

**USE OF DATA QUALITY ASSESSMENT TO EVALUATE THE LEVEL OF HIV
DATA COMPLETENESS IN MINISTRY OF HEALTH FACILITIES:
A CASE STUDY OF EASTERN PROVINCE-KENYA**

KIRIMI PETER MUTUGI

**A Dissertation Submitted to the Department of Horticulture in partial fulfillment of the
requirement for the Master of Science Degree in Research Methods at Jomo Kenyatta
University of Agriculture and Technology**

2013

DECLARATION

I hereby declare that this dissertation is my original work and has not been submitted for the award of a degree in any other university.

Signed_____ Date_____

KIRIMI PETER MUTUGI

Registration No.AG332/1822/2010

This dissertation has been submitted with our approval as university supervisors:

Signed_____ Date_____

Dr. EDWARD MAMATI

Jomo Kenyatta University of Agriculture and Technology

Signed_____ Date_____

Dr. SANDRA MUDHUNE

International Center for AIDS Care and Treatment Kenya

ACKNOWLEDGEMENTS

This research would not have been possible without the support of many people. I would like to thank my supervisors Dr. Edward Mamati and Dr. Sandra Mudhune for their guidance and constructive criticism. I can't forget to thank Jomo Kenyatta University of Agriculture and Technology (JKUAT) Research Methods students (second cohort), ICAP Monitoring, Evaluation and Research team and Kenyatta National Hospital and University of Nairobi Ethics and Research Committee (KNH/UON ERC) for their support, my family for their constant inspiration and guidance that kept me focused and motivated. Finally, my greatest regards goes to the Almighty God for bestowing upon me the courage to face the complexities of life and complete this dissertation successfully.

ABSTRACT

Human Immunodeficiency Virus (HIV) and Acquired Immune Deficiency Syndrome data feedback and decision making is enhanced by data quality assessments which are done using well designed tools and methods. HIV data quality is useful in HIV prevention, care and treatment. The main objective of the study was to evaluate the quality of data collected in Ministry of Health facilities offering HIV prevention, care and treatment services in Eastern Province of Kenya. Assessment of factors that contribute to HIV data incompleteness in Ministry of health facilities was done by interviewing 60 health care providers and patients records reviewed in 42 health facilities to determine whether any difference exists in level of HIV data completeness in Eastern Province-Kenya. Data on perceptions behind missing HIV data entries in patients' records and competence in HIV data quality assessment for completeness was obtained by administering a semi-structured questionnaire to health care providers. Logistic regression model was used to create a relationship between training of health care providers in health management information systems (HMIS), experience in filling patient records, gender and respective competence on how to conduct HIV data quality assessment for completeness. Second, patient records (Blue cards) were evaluated for complete entries and the level of data completeness was determined. This was done by evaluating the patients' records for complete entries in the fields of weight, new opportunistic infections (OI), pregnancy status and Tuberculosis (TB) screening. The level of HIV data completeness varied in different HIV data elements. Weight had the least difference in level of data completeness at 6.89% while screening of new opportunistic infection and pregnancy status for patients had the biggest difference at 10.86% and 10.77% respectively. It was evident that training health care providers in health management information systems ($P=0.007$) and having had an experience in the comprehensive

care clinic ($P < 0.000$) contributed significantly to competence on how to conduct a HIV DQA for completeness. It was noted that many HIV data collection tools for documentation 26 (43%) and few record officers 24 (40%), were perceived as the major causes of data incompleteness in patient records. Documentation of patients' weight, pregnancy status, screening of new opportunistic infections and tuberculosis in HIV records was incomplete. Many HIV data collection records for documentation and few record officers assigned to work in the comprehensive care clinics were identified as the major causes of missing entries in HIV records. Training health care providers in health management and information systems and having had worked in the HIV clinic for at least ten months contributed positively in competence of conducting a structured HIV DQA for completeness.

ABBREVIATIONS AND ACRONYMS

AIDS-Acquired Immune Deficiency Syndrome

ANSI- American National Standards Institute

ART-Antiretroviral Therapy

CDC-Centre for Disease Control and Prevention

DHRIO- District Health Records and Information Officer

DQA-Data Quality Assessment

DQO- Data Quality Objectives

EDI-Electronic Data Interchange

ERC- Ethics in Research Committee

GDP-Gross Domestic Product

GOK-Government of Kenya

HIS-Health Information Systems

HIV-Human Immunodeficiency Virus

ICAP-International Centre for AIDS Care and Treatment Programmes

KDHS-Kenya Demographic and Health Survey

KHPF-Kenya Health Policy Framework

KNBS- Kenya National Bureau of Statistics

MRC-Ministerial Reform Committee

NGO-Non Governmental Organization

OPD- Out Patient Department

PHRIO-Provincial Health Records and Information Officer

SOP- Standard Operating Procedures

TB-Tuberculosis

UNAIDs-United Nations program on Acquired Immune Deficiency syndrome

UON- University of Nairobi

VCT-Voluntary Counseling and Testing

DEFINITION OF TERMS

Patient records review:	A process of checking through HIV primary and secondary data collection tools to cross check the quality of care given to the patients either through accurate and complete documentation or timely services
Data Clarity:	The degree of clarity with which the data, assumptions, methods, quality assurance, sponsoring organizations, and analysis employed to generate the information is documented.
Data Completeness:	The extent to which the expected attributes of data are provided.
Data Quality Assessment:	An audit of data against set standards to assess its quality in terms of timeliness, accuracy, completeness and relevance.
Evaluation:	Episodic or periodic assessment of the relevance, effectiveness and efficiency of the project interventions and its impact in relation to stated objectives.
Monitoring:	The process of routinely collecting, storing, analyzing and reporting programme or projects information used to make decisions for project management
Mother to child transmission:	Transmission of HIV from an HIV-positive woman during pregnancy, delivery or breastfeeding to her child.
Opportunistic infections:	These are infections that take advantage of the weak immune system to cause disease.
Patient blue card:	A primary data collection tool used by a health care provider to record all data elements regarded as important in management of HIV positive patients in every clinic day visit. i.e. Evaluation of Weight, pregnancy

new opportunistic infection and Tuberculosis. It's used mostly for HIV positive patient's routine data collection.

Variability:

The extent to which the variability (quantitative and qualitative) in the information or in the procedures, measures, methods, or models are evaluated and characterized.

TABLE OF CONTENTS

DECLARATION	ii
ACKNOWLEDGEMENTS	iii
ABSTRACT.....	iv
ABBREVIATIONS AND ACRONYMS	vi
DEFINITION OF TERMS	viii
TABLE OF CONTENTS.....	x
LIST OF TABLES	xii
LIST OF FIGURES	xiii
CHAPTER ONE: INTRODUCTION	1
1.1 Background of the problem.....	1
1.1.1 Impact of HIV and AIDs to the economy	2
1.1.2 Importance of data quality in HIV Programmes	3
1.1.3 Data in HIV and AIDS Programmes.....	3
1.2 Statement of the problem	5
1.3 Justification of the study	6
1.4 Objectives of the study.....	6
1.4.1 General objective	6
1.4.2 Specific objectives	6
1.5 Hypothesis.....	7
CHAPTER TWO: LITERATURE REVIEW	8
2.1 Data Quality Dimensions	8
2.2 Steps in Data Quality Assessment	13
2.5 History of HIV Data in Kenya	18
2.6 Factors Affecting HIV Data Quality in Kenya	19
CHAPTER THREE: RESEARCH METHODOLOGY.....	24
3.1 Study design.....	24
3.2 Study population	24
3.2.1 Health Facilities.....	24

