Data Quality Assessment Framework (DQAF) concepts, mapping relevant data sources and identifying data gaps against the desired outcome, conducting a situation analysis through a series of Data Quality Assessment (DQA) for various required data sources and producing the DQA Reports, including recommendations for improvement. This formulated a view that assisted in understanding opportunities and strengths of the methodology used for data collection and how it impacted data quality.

This research aimed at assessing the impact of data collection methodology on data quality that was used during clinical research whose aim was to evaluate PMTCT in Malawi. The main objectives were to assess the impact of data collection methodology on data quality in the clinical evaluation domain, establish the actual data quality gaps, the causes and impact on final income in order to establish possible strategies to improve data quality.

5.2 Opportunities associated with use of OCR technologies as a data capturing and transcription methodology

As presented in the findings Chapter 4, users indicated that they experienced more advantages in using OCR than disadvantages. The adoption of Optical Character Recognition (OCR) technologies within the context of health centers and Prevention of Mother-to-Child Transmission (PMTCT) initiatives presented significant opportunities for data management and process optimization. By employing OCR, health centers can

streamline the capturing and transcription of critical patient information, medical records, and PMTCT-related documentation. One key finding in Chapter 4 indicated that OCR is more efficient. (HP, 2012) agrees that OCR has the ability to digitize enormous paper documents within a short time. It is easier to store, search, and access data.

This digitization not only accelerates the transition from paper-based to electronic records but also minimizes the risk of human error during data entry. Such accurate and efficient data management is particularly vital for PMTCT programs, where precise information exchange and timely decision-making are crucial to prevent mother-to-child transmission of diseases. Moreover, OCR's capacity to swiftly process extensive datasets equips healthcare facilities to manage the substantial amount of information generated by PMTCT initiatives. Nonetheless, it's essential to acknowledge potential challenges such as handling handwritten notes and maintaining data security in compliance with healthcare regulations. A comprehensive exploration of these opportunities, coupled with a nuanced understanding of the healthcare and PMTCT contexts, is paramount to harnessing the full potential of OCR technologies in this domain.

Another advantage is that OCR is easier to implement as it is configurable. Convenient for evaluations with tight deadlines in clinical evaluations. OCR has capability to validated extracted information to ensure all content accurate. At every stage of the capture process, OCR validates extracted information to ensure that all content is fully authenticated and accurate before it is delivered into business processes and repositories, reducing downstream processing exceptions. OCR provides the flexibility to create complex business rules and logic to implement checksum algorithms, perform cross-field validation, and verify information against external data sources using IDOL connectors. Example in the **Figure 11** show how OCR corrects the errors through Data Verifier.

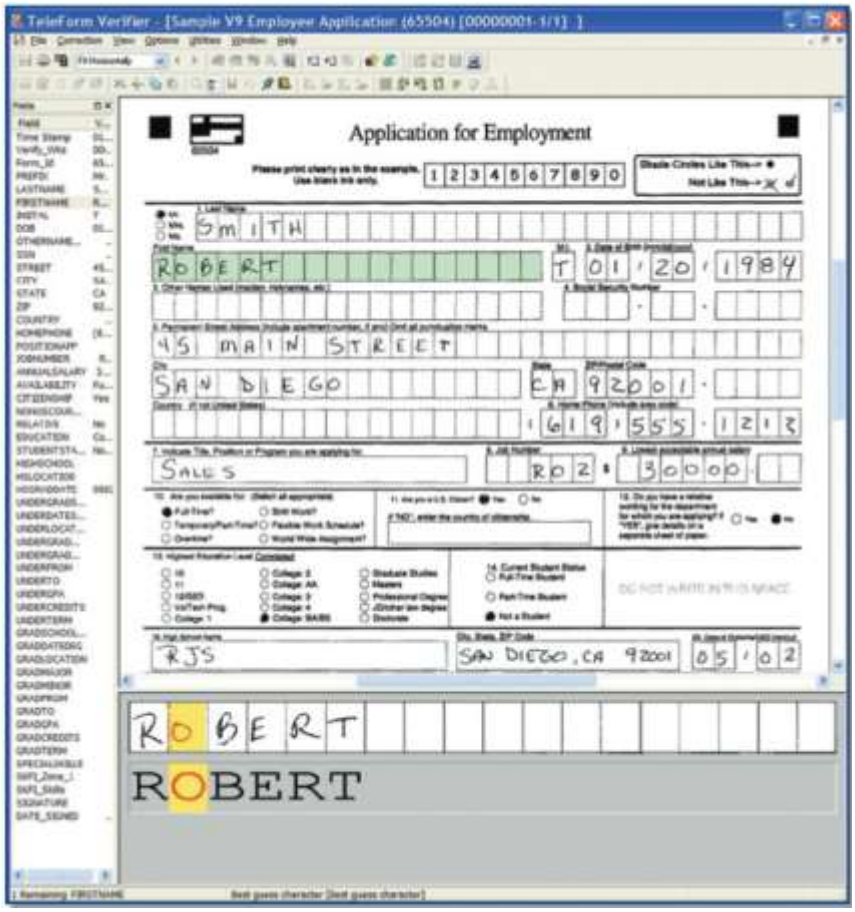


Figure 11: Teleform Data Verifier

5.3 Challenges associated with use of OCR technologies as a data capturing and transcription methodology

The findings revealed that most of the challenges encountered using this methodology are mostly the data collection processes and workflows but not necessarily the Teleform technology using OCR. However, despite its efficiencies, OCR fails to read faintly filled questionnaires, making it difficult to extract responses from the questionnaires. OCR has challenges dealing with diversity and quality of text images, different languages, scripts, fonts, sizes, orientations, layouts, and backgrounds (Nicolls, 2019).

Scanner configuration is adapted to black ink and if a different ink or pencil is used, will not read properly therefore inserting wrong data. OCR works best with good quality typed documents. Bad handwritten documents were not easily read by OCR software because, according to (Gilani, 2015), bad handwritten documents are not easily read by OCR. The

other challenge is that the Scanner configuration is adapted to black ink and if a different ink or pencil is used, OCR will not read it. It also proved to be difficult for OCR to read questionnaires that were filled wrongly and then corrected during interviews especially if they were not neatly crossed out. OCR failed to read the correct responses.