3.2.2 Health care providers.....	24
3.3 Exclusion criteria	25
3.3.1 Health Facilities.....	25
3.3.2 Health care providers.....	25
3.4 Sampling design.....	25
3.4.1 Sampling frame	25
3.4.2 Sampling technique.....	26
3.4.3 Sample size.....	27
3.5 Research procedures	27
3.6 Data collection methods.....	28
3.7 Data Analysis Methods	28
CHAPTER FOUR: RESULTS	30
4.1 Evaluating the level of HIV data completeness	30
4.2 Perceptions regarding HIV data incompleteness	31
4.3 Experience in handling HIV data collection tools	33
4.4 Logistic regression analysis and modeling of competence on how to conduct a HIV DQA for completeness by health care providers.....	34
CHAPTER FIVE: DISCUSSION	37
CHAPTER SIX: CONCLUSION	40
CHAPTER 7: RECOMMENDATIONS.....	41
REFERENCES	42
APPENDICES	47
Appendix 1: Consent form for Health care providers.....	47
Appendix 2: KNH/UON ERC Approval Letter.....	49
Appendix 3: Questionnaire on HIV DQA for completeness competence and perceptions behind missing HIV data entries.	51
Appendix 4: Chi square and logistic model syntax	53
Appendix 5: A participant’s Master list template for data entry.....	54
Appendix 6: Ministry of Health facilities where HIV data completeness was evaluated.....	55

LIST OF TABLES

Table 1: Pooled HIV data elements completeness for different Ministry of Health facilities	30
Table 2: Calculated Chi square results (P-value) for Gender, experience and training on HMIS.....	34
Table 3: Logistic regression predictor variables and their respective coefficients.....	35

LIST OF FIGURES

Figure 1: Proportion of missing HIV data entries in patients blue card	31
Figure 2: Perceptions regarding HIV data incompleteness in health facilities.....	32
Figure 3: Duration worked in months by health care providers in the comprehensive care clinic.....	33

CHAPTER ONE: INTRODUCTION

1.1 Background of the problem

Accurate, timely and accessible health care data play a vital role in the planning, development and maintenance of health services. Quality improvement and the timely dissemination of quality data are essential if the health care authorities wish to maintain health care at an optimal level. In the recent past, data completeness has become an important issue not only because of its importance in promoting high standards of patients care but also because of its impact on government budgets for maintenance of health services. (WHO, 2003)

Studies have shown that there are several consequences of applying poor quality data in an organization. For instance, according to Lee and Strong (2004) in their study, poor quality data can have far reaching social and economic repercussions to the organization such as misguided decision-making, reduced efficiency and increased operational costs. The most salient consequence is an increase in operational costs because significant resources including time and funds are spent trying to detect and correct errors. According to Häkkinen and Hilmola (2008) the economic implications of using low quality data may be more evident in operational disruptions. Studies have also shown that poor quality data has impact on the organization's culture because data is a key input in the decision-making process. Some researchers have also proved that poor quality data compromises the confidence and acceptability of such data among its users.

1.1.1 Impact of HIV and AIDs to the economy

Acquired Immune Deficiency Syndrome has the potential to create severe economic impacts in many African countries. It is different from most other diseases because it strikes people in their most productive age. The effects vary according to the severity of the AIDS epidemic and the structure of the national economies. Although Kenya's economy is somewhat diversified in terms of Gross Domestic Product (GDP), agriculture is the predominant economic activity. The sector accounts for about 25% of GDP and 70% of export earnings. About 80% of all Kenyans live in the rural areas. Of these, 90% earn their livelihood from agriculture. The main agricultural exports are tea, coffee, and horticultural products, which together account for 45% of total export. Tourism is an important industry, contributing another 19% to overall GDP, and an important source of foreign exchange. Although GDP grew at about 2.5% between 1990 and 1997, the population grew at 2.6%, so that per capita income decreased over that time period. According to Lori, John and David (1999), the economic effects of AIDS are felt first by individuals and their families, and then ripple outwards to firms and businesses and the macro-economy. The two major economic effects are a reduction in the labor supply and increased costs of production. The loss of young adults in their most productive years affect overall economic output and if AIDS is more prevalent among the economic elite, then the impact may be much larger than the absolute number of AIDS deaths indicates. The direct costs of AIDS include expenditures for medical care, drugs, and funeral expenses. Indirect costs include lost time due to illness, recruitment and training costs to replace workers, and care of orphans. If costs are financed out of savings, then the reduction in investment could lead to a significant reduction in economic growth

1.1.2 Importance of data quality in HIV Programmes

Data incompleteness in HIV can hurt program in several ways. Programmes that do not take data quality into account, may have to use additional resources to correct the data. They can also experience reduced stakeholder confidence and support, miss opportunities to identify areas of strength or gaps in programme activities as well as face the undesirable consequences of inappropriate decisions based on poor data. A data management system is the set of standard operating procedures (SOPs) and actions put in place to track the data flow and reduce the probability of errors being introduced into the data at any stage. HIV data may entail visit dates for HIV care and treatment, Anti retroviral start dates, weight of the patient at every clinic visit, assessment of pregnancy in adult female patients at every clinic visit, Tuberculosis status at every patient clinic visit as well as assessment of new opportunistic infections amongst others.

According to Bennett, Mate, Mphatswe, Barker and Rollins (2009), in their study on challenges for routine health system data management in a large public programme to prevent mother-to-child HIV transmission in South Africa, there was large variation in the completeness of data reporting for each element in District Health Management Information System (DHIS). South Africa and Kenya share the same model of health care systems making it important to evaluate the level of data completeness in HIV data elements in data collection tools across the health facilities.

1.1.3 Data in HIV and AIDS Programmes

There are two types of HIV data that are available in Kenya; primary and secondary data. HIV primary data are obtained from surveys, focus groups, personal interviews, experiments and

observational studies. According to Kenya Demographic Health Survey (KDHS) that was conducted in 2009, HIV primary data can as well be collected routinely during patient's normal clinic visits. It was noted in KDHS report that routine data are readily available, exhibits low costs in collection, are useful for initial assessment and can provide baseline information on expected levels of health or disease. HIV primary data has several limitations which include the amount of time and money spent designing data collection instruments, selection of the sample, the pretesting of the instruments to work out sources of bias, administration of the instrument and entry and collation of data. Routine data are not always up to date, lack completeness; important variables may not be collected and sometimes can be subjected to manipulation.

HIV secondary data may be collected from chart reviews, data collection and collation registers in health facilities and study materials. HIV secondary data may lack completeness, consistency, reliability and may exhibit other confounding factors like sample selection bias, source choice bias and in time series some observations may drop out over time.

Manjunath Manjunath, Ravindra and Ravikumar (2011) states that the primary goal of a Data Quality Assessment (DQA) is to determine if the data set is usable for its intended purpose and be used with a required degree of confidence. Also, timely and accurate data assessment is necessary to ensure that field and analytical activities produce data that meet the project Data Quality Objectives (DQOs). The evaluation and assessment of data quality include four separate activities: data verification, data validation, statistical data quality assessment, and peer review.

Although most data gathering disciplines treat error as an embarrassing issue to be done away with, the error inherent in data deserves closer attention and public understanding because error

provides a critical component in judging fitness for use. The main purpose of data quality is to correct the data which is missing or incorrect or invalid in some perspective (Pipino *et al.*, 2002)

1.2 Statement of the problem

Eastern province health department has a problem of poor quality data from health facilities. The data is inaccurate, inconsistent, incomplete and untimely. District Health Information Records 2010-2011 reveals variations and inconsistencies in HIV data completeness. According to Data quality report card-Kenya (Kenya Ministry of Health Abstract, 2011), Eastern province was cited as one of the provinces with health facilities with HIV data completeness below 90%. Although a lot is being done to ensure that HIV information is documented for patient's medical history and follow-ups, erroneous data recording, missing information, unreadable and inconsistent information in HIV data collection tools makes prevention, care and treatment services become difficult.

According to NACC and NASCOP (2012), the number of People Living with HIV in Eastern province by 2010 was 51,000 and the projections for the cumulative numbers by the end of 2012 were 65,000. Failure to document HIV data parameters completely can increase the risk of Mother to Child Transmission of HIV and increased mortality rates. In addition, lack of proper documentation of HIV data can lead to missed opportunities in screening the presence of any Tuberculosis and other Opportunistic Infections which are the leading causes of mortality cases in people living with HIV.

Data completeness in HIV programs is enhanced by data quality assessment for completeness which improves prevention, care and treatment services for People Living with HIV. Therefore,

this study seeks to investigate the level of HIV data completeness, factors that lead to HIV data incompleteness and health care providers' competence in conducting a HIV data quality assessment for completeness across the Ministry of Health facilities.

1.3 Justification of the study

Treatment of people living with HIV is mostly based on their medical history experiences. Patient's medical history is provided in HIV data collection tools. HIV data should be free of erroneous recording, missing information, unreadable and inconsistent information. HIV data completeness enhances effective prevention, care and treatment services to people living with HIV.

1.4 Objectives of the study

1.4.1 General objective

To evaluate the quality of data collected in Ministry of Health facilities offering HIV prevention, care and treatment services in Eastern Province of Kenya.

1.4.2 Specific objectives

- i. To establish the level of HIV data completeness in patient blue cards across the Ministry of Health facilities in Eastern Province of Kenya.
- ii. To identify factors that lead to missing data values in various HIV data collection tools across the Ministry of Health facilities in Eastern Province of Kenya.

- iii. To determine the effect of training health care providers in health management information systems on competence in conducting HIV data quality assessment for completeness.
- iv. To determine the effect of health care providers experience in filling health records and gender on competence in conducting HIV data quality assessment for completeness.

1.5 Hypothesis

Null hypothesis:

- i. There is no data incompleteness among patient blue cards across the Ministry of Health facilities in Eastern Province of Kenya.
- ii. There are no factors that lead to missing data values in various HIV data collection tools across the Ministry of Health facilities in Eastern Province of Kenya.
- iii. Training health care providers in health management and information systems doesn't have effect on their competence in conducting a HIV data quality assessment for completeness.
- iv. Experience in health care provision has no effect on the competence on how to conduct a HIV data quality assessment for completeness.

CHAPTER TWO: LITERATURE REVIEW

2.1 Data Quality Dimensions

There are diverging views on what data quality means, although the classical definition refers to the degree to which the data being used serves the needs of the user (Haug and Arlbjørn, 2011). The primary goal of data quality evaluation is to determine whether the data set is usable for its intended purpose and can be used with the required degree of confidence. Timely and accurate data assessment is necessary to ensure that field and analytical activities produce data that meet the project data quality objectives. The evaluation of data quality includes four separate activities: data verification, data validation, statistical data quality assessment, and peer review (Haug and Arlbjørn, 2011).

According to EPA (2006), production of high quality statistics depends on the assessment of data quality. Without a systematic assessment of data quality, the statistical office will risk to lose control of the various statistical processes such as data collection, editing or weighting. Data quality assessment is a precondition for informing the users about the possible use of the data, not to mention that it provides safe grounds on which results may be published with or without a warning. Indeed, without good approaches for data quality assessment statistical institutes are working in the blind and can make no justified claim of being professional and delivering quality data at all.

EPA (2006) shows that poor data quality tends to have undesirable impact on its effectiveness. Consequently, researchers tend to ensure that quality data exhibits accuracy; uniqueness;

completeness; consistency; conformance; currency; referential integrity, and timeliness. These different dimensions are used in assessing data quality.

According to Pipino *et al.* (2002) in their discussion on data quality dimensions, the first quality aspect is the free-of-error dimension, which represents data correctness. This is the correctness of the data being reported, and it should reflect the object or event being described in the real world. In many instances, accuracy of data refers to how values agree with the reference data; hence, minimizing errors to a negligible level. If one is counting the data units in error, the metric is defined as the number of data units in error divided by the total number of data units subtracted from 1. In practice, determining what constitutes a data unit and what is an error requires a set of clearly defined criteria. For example, the degree of precision must be specified. It is possible for an incorrect character in a text string to be tolerable in one circumstance but not in another.

According to Loshin (2006), uniqueness of data is the degree to which the entities modeled by the data are captured and represented in a unique manner. This dimension seeks to ensure that no entity is captured more than once within a set of data, and that a single key can be uniquely used to access the entity. This dimension is usually monitored through a static assessment or ongoing monitoring process. A static assessment involves conducting a duplicate analysis on the data to establish duplicate records that may exist, whereas ongoing monitoring process entails providing an identity match and resolution effect when creating the records to identify exact or likely matching records.

Pipino *et al.*, (2002) explained the completeness dimension as another aspect of quality data which can be viewed from many perspectives, leading to different metrics. At the most abstract level, one can define the concept of schema completeness, which is the degree to which entities and attributes are not missing from the schema. At the data level, one can define column completeness as a function of the missing values in a column of a table. This measurement corresponds to Codd's column integrity, which assesses missing values. A third type is called population completeness. For example, if a column should contain at least one occurrence of all 50 states, but it only contains 43 states, then there is a case of population incompleteness. Each of the three types (schema completeness, column completeness, and population completeness) can be measured by taking the ratio of the number of incomplete items to the total number of items and subtracting from one. Generally, the completeness of data is the extent to which the expected and mandatory characteristics of the data are captured.

Bennett *et al.*, (2009) in their study noted that there was large variation in the completeness of data reporting for mother to child data elements that were assessing for programme outcome. The study analyzed six data elements from 316 sites reported over 12 months to the district health information system. The study showed that the data were complete only half the time (50.3%). There was variation in the districts that were included in the study. It was rather evident that the levels of data incompleteness, inconsistency and or inaccuracy prevailed in data collection and reporting tools. Bennett *et al.* (2009) continues to stipulate that consistency of data increases its reliability because the data does not change regardless of the number of times it has been used or irrespective of the user utilizing the data. It has been observed that consistency of data may not necessarily mean correctness of the data.

Pipino *et al.* (2002) states that good quality data should also possess timeliness attributes in terms of how much lag time to expect before the information becomes accessible and available for use. This dimension is measured in terms of the lag between when information is anticipated and when it becomes available to the user.

According to McGilvray (2008), the currency attribute in data quality refers to the extent to which information is up to date with the real world it seeks to model. Current data is up to date with the real world despite the time-related changes that may have occurred. This attribute is expressed as a function of the expected rate of frequency at which the different elements of data will be refreshed to ensure that it is up to date. This process may be conducted manually or by an automatic process. Essentially, currency of data is the lifetime of the data after which they must be refreshed to ensure they reflect the real world objects or events.

Cervo and Allen (2011) states that data conformance dimension contains several metadata attributes relating to its data type; storage formats; precision; domain ranges, format patterns, among others. However, sometimes it becomes necessary to use multiple quality indicators (variables); hence, the minimum or maximum operation may be applied. One computes the minimum (or maximum) value from among the normalized values of the individual data quality indicators. The minimum operator is conservative in that it assigns to the dimension an aggregate value no higher than the value of its weakest data quality indicator.

Fennel (1988) argues that the maximum operator proves useful in more complex metrics applicable to the dimensions of timeliness and accessibility. Among other factors, it may reflect an individual's assessment of the credibility of the data source, comparison to a commonly accepted standard, and previous experience. Lee and Strong (2004) define believability as "the extent to which data are accepted or regarded as true, real and credible". Their survey shows that data consumers consider believability as an especially important aspect of data quality. Besides, the authors characterize believability as an intrinsic data quality dimension. Believability is itself decomposed into sub dimensions. Lee and Strong (2004) propose three sub dimensions, namely believability: of source, compared to internal common-sense standard, and based on temporality of data.