The PMTCT data processing team also experiences challenges in processing some Questionnaires due moisture. Moisture in the context of TeleForm OCR can be caused due to several issues, some of them indicated below: -

- **Image Quality**: Moisture on paper documents can0 lead to smudging, blotches, and other forms of distortion. This can result in poor image quality during scanning, making it difficult for the OCR software to accurately recognize characters.
- **Character Recognition Errors**: Moisture can cause the ink or toner on printed documents to run, making characters unclear or even illegible. OCR software relies on distinct and well-defined characters to accurately recognize text. If moisture distorts these characters, it can lead to errors in character recognition.
- **Document Integrity**: Moisture can damage paper documents, potentially causing them to wrinkle, tear, or degrade. This not only affects OCR but also poses a risk to the physical integrity of the documents themselves.
- **Data Loss**: If the moisture damage is severe, portions of the text might become completely unreadable. This can result in missing or incomplete data during OCR.
- **Misalignment**: Moisture can cause paper to expand or contract, leading to misalignment of text or other elements on the page. Misaligned text can confuse OCR software and lead to incorrect character recognition.
- **Processing Delays**: Documents that are damaged by moisture might require additional handling and preparation before scanning, slowing down the OCR process.

5.4 Causes of Data Errors

5.4.1 Study Personnel

Study personnel play a bigger role in making sure data collected is of high quality. A lot of data was collected from the health facilities and laboratories. MOH program staff, besides their normal routine work, were responsible in undertaking the evaluation activities

which involved completing research registers and forms (questionnaires) which were specifically designed for the clinical research while administering the participant in a form of interview. The process required that they carry out testing of the participant, both mother/caregiver and the child, fill out a lab form and special designed research logbooks and then stick unique identifiers in all these forms (lab form and questionnaire). These activities were done simultaneously with their routine work which required them to also complete routine registers that are used by MOH. As put across by Edward Nicol (2013), Nurses have multiple responsibilities, including clinical responsibilities, which may interfere with the time they allocate for data responsibilities and affect time they allocate for data collection. Clinic staff may value the care of patients over data collection.

For this evaluation, facility study personnel were overwhelmed with the other work therefore compromising quality of administering the evaluation study. The findings through observation, revealed that study personnel at times stuck wrong stickers e.g. child stickers on mother forms and vice-versa. In other cases, some would stick barcode stickers for another participant. This required Study Supervisors to carry out thorough quality checks before collecting the forms. The challenge was that this process was done manually, prone to errors. This problem contributed to mismatch and eventually to data inconsistency at the point of processing as indicated on figure 6 on chapter 4.

5.4.2 Erroneously Recording at Source

The completed forms/questionnaires were then processed at MSH Offices (South and Central using Teleform. TeleForm is a form processing application which performs several tasks including creating machine-readable data forms, scans images, pulls images created by other applications, allows users to review and correct data, read data from completed forms and create databases to contain the data collected and store it in the appropriate database. Teleform is designed to reduce data entry and manual processes associated with paper based questionnaires. Trapping and correcting of errors was done during transcription. Those errors were only corrected in the database but not on the forms

and facility registers. Additional data was extracted from mothers ART cards and Electronic Management Records system.

The data was processed retrospectively. The database contained erroneous records and blanks(incomplete) which could not be cleaned because in other cases it was difficult to trace back the patients to have their data corrected or completed. This was mainly caused by mothers/caregivers who deliberately gave wrong information all together including their locations in order to conceal information. Some participants moved to other locations before the end of the study without informing the facility, they could not be traced.

5.4.3 Loss of Source Documents

In other cases, evaluation questionnaires that were completed at facilities were lost before they could be collected by the study teams. In this case, the main registers could have information recorded for the participant without the questionnaire. The questionnaire had more information than the registers. In other cases, registers were torn or lost leaving them with no information to refer to when patients come for visits or when study coordinators wanted to verify some information. This caused inaccuracies as there was no reference data. The facility staff indicated that it proved to be difficult to verify the data on the registers.

5.4.4 Unclearly Filled Questionnaires

Some questionnaires were faintly filled or another ink apart from black was used. Teleform fails to read those faintly filled questionnaires, making it difficult to extract responses from the questionnaires because scanner configuration is adapted to black ink and if a different ink or pencil is used, Teleform will not read it. It was also discovered that Teleform could not read questionnaires that were not neatly crossed out during interviews. Teleform fails to read the correct response, thereby transmitting wrong data into the database. This contributed to data inconsistencies during scanning and transcription. In such situations, default values (e.g. XXXX, 9999 or blanks) were inserted.

In conclusion, the 3-core dimension of data quality as guided by DAMA (2013)were affected. The researcher was able to summarize data quality issues in this **Table 5**:

Table 5: Summary of Data Quality issues

Data Quality Dimension violated	**Root Causes**	**Data Quality Errors**
Consistencies	Loss of source documents	Source documents not equal to record in database
Accuracy	Unclearly/ Wrongly Filled Questionnaires	Wrongly transcribed data that was not able to be cleaned.
Completeness	Data gaps	Missing/loss of source documents

5.5 Impact of the data errors on the final outcome

The finding revealed the following impact on the use of the. methodology:

- During evaluation of PMTCT, evaluation questionnaires that were completed by caregivers were kept at the facility together with blood samples until evaluation teams collected them. Data processing was done retrospectively. Weekly, supervision teams would then collect the forms to the MSH Offices (Blantyre and Lilongwe) for processing. Quality checks were also conducted during supervision to see if the forms were filled correctly, all forms with queries were not picked. This impacted the evaluation in a way that it took long to reach the targeted enrollment numbers. In such cases, the facility staff were tasked to consult the participant for more details or corrections in their subsequent visits, which was not easily done as some participants could not return to the facilities because of different reasons, this contributed to data incompleteness.