Jones and Lewis (2004) state that quality reports are a significant data quality assessment especially for producers and users of official statistics because they report on the quality of a survey or statistical product. Quality reports can either be user-oriented reports or producer-oriented. User-oriented reports provide a means by which the user can evaluate the strengths and limitations of statistics, as well as competence on how to use the statistics properly. On the other hand, producer-oriented reports provide a means for producers to analyze the results of earlier product developments, and to pick up areas that need further improvements. They provide information based on parameters such as sampling and no-sampling errors; indicator examples; user satisfaction in terms of relevance; key production dates to capture the timeliness dimension; changes that have occurred; coherence with related sources; survey and methods employed, among others.

Quality indicators are also used as DQA tools. They are specific, measurable elements used to characterize the quality of data in statistical practice (Jones and Lewis, 2004). Statistical indicators are usually calculated in order to characterize a real-world phenomenon. Indicators alter the measurability level to interval scale or ordinal from nominal to establish indicators that relate to the real phenomenon closely. Sometimes, they do not measure quality directly; instead, they provide information that gives insight into the quality of the data.

Jones and Lewis (2004) continue to stipulate that quality indicators are categorized as orientation, reference, or quality components. Orientation indicators refer to those that measure quality of data from the producer's perspective. Reference indicators are those that measure a statistical product's quality, for instance, a product variable could be used to measure the quality of a process in statistics such as data collection.

2.2 Steps in Data Quality Assessment

Data quality assessment provides a platform on which users of the data are adequately informed of its possible uses; hence, they are in a better position to draw justifiable conclusions. According to EPA (2006), DQA involves evaluating real-world data scientifically and statistically to determine whether the data meets the objectives of a project in terms of type, quantity, and quality to support the intentions of the user. The technique is based on the principle that quality data is useful only if it meets the user's needs.

EPA (2006) shows that the first step in the DQA process is to review the project's data quality objectives established during the planning stage to assure that they are still applicable. If DQOs were not developed during the planning stage, the objectives should be defined prior to evaluating the data. Understanding the objectives of the study gives the user insight on the purposes of data collection, as well as providing a platform for understanding the qualitative and quantitative principles for assessing the quality of the data collected. The sampling design gives the user relevant information including the assumptions made when designing the sample under observation, not forgetting the relationship between the objectives of the study, and the assumptions of the design. This stage serves as preparation for subsequent steps of the assessment, and preliminary data review. For example, when evaluating data for use in making decisions, the DQOs should define the statistical hypothesis and specify the tolerable limits on decision errors for estimation problems and the acceptable confidence or probability interval width. Review the sampling design and data collection documentation for consistency with the DQOs.

The second step in the DQA process is to conduct a preliminary data analysis to gain an overview of the sampling results and to identify trends or anomalies that could have adverse effects on data analysis (EPA, 2006). Essentially, the reviewer seeks to understand the calculations of the statistical quantities and any graphical representations to understand the data structure that emerges; thus, assessing the strengths and drawbacks of using the data. This is accomplished by reviewing the data validation reports, calculating of basic descriptive statistics, and graphing or plotting the data. At this stage, the reviewer uses statistical quantities from the sample such as median, mean, variance, and range to describe the data numerically. The

estimates are then used to make conclusions about the populations from which the sample was obtained (EPA, 2006). On the other hand, the reviewer uses graphical representations to identify relationships and trends, confirm, or dispute the underlying assumptions, and to identify potential challenges.

In the third stage, the reviewer conducts a quality assurance review using laboratory quality control reports, quality assurance reports, or validation reports to identify the actual data collection procedures, potential problems, and quality control failures.. If there is no evidence of data validation, then, the DQA process must stop until the data are validated. If data qualifiers were applied to the data, then the underlying reasons for those qualifiers and any confirmation results should be assessed to determine the usability of data based on the project DQOs. Communication with data validators may be required to verify or confirm anomalies in the data set.

Beckman (1983) points out that after conducting a quality assurance review, then, investigation for historical outliers should be done. This could be done by use of histograms and box plots to investigate on data normality. The reviewer then checks the data to ensure there are no anomalies. Data anomalies that may be seasonal or project specific should be noted. For example, a weight loss of a patient due to poor feeding trends , famine or an increase in weight of a patient as a result of being pregnant. In this case, the project manager should make a decision on whether or not the data should be accepted as is or qualified.

The next step involves checking the standards and compliance of the data. The reviewer checks data against applicable state/government policy criteria or action levels for compliance. According to Chang (2007), the reviewer should confirm any deviations on criteria or action levels that may be suspect or challenged, and provide appropriate comments in the final report.

According to Tayi and Ballou (1998) descriptive statistics are necessary to quantify characteristics of the data in scientific studies. For example, when comparing a population at a normal site to one at a contaminated site, calculating descriptive statistics will aid in the identification of population differences. Some useful statistical quantities are: number of observations; measures of central tendency such as mean, median, or mode; measures of dispersion, such as range, variance, standard deviation, coefficient of variation, inter quartile range; measures of relative standing, such as percentiles and quartiles; measures of distribution symmetry or shape; and measures of association between two or more variables, such as correlation.

According to Lee and Strong (2004), the next stage is to draw graphs from the data to identify patterns and trends, relationships, or potential anomalies that might go unnoticed by just looking at tables of data. Graphs can quickly disprove or confirm hypotheses. The reviewer must then select the statistical method that best suits the data. According to Lunani *et al* (1997), once the characteristics of the data are known; the appropriate statistical method can be chosen to draw conclusions from the data. The method selected should be influenced by the sampling design used for data collection, assumptions made when determining the objectives, and the type of

distribution of the data. Statistical methods should provide appropriate procedures for summarizing and analyzing the data based on the Data Quality Objectives (DQOs), the sampling design, the preliminary data review, assumptions made in setting the DQOs, and assumptions necessary for analyzing the data. Tests should be robust, or not seriously affected by moderate deviations from the underlying assumptions. The reviewer should note any sensitive assumptions where relatively small deviations could jeopardize the validity of the test results (Lunani *et al.*, 1997).

The next step involves selecting the appropriate set of procedures to be used in summarizing and conducting data analysis. This stage also involves identifying the critical assumptions based on the statistical test. The reviewer should assess the validity of the statistical test chosen by examining whether its underlying assumptions hold, or whether departures are acceptable, given the actual data and other information about the study. Minor deviations from assumptions are usually not critical, as the robustness of the statistical technique used is sufficient to compensate for such deviations. If it is determined that one or more of the assumptions are not met, then either a different statistical method is selected, or additional data may be collected to verify the assumptions. For example, if the assumption of normality in the distribution of data is not met, then a non-parametric test can be used, or the data can be transformed (Lunani *et al.*, 1997).

In the final step of the DQA process, the appropriate statistical tests are performed so that the data user can draw conclusions that address the project objectives. This step involves undertaking the relevant statistical tests calculations, from which conclusions are drawn. If the

first four steps have been performed, then the calculations and conclusions will be scientifically defensible. If the sampling design is to be used again, the performance of the design should be evaluated (Lunani *et al.*, 1997).

US EPA (2006) notes that the process of conducting data quality assessment is critical because it mitigates the possibility of collecting poor quality data that can hamper an organization's activities by casting doubt on the data that feeds both operational and analytical processes. Yet, the absence of clearly defined measurements to demonstrate how organization impacts are attributable to erred data; prevents developing an appropriate organization case for introducing data quality management improvement. When resources are allocated to correcting "bad data" without being able to evaluate the root causes there is no ability to objectively evaluate the relationship between poor data quality and organization performance. This suggests a need for an approach that can quickly identify high priority data issues whose remediation can be justified.