- Teleform fails to read faintly filled questionnaires, making it difficult to extract responses from the questionnaires. This required the data processing team to manually capture the data, thereby delaying the data entry processing.
- The Scanner configuration is adapted to black ink and if a different ink or pencil is used, Teleform will not read it. This required that the evaluation teams consistently checked and provided required stationery to all the sites. At times, it was not possible to use the black ink especially when they had run out of them. This data was unusable as it was not read, creating more gaps in the database. This had an impact on data analysis especially if critical data elements were missing.
- It also proved to be difficult for Teleform to read questionnaires that are filled wrongly and corrected during interviews especially if they were not neatly crossed out. Teleform failed to read the correct response, thereby transmitting wrong data into the database. Wrong data always misrepresent facts, evaluators make wrong decisions.

5.6Strategies to improve Data Quality

5.6.1 Data Cleaning Strategy

- Learning from the literature and the finding of this research, most data quality issues are attributed to lack of data cleaning. Figure 13 below illustrated the data cleaning process as a conceptual recommendation:

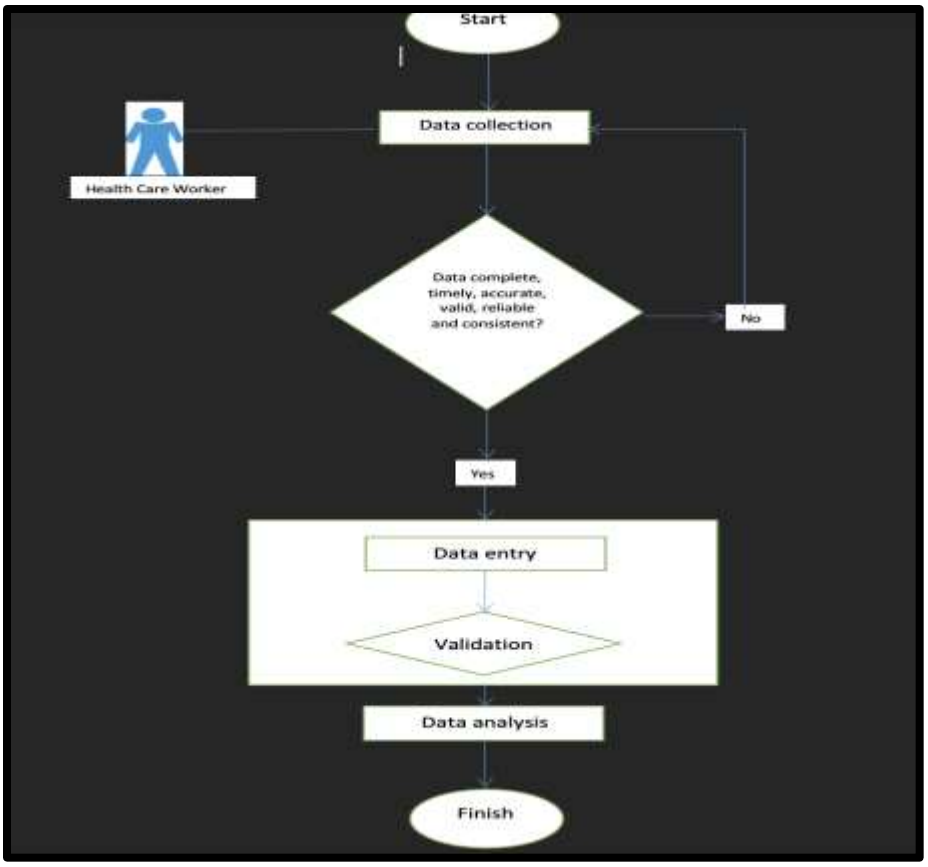


Figure 12: Data Cleaning Process

Data audits, data verification and data quality assessments are required in ascertaining high quality data.

5.6.2 Document Management

To mitigate the effects of moisture on TeleForm OCR:

- Ensure that documents are stored in a dry environment to prevent moisture buildup.
- Handle documents with care to avoid smudging or further moisture exposure.
- If documents are already damaged by moisture, consider using document restoration techniques before scanning.
- Use high-quality scanning equipment to capture the best possible images, even under less than ideal conditions.
- Depending on the severity of the moisture damage, manual data verification and correction might be necessary after OCR to ensure accuracy.

5.6.3 OCR Technology Improvements

OCR requires additional AI or advanced technology to recognize complex data types in order to improve accuracy (Biondich PG, 2002).

- It can detect and recognize characters with higher accuracy, even in cases where the text is distorted, low-quality, or obstructed by background noise.
- Improves processing time with machine learning technics
- It also Improves scanning of diverse text images – detection of patterns

It's worth noting that OCR technology has improved over the years and can handle some level of distortion caused by moisture or other factors. However, preventing moisture-related issues from the outset will still lead to better results and efficiency in the OCR process.

5.7 Conclusion

Methodology used to collect and capture data has an impact on data quality. Much as Teleform reduces manual data entry and processing and is very effective in data entry, there were numerous data errors that originate at the point of collection in facilities and were eventually transmitted into the database. Considering using Electronic Data Capture (EDC) systems that manage digital data from various channels into a centralized platform can improve quality of data and efficiency of capturing throughout clinical research.

CHAPTER 6

CONCLUSIONS AND RECOMMENDATIONS

6.1 Introduction

This chapter encompasses a concise overview of the study and its discoveries, draws comprehensive conclusions, and provides valuable recommendations. It tackles the research questions posed in the study by delving into the questions posed through the questionnaire. In forming conclusions, the findings also incorporate the insights gleaned from the literature review.

6.2 Conclusions

6.2.1 Opportunities and Challenges associated with use of OCR technologies

6.2.1.1 Opportunities

The researcher's findings on opportunities and challenges of data collection methodology indicates that there are more advantages when using OCR technologies for data capturing as opposed to its disadvantages.

One of the major benefits of using OCR is its ability to automate the process of data entry and text recognition, saving time and reducing the risk of errors. This is because of its ability to extract information from scanned documents and insert it into a database within a shortest period.The outcomes demonstrated that OCR offers high efficiency and user-friendliness. Similarly, literature above confirms that OCR technologies are influencing high productivity, superior data security and aiding in cost reductions (IBM, 2022).

Another benefit of OCR is its ability to digitize paper documents, making them easier to store, search, and access. This was particularly useful since the evaluation team had to deal with large volumes of paper documents daily, as it eliminates the need for physical storage space and makes it easier to find the information needed by just searching the electronic folders.OCR plays a pivotal role in not only extracting information but also in the critical task of validating the accuracy of the extracted content. This validation process is essential to confirm that the information obtained through OCR is reliable and error-free. By meticulously cross-referencing the extracted data with the original source, OCR helps to identify and rectify any discrepancies or inaccuracies that might have arisen during the recognition process. This validation step significantly enhances the overall reliability and integrity of the digitized data, thereby reinforcing the effectiveness of OCR technology in facilitating precise and dependable information processing. Jameson (2022) further adds that automated data capture scores over manual data capture by eliminating the risk of error. Since human effort is reduced, the accuracy of data is heightened and the cost incurred to rectify an error is almost nil.

6.2.1.2 Challenges

OCR technology encounters a multitude of challenges when it is confronted with a diverse and varied landscape of text images. These challenges encompass factors such as the quality of images, the presence of various languages, an array of scripts and fonts, distinct sizes of text, varying orientations, intricate layouts, and an assortment of backgrounds (Nicolls, 2019).

The configuration of the scanner is tailored to effectively process documents with black ink, and any deviation from this standard, such as using a different ink color or pencil, leads to inaccurate readings, resulting in the insertion of incorrect data. It's imperative to recognize that the optimal performance of OCR is closely tied to high-quality typed documents. Conversely, OCR software encounters significant challenges when it comes to deciphering poorly handwritten documents, a limitation that has been highlighted by Gilani (2015).

OCR has limitations in accurately identifying certain characters. Notably, errors tend to emerge more prominently when distinguishing between characters such as an uppercase "I" and a "1", or a "b" and an "8". It's worth noting that the application of OCR software can also give rise to transcription errors, potentially leading to the misinterpretation of content (Nicolls, 2019).

6.2.2 Causes of data errors

It was observed that some of the facility staff enrolled to undertake this study did not understand the protocol as such, they were making some errors especially sticking to barcodes which were used for identification. Much as ORC reduces manual data entry and processing, there are numerous data errors that originate at the point of collection in facilities and are eventually transmitted into the database. This is because humans are prone to making errors, and even a small data set that includes data entered manually by humans is likely to contain mistakes (Tozzi, 2022). Some fields are erroneously recorded at source during data collection. It was observed, during the study, that some questionnaires were wrongly filled at facility level due to several reasons including lack of commitment to the study, overwhelming study stuff because of many clients to attend to, lack of adherence to protocols and lack of knowledge.

Data errors are also caused due to limitations of OCR in accurately identifying certain characters. Notably, errors tend to emerge more prominently when distinguishing between characters such as an uppercase "I" and a "1", or a "b" and an "8". It's worth noting that the application of OCR software can also give rise to transcription errors, potentially leading to the misinterpretation of content (Nicolls, 2019).

Some data errors occur due to gaps and inaccuracies in the database due to missing information some of which cannot be cleanable, as it proves to be difficult to trace the patients as data processing is done retrospectively when the patient is gone and this poses errors in data capturing.

Certain data errors encountered when utilizing OCR technology can be attributed to the occurrence of data loss as a consequence of specific factors. One such factor is the destruction or alteration of papers during the scanning process, resulting from the presence of scanner moisture. In instances where papers are exposed to moisture, they can shrink or undergo distortion, ultimately leading to the loss of essential content. This deterioration not only compromises the readability of the text but also hinders the accurate recognition capabilities of OCR.

6.2.3 Impact of data errors on final outcome

The impact of data errors resulting from OCR technology reverberates beyond the initial recognition process. The need for additional time and resources to rectify and address queries stemming from these errors can lead to substantial costs incurred by organizations. One of the foremost challenges associated with employing OCR lies in the meticulous identification and subsequent elimination of errors. The process of differentiating between characters, ensuring accuracy, and rectifying misinterpretations necessitates a dedicated effort, often consuming valuable time and human resources. Jenkins (2014)underscores this challenge, highlighting the crucial task of effectively managing errors within the OCR framework.

When the database contains some erroneous records which cannot be cleaned, the data entered is considered as wrong data. Wrong data always misrepresent facts and make users/evaluators make wrong and costly decisions. Misread characters have the potential to distort the information's intended meaning, which could also then cascade into flawed decision-making processes and reporting. Consequently, meticulous proofreading and corrective measures become imperative to mitigate the adverse impact of these character recognition limitations and safeguard the integrity of the digitized content.

6.2.4 Strategies to be used to address the challenges associated with OCR and harness its opportunities in order to improve for future use

Strategies to address the challenges associated with OCR and capitalize on its potential for future enhancements involve a proactive approach that taps into advanced technologies.

One notable avenue is the integration of additional artificial intelligence (AI) components or more sophisticated technology layers. This approach is particularly relevant when dealing with complex data types that traditionally pose difficulties for OCR accuracy. By incorporating advanced AI algorithms, OCR systems can attain a higher level of adaptability and nuanced recognition capabilities, enabling them to decipher intricate characters and symbols more accurately. The synergy between OCR and AI introduces a dynamic synergy that not only improves character differentiation, such as distinguishing between an uppercase "I" and a "1", but also enables OCR to navigate intricate distinctions like those between a "b" and an "8". This endeavor can significantly elevate the accuracy and reliability of OCR outputs, paving the way for more dependable digitized data. As organizations recognize the potential of fusing OCR with advanced technology, they can enhance their capabilities to interpret and process data, ultimately ushering in a new era of refined accuracy and usability for OCR applications.