2.5 History of HIV Data in Kenya

Past attempts to impose rigid data based standards in the application of clinical judgment have met resistance from healthcare providers, and thus the need for rigorous data quality assessment was of little consequence. Today, with the increasing deployment of decision support systems at the provider level, application of data quality improvement is less likely to be found objectionable in the establishment of standards for information application and management in health care, in the care and treatment of patients as well as the management of the system where

they are treated. Ballou (1985) notes that such information-based scheme requires a system of continuous data collection, evaluation, feedback, and adjustment of the health care process.

A number of quality improvement measures have been attempted in response to this development, including the use of health outcomes data in clinical practice, which are objective measures of post-treatment health, based on defined criteria. Fennel (1988) states that large outcomes databases of similar procedures are maintained and analyzed statistically to provide the patient with an objective prognosis for success or failure. Therefore, a patient contemplating heart surgery, for example, can compare him to other patients with similar characteristics who have undergone the surgery, and obtain a statistical prediction of success with the surgery, as well as chances for survival without it.

2.6 Factors Affecting HIV Data Quality in Kenya

HIV data in Kenya is obtained from the National Health Information System, which is a combination of health statistics from Kenya National Bureau of Statistics (KNBS); Ministry of Health, research institutions; private health facilities; faith-based organizations, among others. There have been concerns that the tools in use for HIV data collection are not effective, with researchers arguing that they are not complete mainly because of the numerous cases of missing values as far as HIV statistics are concerned. For instance, Otworld *et al* (2007) conducted a study on 332 VCT centers in Kenya in 2004, and found that 298 units or 89.76% of the sample reported some missing data. However, they managed to update their records by referring to the logbooks and completed client forms. This study also revealed a big discrepancy between actual

records, and what was reported in the national database. Essentially, there were 193,960 client records as reported by the national database from the second quarter of 2001 to the second quarter of 2004. This study revealed a further 220,944 records upon visiting the 332 VCT centers, which clearly shows that there were several missing records, and the outcome is that the government ended up with an underestimation of the number of HIV positive individuals.

The following have been identified as key factors that affect the quality of HIV data in terms of completeness and effectiveness; A report compiled by AMREF (2010), Kenya revealed that shortage of staff in the national HIS adversely affects the quality of the data collected. For instance, the institution had only 17 staff in 2010, with key employees being the Provincial Health Records and Information Officer (PHRIO) and the District Health Records and Information Officer (DHRIO) who coordinate HIV and AIDS data at provincial and district levels respectively. Consequently, staff members are expected to double up their roles; thus, increasing the likelihood of compromising the quality of data collected as employees concentrate more on their main tasks than on recording data. Moreover, there is a tendency to hire inadequately trained officers to collect data at various levels, which is likely to interfere with the quality of the data.

Patients must be assured that the health information they share with healthcare professionals will remain confidential. Without such assurance, patients may withhold critical information that could affect the quality and outcome of care. According to Amadeus International (2009), the privacy and confidentiality of patient health information has been protected by a patchwork of federal and state laws and regulations, facility policy, professional standards of practice, and codes of ethics. The recently passed Standards for Privacy of Individually Identifiable Health

Information under HIPAA establishes requirements for the protection of health information maintained by health plans, healthcare clearinghouses, and healthcare providers who transmit certain transactions electronically. These covered entities will likely need to establish or modify existing policies and procedures to comply with this new legislation.

The article stresses on three major areas: Confidentiality, for the purpose of this article, is the practice of permitting only certain authorized individuals to access information with the understanding that they will disclose it only to other authorized individuals as permitted by law. For example, HIV information may not be released without specific consent. Privacy, for the purpose of this article, is an individual's right to control his or her protected health information. Security is the protections or safeguards put in place to secure protected health information (Amadeus International, 2009)

According to Alfano and Sandra (2007), healthcare organizations are entrusted with the most private information of their HIV patients and employees. They have a legal, moral, and ethical duty to protect all clinical and research information by ensuring that security and privacy safeguards are in place. A higher degree of control is necessary to prevent unauthorized access to especially sensitive information. This is particularly true within the context of the electronic health record (EHR).

Alfano and Sandra (2007) continue to state that the practice of controlling unauthorized access to health information identifies categories of health information that are accorded special protections under law or may otherwise require a higher degree of security. It recommends system features and practices that will help secure sensitive information in EHRs and afford

organizations satisfactory assurances that sufficient safeguards shield this information from misuse hence ensuring data completeness in our health facilities.

AHIMA (2006) states that by training physicians, nurses, health workers, and administrators, new generations of healthcare providers are learning to deliver comprehensive, community-based care key to it being ensuring that data is well collected, stored and transferred when necessary in the right manner. Training is one of the most important ways of reinforcing, replicating, and transmitting correct patient information that needs to be handled with care, confidentiality and in a secure manner. Trainings on data collection, entry and data storage systems are important to enhance data completeness. However, in order to successfully exchange health information and reap the benefits of health information systems, organizations must maintain accurate patient identification information. Patient identification integrity is a complex concept, and one that is not well understood throughout the healthcare industry. Many policy makers and industry leaders do not fully comprehend the negative effects of inaccurate patient identification information for even basic health information interchange.

Seemingly mundane errors such as transpositions in a patient's birth year, misspellings or culturally acceptable spellings of a patient's last name, and nicknames or default Social Security numbers (SSNs) can play havoc with successfully linking electronic records across clinical and administrative systems. All these errors can be corrected by proper trainings to all the health care providers. AHIMA (2006) notes that recent research has shown that good exam results are not enough to ensure the expected results are achieved in the health care facilities. Relevant work experience is as vital when handling patient's information and especially how to gain confidence

from the patient so that s/he is able to release sensitive information. Health Care providers should be trained on the health management information systems prior to handling patients so that they are prepared to handle patient's information. Organizations handling health information should keep on training their staff on the current trends so that Kenya is at par on health information handling with other global partners.

CHAPTER THREE: RESEARCH METHODOLOGY

3.1 Study design

The study was a descriptive cross sectional design in which 60 health care providers and patients' records were reviewed in 42 Ministry of Health facilities. The study sought to establish the level of HIV data completeness and identify perceptions behind missing HIV data. In addition, the study also sought to establish the effect of health care providers experience in filling health records, gender and training them in health management information systems on competence of how to conduct a HIV data quality assessment for completeness. This was done by administering an anonymous questionnaire to health care providers (Appendix 3). Secondly, from each health facility, ten percent of patient records were randomly identified and evaluated for missing HIV data. Among the HIV data considered for evaluation of missing entries were weight, pregnancy status for women above the age of fifteen years, presence of opportunistic infections and patients Tuberculosis status in all the patients' clinic visits.

3.2 Study population

3.2.1 Health Facilities

Ministry of health facilities that offer comprehensive care clinic services formed the study population.

3.2.2 Health care providers

Health care providers who work in the Ministry of Health comprehensive care clinics in the ICAP supported health facilities formed the study population.

3.3 Exclusion criteria

3.3.1 Health Facilities

Health facilities that are not supported by International Centre for AIDs Care and Treatment Programmes (ICAP) Kenya were not included in the study. The study sample was determined from 87 health facilities that are supported by ICAP.

3.3.2 Health care providers

Health care providers that were not involved in HIV related primary and secondary data collection at facility level were excluded from the study. Medical officers of health and other medical matrons who involved themselves more in administrative matters than medical service practice at Comprehensive Care Clinic were also not included in the study.

3.4 Sampling design

3.4.1 Sampling frame

3.4.1.1 Health care providers

Employees who work in Ministry of Health facilities that are supported by ICAP and handle HIV data collection at primary and secondary levels formed a sampling frame. They included Health records officers, clinicians and data entry clerks. A self administered semi-structured questionnaire was distributed to sampled employees (Appendix 3). All participants were required to complete a consent form which entailed the requirements and ethical standards in relation to patients' information, benefits and risks of being in the study (Appendix 1).