Strategic incorporation of machine learning techniques stands out as a pivotal approach. By amalgamating OCR with machine learning, a symbiotic relationship emerges that promises heightened accuracy and expanded capabilities. The adaptability of machine learning empowers OCR systems to grapple with intricate character distinctions and variations in fonts, layouts, and styles. This synergy enables OCR to evolve through the assimilation of diverse datasets and the accommodation of multifaceted textual formats, thereby elevating accuracy levels, even in the realm of handwritten content.Allocating data entry services especially data clerks to assist the nurses and clinicians during service delivery, making sure data is accurate and complete and all forms/registers are filled correctly. This will assist in filling gaps and fixing inaccuracies in the database caused by missing information.

6.3 Recommendations

To enhance data quality in clinical research, ongoing strategies such as orientation, data reviews, quality assessments, and validations should be adopted. Government staff can be motivated with non-monetary incentives like training, thus improving data collection.

Chapter 4 underscores the impact of data errors on outcomes, prompting a focus on data quality issues in the evaluation study results. However, interviews faced common issues like bias, poor recall, and inaccurate articulation, especially when facility staff were involved.

Human resources deeply affect data quality. Regular training and orientation for data collectors are vital for improved data collection quality. Incorporating data validation within collection tools can minimize errors, while data quality assessment evaluates validity, accuracy, consistency, and timeliness, crucial for effective data utilization.

Further research is warranted to gauge data quality in clinical research that involves extensive data and utilizes OCR technology, compared to alternatives like Electronic Data Entry, to enhance decision-making efficacy. Real-time data collection tools can notably reduce correction errors by embedding protocols and guidelines within validations.

Integrating AI capabilities, OCR systems can transcend the limitations posed by complex data structures, ensuring higher accuracy and precision. Advanced algorithms, fueled by machine learning and neural networks, empower OCR to navigate intricate character differentiations and variations, enhancing its adaptability to diverse fonts, layouts, and languages. This fusion of OCR and cutting-edge technology holds the promise of revolutionizing data recognition, allowing organizations to seamlessly process and interpret complex data types, thereby fostering more robust decision-making processes and insights.

Establishment of rigorous scanning protocols and environmental controls. Implementing thorough procedures to prevent moisture exposure during the scanning process is essential. This could involve maintaining controlled humidity levels in the scanning area and ensuring that papers are adequately protected from any potential sources of moisture. Additionally, adopting high-quality scanning equipment with moisture-resistant features can provide an added layer of protection. Regular maintenance and inspection of the scanning equipment, along with proper handling of documents before and during scanning, should be emphasized to prevent the occurrence of paper shrinkage and subsequent data loss. By prioritizing these precautions, organizations can significantly reduce the risk of

data errors arising from scanner-induced moisture damage and uphold the accuracy and integrity of OCR-driven data digitization processes.

REFERENCES

S, J. (2022, October 13). *Data Capture: Definition, Process,Methods, and Benefits*. Retrieved from Nanonets AI & Machine Learning: https://nanonets.com/blog/what-is-data-capture/amp/

C, T. (2022, November 14). *Understanding Data Quality: How Data Quality Problems Arise*. Retrieved from precisely: www.precisely.com

NLM, N. L. (2009). *Ensuring the Integrity of Research Data*. Retrieved from https://www.ncbi.nlm.nih.gov/books/NBK215260/

JAY, W. (2021). *WILL*. L CITY: S95.

Beverly, W., & Alphonso, O. (2014). Surveying Adolescents: The Impact of Data Collection Methodology on Response Quality. *The Electronic Journal of Business Research Methods* .

Vaus, D. (2001). *Research Design in Social Research.* London: SAGE.

Nicolls, D. (2019, June 25). *6 Glaring Limitations of OCR for Identity Verification* . Retrieved from Jumo: https://www.jumio.com/limitations-ocr-technology/

Monique, L., & al, e. (2018). Prevention of mother-to-child transmission of HIV: a cross-sectional study in Malawi. *Bulletin of the World Health Organisation*.

HP, C. (2012). *Autonomy.*

NSO. (2018). *MALAWI POPULATION AND HOUSING CENSUS REPORT* .

UNDP. (2016). *Human Development Report* .

Knoema. (2020). *World Atlas*. Retrieved from https://knoema.com/atlas/Malawi/GDP-per-capita

WHO. (2003, June 30). *Consolidated guidelines on the use of antiretroviral drugs for treating and preventing HIV infection (2013).* Retrieved from www.who.int: https://www.who.int/publications/i/item/9789241505727

MoH. (2014). *NATIONAL EVALUATION OF THE MALAWI PMTCT PROGRAM (NEMAPP).* Lilongwe: Ministry of Health.

key-advantages-ocr-based-data-entry.php. (n.d.). Retrieved from Flatworld Solutions : https://www.flatworldsolutions.com/data-management/articles/key-advantages-ocr-based-data-entry.php

Barchard, K. (2011). Preventing human error: The impact of data entry methods on data accuracy and statistical results. *Computers in Human Behavior*.

Daniel J Wilson, B. K. (1999). Accuracy of Digitization Using Automated and Manual Methods. *Oxford Academic*.

Solomon M, A. M. (2021). Data quality assessment and associated factors in the health management information system among health centers of Southern Ethiopia. *National Library of Medicine*.

Yen PY, K. M. (2017). Understanding and Visualizing Multitasking and Task Switching Activities: A Time Motion Study to Capture Nursing Workflow. *National Library of Medicine*.

Ping et al. (2009). The development and evaluation of a PDA-based method for public health surveillance data collection in developing countries. *International Journal of Medical Informatics*.

Elodia Cole, E. D. (2006). A comparative study of mobile electronic data entry systems for clinical trials data collection. *National Library of Medicine*.

Jenkins TM1, E. A. (2014). Evaluation of a Teleform-based data collection system: a multi-center obesity research case study. *PubMed*.

Sebastian-Coleman, L. (2013). *Data Quality for Ongoing Improvement: A Data Quality Assessment Framework.* Burlington, Mass: Morgan Kaufmann Publishers.

Chaulagai CN, e. t. (2005). Design and implementation of a health management information system in Malawi: issues, innovations and results. *NCBI*.