3.4.1.2 Health Facilities

Health facilities that offer care and treatment services to People Living with HIV and are supported by ICAP formed the sampling frame. From each health facility, patients blue cards formed the sampling frame for review of HIV data completeness.

3.4.2 Sampling technique

3.4.2.1 Health Facilities

Purposive sampling was used to select the Health facilities, in this case, those that were being supported by ICAP. Then random sampling was used to select 42 health facilities and patients' blue cards randomized using computer generated numbers and ten percent of the randomized blue cards picked for evaluation of HIV data completeness.

3.4.2.2 Health Care Providers

Purposive sampling was used to select the health care providers; in this case, those who worked in ICAP supported comprehensive care clinics. Then simple random sampling was used to ensure an equal representation of purposively selected health care providers.

Mugenda and Mugenda (1999) provides the following automated excel spreadsheets randomization formula that was used to randomize the number of health care providers, health facilities as well as patients records unique numbers. {= RANDBETWEEN (top, bottom)} whereby 'top' represents the first number allocated to the first participant and 'bottom' represents the last number in the list as allocated for randomization. From the randomized numbers that are assigned to participants, 60 health care providers were then picked at random. Fog and Agner (2004) states that drawing a random sample helps to reduce the sampling error.

3.4.3 Sample size

Cooper and Schindler (2006) refer to the economic advantage of taking a sample rather than a census. According to Mugenda and Mugenda (1999), the size of the sample was chosen using the following formulae:

$$\underline{n} = n / \{(1+n)/N\}$$

a) Number of Health Care Providers study participants

$$60 = 113 / \{(1+113/128)\}$$

b) Number of health facilities selected for evaluation of patient records

$$42 = 81 / \{(1+81/87)\}$$

Where \underline{n} = the sample size

n is desired sample size when population is less than 10,000

N is the population

3.5 Research procedures

The respective questionnaires were first pretested on three health care providers each from a different health facility. This confirmed how the questions were understood by the participants. The information that was obtained by the questionnaire included participant's gender, department of duty and the duration worked in the department of duty. In addition, the question on whether the health care provider is trained in HMIS and the causes of data incompleteness in the comprehensive care clinic was also asked. A database for data entry, analysis and presentation was developed using Epi Info statistical software's. The interviewer then issued out the questionnaires and had them dropped off in a collection box in the respective offices after they were filled in by the respondents. The interviewers were on the sites during data collection to collect and ship research tools for data entry. A master list (Appendix 5) was developed to

help track data entry process. Data analysis and modeling of how experience, training in health management information systems and gender affect competence on how to conduct a HIV DQA for completeness was done. Archiving of used research master list and questionnaires was there after done for future reference. Experience was depicted by the number of months a health care provider has worked in the comprehensive care clinic. Patient records were evaluated for data completeness and the missing HIV data tabulated. The data was then presented in tables, figures and charts.

3.6 Data collection methods

Primary data from health care providers was collected using a self administered semi structured questionnaire which contained both closed and open ended questions. Information obtained from the questionnaire addressed the experience of health care providers in comprehensive care clinic as well as the perceptions regarding missing HIV data values in HIV data collection tools (Appendix 3). Data completeness was addressed by reviewing patients' blue card for data completeness. The proportion of patient records that were evaluated for missing values was 10 percent of the cumulative records in each health facility in the 42 health facilities. Tallying of the missing HIV data for all the entries that were evaluated was done and proportion missing determined

3.7 Data Analysis Methods

Each question from the questionnaire was analyzed separately by getting the sums, proportions and percentages. A Pearson's Chi-square test was done to test the association between the groups in experience, gender, training in health management information systems of the health care

providers and their respective competence in conducting a HIV data quality assessment for completeness.

Then, a binomial logistic regression was used to model the competence of health care providers on how to conduct a HIV data quality assessment for completeness against gender, training of health care providers in HMIS and experience in filling HIV data collection tools depicted by number of months worked in the facility as the logistic regression model predictor variables.

$$\text{logit}(p) = \log(p/(1-p)) = \beta_0 + \beta_1 * x_1 + \dots + \beta_k * x_k$$

Where:

Log (p/ (1-p)) is the odds ratio

β_0 is the constant

$\beta_1 * \dots * \beta_k$ are the model predictor variables coefficients

$x_1 \dots x_k$ are the predictor variables

Qualitative responses from open ended questions were grouped together and respondents with similar opinions were tallied, grouped and results presented in figures. Health facilities average complete entries for weight, pregnancy, TB screening and new opportunist infection were tabulated and presented in a chart. Difference in the total entries evaluated for completeness and the entries complete was tabulated and the outlying health facilities identified. Data was considered incomplete if there was any missing entry i.e. below 100% level of completeness. In Kenya, HIV data is expected to be 100% complete this being according to UNAIDs (2010).

CHAPTER FOUR: RESULTS

4.1 Evaluating the level of HIV data completeness

It was noted that from the 42 Ministry of Health facilities, Kitui District Hospital, Matuu District Hospital and Ekalakala Health Centre had the most extreme data variations in total HIV data entries evaluated and entries complete (Appendix 6). The number of patient's visits should match the number of entries in the patients' records. Variations noted were presented in the table below:

HIV data elements	Total Visits			% Missing entries
	checked for data completeness (n)	Entries Present	Missing Entries	
Weight	3829	3565	264	6.9%
Pregnancy	3065	2735	330	10.8%
TB status	3829	3535	294	7.7%
New OI	3829	3413	416	10.9%

Table 1: Pooled HIV data elements completeness for different Ministry of Health facilities

Figure 1 below shows the proportions of missing entries in pregnancy for women above the age of fifteen years, tuberculosis screening, new opportunistic infection screening and weight documentation at every visit.

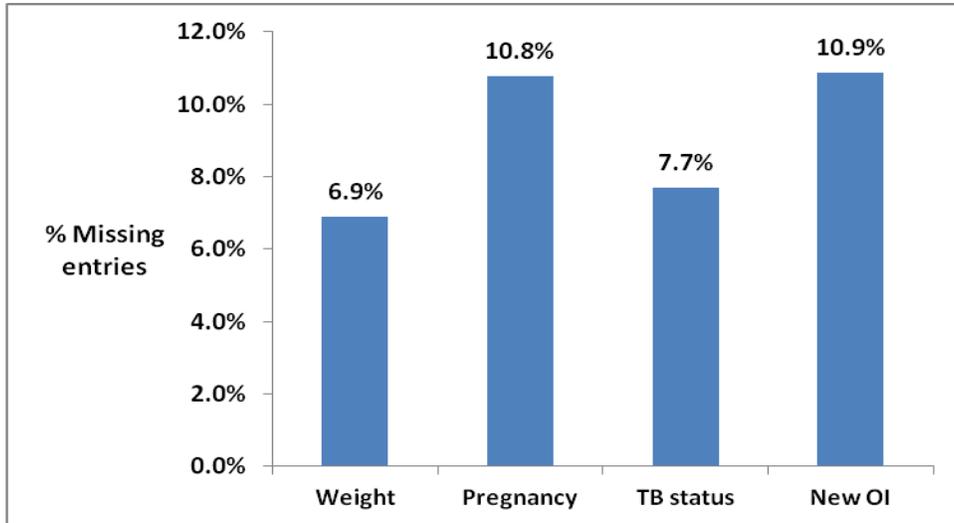


Figure 1: Proportion of missing HIV data entries in patients blue card

Weight had the least percentage difference in total entries evaluated for completeness and entries complete (6.89%) followed by Tuberculosis at 7.68%. Screening of pregnancy status had the highest percentage difference in total entries evaluated for completeness and entries complete at 10.86%.

4.2 Perceptions regarding HIV data incompleteness

Health care providers who worked in the comprehensive care clinic for people living with HIV give the following perceptions behind missing entries in HIV data records.