Edward Nicol, e. t. (2013). *Human Factors Affecting the Quality of Routinely Collected Data in South Africa.* IOS Press.

Redman. (2001). *DATA QUALITY, The field Guide.* Boston: Digital Press.

Chen. (2014). A Review of Data Quality Assessment Methods for Public Health Information Systems. *International Journal of Environmental Research and Public Health*.

Ministry of Finance, E. P. (2014). *2014 MILLENNIUM DEVELOPMENT GOAL REPORT FOR MALAWI* . Lilongwe: Malawi Government.

Chou, E. A. (2012). *Trends in Martenal Mortality : 1990 to 2010.* Geneva 27: World Health Organization.

Alverez, e. t. (2010, October 2). GS1 Data Quality Framework Version3.0.

Rogders. (2014, September). Theory of Change: Methodological Briefs Impact Evaluation No.2. Frorence, Italy.

Hong, C. (2014). A Review of Data Quality Assessment Methods for Public Health Information Systems. *International Journal of Environmental Research and Public Health*.

Gourlay, e. t. (2015). CHALLENGES WITH ROUTINE DATA SOURCES FOR PMTCT PROGRAMME MONITORING IN EAST AFRICA: INSIGHTS FROM TANZANIA. *Global Health Action*.

Arts, e. t. (2002). Defining and improving data quality in medical registries: a literature review, case study, and generic framework. *PubMed*.

Mphatswe, W. E. (2011, 12 05). Improving public health information: a data quality intervention in KwaZulu-Natal, South Africa. *Bulletin of the World Health Organization*.

Mikkelsen, G. (2001). *Concordance of information in parallel electronic and paper based patient record.* pubmed.

Thomas, C. R. (2001). *DATA QUALITY, The field Guide.* Boston: Digital Press.

Wenfei, F. (2012). *Data Quality: Theory and Practice.* Berlin Heidelberg: Springer-Verlag B.

Nahm, M. (2012). *Data Quality in Clinical Research* . London: Springer-Verlag.

DAMA, U. (2013, October). Data Quality Dimensions.

Sattler, K. (2009). *Data Quality Dimensions.* Springer, Boston, MA: Encyclopedia of Database Systems.

(USAID), U. S. (2015). Data Quality Metrics. *RHIS Data Quality* (pp. 7,20). North Carolina: Carolina Population Center.

Hong Chen, E. A. (2014). A Review of Data Quality Assessment Methods for Public Health Information Systems. *International Journal of Environmental Research and Public Health*.

Maydanchik, A. &. (2008). *Data Quality Assessment - Practical Skills.* TDWI.

Bari Fijo, E. a. (2010). *How to Write a Good Dissertation(1st Edition).* Kampala: New Vision printing and Publishing Company LTD.

Cooper, D. a. (2003). *Business Research Methods. 8th Edition.* Boston: McGraw-Hill Irwin.

Creswell, J. W. (2008). *Educational Research Planning, Conducting and Evaluating Quantitative and Qualitative Research.* International Pearson Merril Prentice Hall.

Milton, N. R. (2003). *Personal Knowledge Techniques.* Nottingham.: University of Nottingham.

Yin, R. K. (2002). *Case Study Research: Design and Methods Thousand Oaks.* Thousand Oaks: SAGE Publications.

Patton, M. Q. (1990). *Qualitative Evaluation and Research Methods.* Beverly Hills: Sage Publications, Inc.

Malawi Statistics. (2013, December 27). Retrieved from unicef.org : www.unicef.org/infobycountry/malawi_statistics.html

MOH, M. (2008). *HIV and Syphilis Sero-Survey and National HIV Prevalence and AIDS Estimates Report 2007.* Lilongwe: NAC.

NSO. (2010). *Malawi Demographic and Health Survey 2010.* Lilongwe: NSO.

avert.org. (2018, December 10). *Global information and education on HIV and AIDS*. Retrieved from HIV AND AIDS IN MALAWI: https://www.avert.org/professionals/hiv-around-world/sub-saharan-africa/malawi

Arrive, e. a. (2007). Prevalence of resistance to nevirapine in mothers and children after single dose exposure to prevent vertical transmission of HIV-1: a meta-analysis. *PubMed*.

WHO. (2006). *ANTIRETROVIRAL DRUGS FOR ANTIRETROVIRAL DRUGS FOR TREATING PREGNANT WOMEN AND PREVENTING HIV INFECTION IN INFANTS.* Switzerland: World Health Organization.

Gliklich, e. a. (2014). *Registries for Evaluating Patient Outcomes: A User's Guide [Internet]. 3rd edition.* Bethesda: Agency for Healthcare Research and Quality (US).

Saveli I. Goldberg, e. a. (2008). Analysis of Data Errors in Clinical Research Databases. *AMIA Annu Symp Proc*.

MOH. (2016). *MALAWI POPULATION-BASED HIV IMPACT ASSESSMENT.* Lilongwe: Malawi Government.

Hamzah, A. A. (2018). Data Capturing: Methods, Issues and Concern. *International Journal of Academic Research in Business and Social Sciences*, 8(9), 617– 629.

MOH. (2018). *HIV and Syphilis Sero-Survey and National HIV Prevalence and AIDS Estimates Report 2007*. Lilongwe: NAC.

Beverly, W., & Alphonso, O. (2014). The Impact of Data Collection Methodology on Response Quality. *The Academic Conferences & Publishing International Ltd (ACPIL)*.

Syed, M. A., & al, e. (2016). Digitizing Data Collection and Impacting Data Management Processes in the Tuberculosis Control Program of Pakistan.

Church, A. H. (2006). ProQuest. *ProQ*.