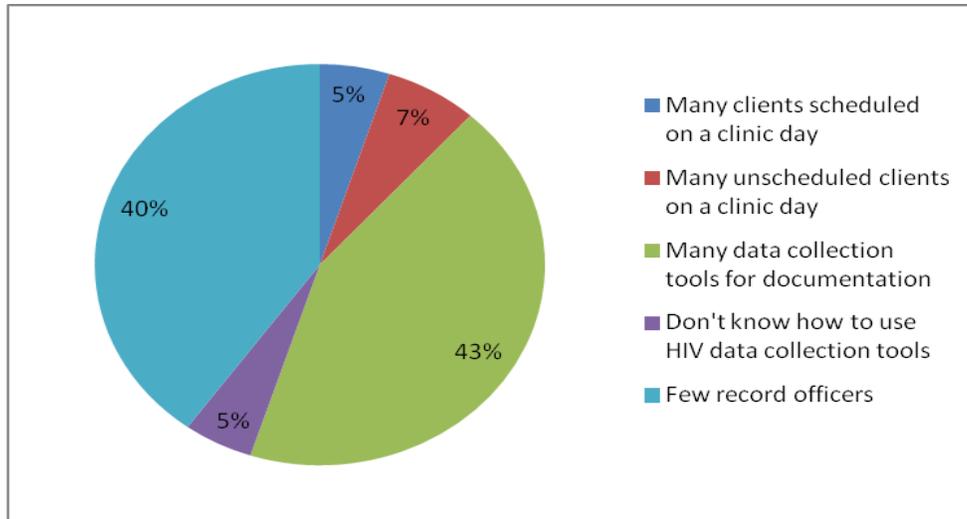


Figure 2: Perceptions regarding HIV data incompleteness in health facilities

Many data HIV collection tools for documentation were perceived as the major cause of data incompleteness in HIV data collection tools at 26 (43 percent), of the overall participants. This was followed by the perception that few data handling personnel's in the comprehensive care clinic at 24 (40 percent) contributed to HIV data incompleteness. Many unscheduled clients attending routine clinic on a certain day 4 (7 percent) was pointed out as one of the reasons for HIV data incompleteness in health facilities while too many HIV data collection tools and health care providers not knowing how to complete HIV data collection tools came last at 3 (5 percent) respectively. The participants could not identify any other factor that they felt was a major contributor of HIV data incompleteness in health facilities.

4.3 Experience in handling HIV data collection tools

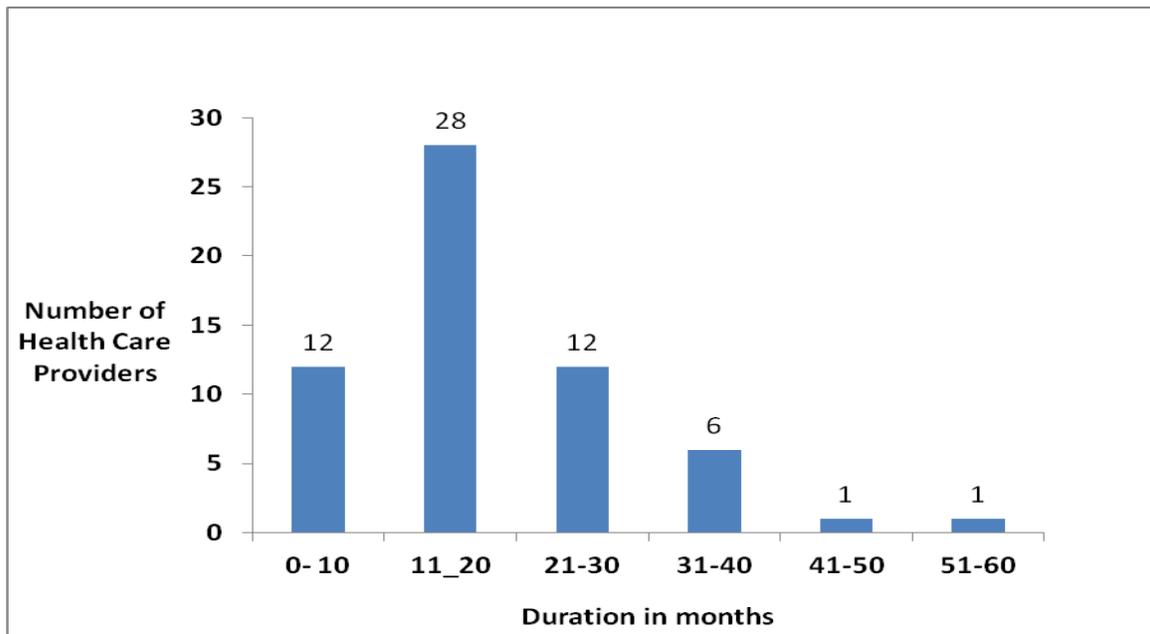


Figure 3: Duration worked in months by health care providers in the comprehensive care clinic

The average duration of participants who had worked in the comprehensive care clinic was 19.4 months with a mode and median class of 11 to 20 months, a minimum months of 3 and maximum months of 60. This shows that 48 participants have had an experience with data handling tools for at least 10 months of working experience in the comprehensive care clinic. There was one health care provider who had worked in the comprehensive care clinic for five years. The total proportion of participants who had ever been trained on HMIS was 53.3 percent which was 32 participants out of 60. Health management and Information systems training entails the standard procedures on how to document HIV records and related reports.

4.4 Logistic regression analysis and modeling of competence on how to conduct a HIV DQA for completeness by health care providers

A chi-square test was calculated at 95% confidence level to determine the independence of the predictor variables before doing a logistic regression analysis for the model. Results are presented in the following table:

		Chi-square (P-value)
Gender	Yes	0.704
	No	
Months of experience	10 Months	0.212
	24 Months	
	36 Months	
Trained in HMIS	Yes	0.078
	No	

Table 2: Calculated Chi square results (P-value) for Gender, experience and training on HMIS

There was no association between the group in the predictor variables since all the Pearson's Chi-square calculated p-value results were not significant ($P > 0.05$). A logistic regression model allows us to establish a relationship between a binary outcome variable and a group of predictor variables. It models the logit-transformed probability as a linear relationship with the predictor variables. A logistic regression model was fitted with health care providers trained in HMIS, gender and experience in filling the HIV data collection tools was used as the predictor variables.

The logistic regression model was fitted using the above constant and coefficients in column B of (Table 3)

	B	S.E.	Wald	df	Sig.	Exp(B)
Trained in HMIS (Yes)	0.129	0.036	2.922	1	0.007	0.359
Gender (Male)	-0.147	0.625	0.056	1	0.814	0.863
Experience (Yes)	0.156	0.026	6.1	1	<0.000	0.998
Constant	1.558	0.716	4.73	1	0.03	4.747

Table 3: Logistic regression predictor variables and their respective coefficients

$$\text{logit}(p) = \log(p / (1-p)) = \beta_0 + \beta_1 * x_1 + \dots + \beta_k * x_k ,$$

Where:

Log (p/ (1-p)) is the odds ratio

β_0 is the constant

$\beta_1 \dots \beta_k$ are the model predictor variables coefficients

$x_1 \dots x_k$ are the predictor variables

$$\text{logit}(\text{Competence in conducting HIV DQA for completeness } (p)) = \log(p/(1-p)) = 1.558 + 0.129(\text{Training in HMIS}) + 0.156(\text{Experience}) - 0.147(\text{Gender})$$

This fitted model denotes that holding experience constant, the odds of a health care provider knowing how to conduct a HIV DQA for completeness and having been trained in HMIS over the odds of not having been trained is $\exp(0.129)$ which is 0.359. This applies also to the odds of having an experience in health service provision, where the odds for knowing how to conduct a HIV DQA for completeness over the odds for not knowing having had an experience is $\exp(0.156)$ which is 0.998.