D, N. (2019, june 25). *6 Glaring Limitations of OCR for Identity Verification*. Retrieved from Jumio: https://www.jumio.com/limitations-ocr-technology/

Arsalan. (n.d.). *Benefits Of OCR-Based Data Entry*. Retrieved from information transformation systems : https://it-s.com/benefits-of-ocr-based-data-entry-2/

Flatworld. (2023). *9 key advantages of ORC-based data entry*. Retrieved from flatworld solutions: https://www.flatworldsolutions.com

IBM, C. E. (2022, january 5). *What Is Optical Character Recognition (OCR)?* Retrieved from IBM: https://www.ibm.com

Gilani, N. (n.d.). *Advantages & Disadvantages of Magnetic Ink Character Recognition*. Retrieved from techwalla: https://www.techwalla.com/

Caeiro, G. (n.d.). *learn 5 reasons why ocr software is expensive*. Retrieved from docdigitizer: https://www.docdigitizer.com

(n.d.).

Jameson, S. (2022, October 13). *Data Capture: Definition, Process,Methods, and Benefits*. Retrieved from Nanonets AI & Machine Learning: https://nanonets.com/blog/what-is-data-capture/amp/

Tozzi, C. (2022, November 14). Understanding Data Quality: How Data Quality Problems Arise. Retrieved from precisely: www.precisely.com

Bibliography

NLM, N. L. (2009). *Ensuring the Integrity of Research Data*. Retrieved from https://www.ncbi.nlm.nih.gov/books/NBK215260/

JAY, W. (2021). *WILL*. L CITY: S95.

Vaus, D. (2001). *Research Design in Social Research.* London: SAGE.

Nicolls, D. (2019, June 25). *6 Glaring Limitations of OCR for Identity Verification* . Retrieved from Jumo: https://www.jumio.com/limitations-ocr-technology/

Monique, L., & al, e. (2018). Prevention of mother-to-child transmission of HIV: a cross-sectional study in Malawi. *Bulletin of the World Health Organisation*.

HP, C. (2012). *Autonomy*.

NSO. (2018). *MALAWI POPULATION AND HOUSING CENSUS REPORT* .

UNDP. (2016). *Human Development Report* .

Knoema. (2020). *World Atlas*. Retrieved from https://knoema.com/atlas/Malawi/GDP-per-capita

WHO. (2003, June 30). *Consolidated guidelines on the use of antiretroviral drugs for treating and preventing HIV infection (2013).* Retrieved from www.who.int: https://www.who.int/publications/i/item/9789241505727

MoH. (2014). *NATIONAL EVALUATION OF THE MALAWI PMTCT PROGRAM (NEMAPP).* Lilongwe: Ministry of Health.

key-advantages-ocr-based-data-entry.php. (n.d.). Retrieved from Flatworld Solutions : https://www.flatworldsolutions.com/data-management/articles/key-advantages-ocr-based-data-entry.php

Barchard, K. (2011). Preventing human error: The impact of data entry methods on data accuracy and statistical results. *Computers in Human Behavior*.

Daniel J Wilson, B. K. (1999). Accuracy of Digitization Using Automated and Manual Methods. *Oxford Academic*.

Solomon M, A. M. (2021). Data quality assessment and associated factors in the health management information system among health centers of Southern Ethiopia. *National Libriry of Medicine*.

Yen PY, K. M. (2017). Understanding and Visualizing Multitasking and Task Switching Activities: A Time Motion Study to Capture Nursing Workflow. *National Library of Medicine*.

Ping et al. (2009). The development and evaluation of a PDA-based method for public health surveillance data collection in developing countries. *International Journal of Medical Informatics*.

Elodia Cole, E. D. (2006). A comparative study of mobile electronic data entry systems for clinical trials data collection. *National Library of Medicine*.

Jenkins TM1, E. A. (2014). Evaluation of a Teleform-based data collection system: a multi-center obesity research case study. *PubMed*.

Sebastian-Coleman, L. (2013). *Data Quality for Ongoing Improvement: A Data Quality Assessment Framework.* Burlington, Mass: Morgan Kaufmann Publishers.

Chaulagai CN, e. t. (2005). Design and implementation of a health management information system in Malawi: issues, innovations and results. *NCBI*.

Edward Nicol, e. t. (2013). *Human Factors Affecting the Quality of Routinely Collected Data in South Africa.* IOS Press.

Redman. (2001). *DATA QUALITY, The field Guide.* Boston: Digital Press.

Chen. (2014). A Review of Data Quality Assessment Methods for Public Health Information Systems. *International Journal of Environmental Research and Public Health*.

Ministry of Finance, E. P. (2014). *2014 MILLENNIUM DEVELOPMENT GOAL REPORT FOR MALAWI* . Lilongwe: Malawi Government.

Chou, E. A. (2012). *Trends in Martenal Mortality : 1990 to 2010.* Geneva 27: World Health Organization.

Alverez, e. t. (2010, October 2). GS1 Data Quality Framework Version3.0.

Rogders. (2014, September). Theory of Change: Methodological Briefs Impact Evaluation No.2. Frorence, Italy.

Hong, C. (2014). A Review of Data Quality Assessment Methods for Public Health Information Systems. *International Journal of Environmental Research and Public Health*.

Gourlay, e. t. (2015). CHALLENGES WITH ROUTINE DATA SOURCES FOR PMTCT PROGRAMME MONITORING IN EAST AFRICA: INSIGHTS FROM TANZANIA. *Global Health Action*.

Arts, e. t. (2002). Defining and improving data quality in medical registries: a literature review, case study, and generic framework. *PubMed*.

Mphatswe, W. E. (2011, 12 05). Improving public health information: a data quality intervention in KwaZulu-Natal, South Africa. *Bulletin of the World Health Organization*.

Mikkelsen, G. (2001). *Concordance of information in parallel electronic and paper based patient record.* pubmed.

Thomas, C. R. (2001). *DATA QUALITY, The field Guide.* Boston: Digital Press.

Wenfei, F. (2012). *Data Quality: Theory and Practice.* Berlin Heidelberg: Springer-Verlag B.

Nahm, M. (2012). *Data Quality in Clinical Research* . London: Springer-Verlag.

DAMA, U. (2013, October). Data Quality Dimensions.

Sattler, K. (2009). *Data Quality Dimensions.* Springer, Boston, MA: Encyclopedia of Database Systems.

(USAID), U. S. (2015). Data Quality Metrics. *RHIS Data Quality* (pp. 7,20). North Carolina: Carolina Population Center.

Hong Chen, E. A. (2014). A Review of Data Quality Assessment Methods for Public Health Information Systems. *International Journal of Environmental Research and Public Health*.

Maydanchik, A. &. (2008). *Data Quality Assessment - Practical Skills.* TDWI.

Bari Fijo, E. a. (2010). *How to Write a Good Dissertation(1st Edition).* Kampala: New Vision printing and Publishing Company LTD.

Cooper, D. a. (2003). *Business Research Methods. 8th Edition.* Boston: McGraw-Hill Irwin.

Creswell, J. W. (2008). *Educational Research Planning, Conducting and Evaluating Quantitative and Qualitative Research.* International Pearson Merril Prentice Hall.

Milton, N. R. (2003). *Personal Knowledge Techniques.* Nottingham.: University of Nottingham.

Yin, R. K. (2002). *Case Study Research: Design and Methods Thousand Oaks.* Thousand Oaks: SAGE Publications.

Patton, M. Q. (1990). *Qualitative Evaluation and Research Methods.* Beverly Hills: Sage Publications, Inc.

Malawi Statistics. (2013, December 27). Retrieved from unicef.org : www.unicef.org/infobycountry/malawi_statistics.html

MOH, M. (2008). *HIV and Syphilis Sero-Survey and National HIV Prevalence and AIDS Estimates Report 2007.* Lilongwe: NAC.

NSO. (2010). *Malawi Demographic and Health Survey 2010.* Lilongwe: NSO.

avert.org. (2018, December 10). *Global information and education on HIV and AIDS*. Retrieved from HIV AND AIDS IN MALAWI: https://www.avert.org/professionals/hiv-around-world/sub-saharan-africa/malawi

Arrive, e. a. (2007). Prevalence of resistance to nevirapine in mothers and children after single dose exposure to prevent vertical transmission of HIV-1: a meta-analysis. *PubMed*.

WHO. (2006). *ANTIRETROVIRAL DRUGS FOR ANTIRETROVIRAL DRUGS FOR TREATING PREGNANT WOMEN AND PREVENTING HIV INFECTION IN INFANTS.* Switzerland: World Health Organization.

Gliklich, e. a. (2014). *Registries for Evaluating Patient Outcomes: A User's Guide [Internet]. 3rd edition.* Bethesda: Agency for Healthcare Research and Quality (US).

Saveli I. Goldberg, e. a. (2008). Analysis of Data Errors in Clinical Research Databases. *AMIA Annu Symp Proc*.

MOH. (2016). *MALAWI POPULATION-BASED HIV IMPACT ASSESSMENT*. Lilongwe: Malawi Government.

Hamzah, A. A. (2018). Data Capturing: Methods, Issues and Concern. *International Journal of Academic Research in Business and Social Sciences*, 8(9), 617– 629.

MOH. (2018). *HIV and Syphilis Sero-Survey and National HIV Prevalence and AIDS Estimates Report 2007.* Lilongwe: NAC.

Beverly, W., & Alphonso, O. (2014). The Impact of Data Collection Methodology on Response Quality. *The Academic Conferences & Publishing International Ltd (ACPIL).*

Syed, M. A., & al, e. (2016). Digitizing Data Collection and Impacting Data Management Processes in the Tuberculosis Control Program of Pakistan.

Church, A. H. (2006). ProQuest. *ProQ*.

D, N. (2019, june 25). *6 Glaring Limitations of OCR for Identity Verification*. Retrieved from Jumio: https://www.jumio.com/limitations-ocr-technology/

Flatworld. (2023). *9 key advantages of ORC-based data entry*. Retrieved from flatworld solutions: https://www.flatworldsolutions.com

IBM, C. E. (2022, january 5). *What Is Optical Character Recognition (OCR)?* Retrieved from IBM: https://www.ibm.com

Jameson, S. (2022, October 13). *Data Capture: Definition, Process,Methods, and Benefits*. Retrieved from Nanonets AI & Machine Learning: https://nanonets.com/blog/what-is-data-capture/amp/

Arsalan. (2021, april 7). *Benefits Of OCR-Based Data Entry*. Retrieved from information transformation systems: https://it-s.com/benefits-of-ocr-based-data-entry-2/

Gilani, N. (2015). *Advantages & Disadvantages of Magnetic Ink Character Recognition*. Retrieved from techwalla: https://www.techwalla.com/articles/advantages-disadvantages-of-magnetic-ink-character-recognition

Caeiro, G. (2018). *learn 5 reasons why ocr software is expensive*. Retrieved from docdigitizer: https://www.docdigitizer.com/blog/why-the-ocr-software-is-expensive/

Kothari, C. (1990). *Research methodology*. Jaipur: New Age International Publisher.

Shamoo Adil, R. D. (2007). Responsible Conduct of Research. *Journal of biomedical optics*.

Tozzi, C. (2022, November 14). *Understanding Data Quality: How Data Quality Problems Arise*. Retrieved from precisely: www.precisely.com

Biondich PG, O. J. (2002). A modern optical character recognition system in a real world clinical setting: some accuracy and feasibility observations. *PubMed*.

Jenn, N. (2006). Designing A Questionnaire. *PMID*.

(n.d.).

APPENDICIES

Appendix A. The Questionnaire

Assessing Data Quality On Teleform Use

Full Name

_____

Position Held

What was your experience with Teleform

_____

Teleform was easy to use

- [ ] Strongly Agree
- [ ] Agree
- [ ] Neatral
- [ ] Disagree
- [ ] Strongly Disagree

Teleform processed data faster

- [ ] Strongly Agree
- [ ] Agree
- [ ] Neatral
- [ ] Disagree
- [ ] Strongly Disagree

There were errors encountered using this methodology

- [ ] Strongly Agree
- [ ] Agree
- [ ] Neatral
- [ ] Disagree
- [ ] Strongly Disagree

Errors/Data issues encountered

- [ ] Miss`ng data elements
- [ ] Inaccuracy
- [ ] Inconsistency