The P-Value for Health care providers having been trained in HMIS is 0.007 which is less than $P=0.05$ at 95% confidence interval. Therefore, we reject the null hypothesis which states that there is no relationship between training of health care providers in health management information systems and their competence in conducting a HIV data quality assessment for completeness. This applies to the experience in filling data collection records where the $P<0.000$ is less than $P=0.05$. Therefore we reject the null hypothesis which stated that experience in health care provision does not determine competence on how to conduct a HIV data quality assessment for completeness.

CHAPTER FIVE: DISCUSSION

The study sought to establish the level of HIV data completeness in health facilities. The results revealed that in all the four HIV data elements that were evaluated for complete entries in 42 health facilities, noticeable difference in level of data completeness was evident. Weight and TB status had least percentage difference in data completeness while new opportunistic infection screening and pregnancy screening for women above the age of fifteen years had the biggest proportion of missing entries. It was noted that inadequate record officers in the comprehensive care clinic were perceived as the major cause of missing HIV data in patients' records. The respondents also noted that many data collection records for documentation were perceived as the second major cause of HIV data incompleteness in HIV data collection records.

Otwombe *et al.*, (2007) in their study on data completeness in 332 VCT centers found that 298 units or 89.76% of the sample reported some missing data values. In this study, it was evident that all the HIV data elements didn't yield the same level of data completeness. It is evident that trained and experienced health care providers in health records and information management could fill data collection records better than those without training and experience in HMIS. This is with reference to logistic regression model where their respective P-values were less than the alpha values. Working in CCC for long durations exposed health care providers to standard practices of filling the health records. This offered health care providers an opportunity to undergo through various HIV data collection records. Gender was not identified as a contributing factor to any competence on how to conduct a HIV DQA for completeness.

Kim *et al.*, (2003) in their study stipulated the cost of using data of poor quality. This study found that quality of HIV prevention; care and treatment and data quality management usually go together. Competence on how to conduct HIV DQA for completeness can lead to proper documentation of HIV collection tools which lead to improved prevention, care and treatment of people living with HIV. It was noted that though health care providers know how to conduct HIV DQA for completeness, they entirely don't know how to design tools that could enable them conduct an assessment on other HIV data collection tools apart from the blue card.

According to Clement and Kato (2010), in improving quality of HIV/AIDs and related data quality in Mayuge district in Uganda, data incompleteness in HIV reporting tools was evident. The reports that were evaluated for completeness were 500 from different departments in 150 health facilities. The writer notes that such levels of data incompleteness were attributable to few health records and information officers in health facilities as well as lack of routine district health management system for data quality assessment. The study recommends routine support supervision for health facilities to aid the already overwhelmed record officers on how to complete the HIV collation, aggregation and reporting forms. In this study, data incompleteness was inherent since all the HIV data elements that were evaluated for completeness were below 100% level of completeness. Few record officers assigned to work in the comprehensive care clinic was identified as a major cause of HIV data incompleteness in comprehensive care clinics.

According to WHO (2003), in their study on quality of HIV data for decision making, noted that poor HIV data are as a result of poorly designed data collection forms, inefficient clerical staff, lack of training in health records and lack of time caused by pressure of work. Poorly trained and

insufficient staffs, lack of understanding of the need for accurate data and lack of understanding of the requirements of data collection and data quality by medical officers, nurses and other health professionals are other causes of poor quality data. In this study, it was noted that trained health care providers in health management and information systems enhanced their competence in conducting a HIV DQA for completeness. This is in respect of logistic regression model that the odds of a health care provider knowing how to conduct a HIV DQA for completeness and having been trained in HMIS over the odds of not having been trained is $\exp(0.129)$ which was statistically significant.

The international journal of public health (WHO, 2005) ,also, cites incompleteness and inconsistencies in the collection, reporting, storage and analysis of social economic data and that statistics office give higher priority to economic data than other social statistics. The report recommends that there is need for a long term commitment to improve training and career structures of statisticians and information technicians working in the health and other social sectors.

CHAPTER SIX: CONCLUSION

Documentation of patients' weight, pregnancy status, screening of new opportunistic infections and tuberculosis in HIV records was incomplete. In Kenya, HIV data is expected to be 100% complete this being according to UNAIDs (2010). Many HIV data collection records for documentation and few record officers assigned to work in the comprehensive care clinics were identified as the major causes of missing entries in HIV data collection tools which accounted for 83% of the overall causes of data incompleteness as mentioned by the health care providers. Training health care providers in health management and information systems and having had worked in the HIV clinic for at least ten months contributed positively in the competence of conducting a structured HIV DQA for completeness.

CHAPTER 7: RECOMMENDATIONS

To improve HIV data completeness, HIV data records for documentation should be reduced and more qualified and experienced data handling personnel's employed in the comprehensive care clinics to aid in quality documentation of HIV data. Updating and reconstruction of secondary HIV data collection tools needs to be done to aid quality and consistent HIV data reporting. To strengthen the system of primary patient data collection, health care providers should be provided with frequent mentorship as well as being trained on how to complete HIV data collection records. More research needs to be done in the area of HIV data collection records and data quality assessment.

REFERENCES

AHIMA e-HIM (Mar. 2006) Work Group on Regional Health Information Organizations (RHIOs). "Using the SSN as a Patient Identifier." *Journal of AHIMA* 77, no. 3

Alfano and Sandra L (2007) "Research Documentation and Data Security." www.info.med.yale.edu/hic/docs/ResearchDoc-DataSecurity.ppt. 29/06/2012

Amadeus International (2009) "21 CFR Part 11 Best Practices." www.amadeussolutions.com/english/practices/bp_21cfr_part11.htm. 29/06/2012

AMREF (2010) *Strengthening HIV strategic information in Kenya – Adam consortium: Report on Organizational Capacity in HIV Monitoring and Evaluation Rapid Needs Assessment. 1-157.*

Barnett V and Lewis T (1978) *Outliers in Statistical Data*, John Wiley, New York, NY 104-116

Beckman J and Cook D (1983) *Outliers*, Technometrics 25:119-149

Bennett B, Mate K, Mphatswe W, Barker P, Rollins N (2009) *Challenges for Routine Health System Data Management in a Large Public Programme to Prevent Mother-to-Child HIV Transmission in South Africa*. African Studies Publishers

Cervo D and Allen M (2011) *Master data management in practice: Achieving true customer MDM*. Hoboken, NJ: Wiley

Chang H (2007) Dynamic robust design with missing data *International Journal of Quality & Reliability Management*, 24(6), 602-616.

Cooper R, and Schindler S, (2006) Research Methods: 19-67

EPA (2006) *Data Quality Assessment: A Reviewer's Guide*. 1-55.

Fennell (1988) Data Quality Assessment for completeness and accuracy. Mainz, Germany, 24-26

Fog and Agner (2004) The non central hyper geometric probability distributions.
www.agner.org/random/theory.29/06/2012

Häkkinen L, and Hilmola O (2008) ERP evaluation during the shakedown phase: lessons from an after-sales division. *Information Systems Journal*, 18(1):73-100

Haug A and Arlbjørn J (2011) Barriers to master data quality *Journal of Enterprise Information Management*, 24(3), pp. 288-303

Jones N and Lewis D (2004) Developing a Handbook on Improving Quality by Analysis of Process Variables. Paper presented at the European Conference on Quality and Methodology in Official Statistics (Q2004), Mainz, Germany, 24-26 May 2004

Kenya Demographic Health Survey (2003) Ministry of Health and WHO report

Kenya, Ministry of health abstract, (2011). Enhancing data quality in health facilities

KHPF, (2011) Analysis of Performance. Analytical Review of Health Progress, and Systems Performance; 1994 – 2010, Ministry of Health and WHO report

Lee Y and Strong D (2004) Knowing-why about data processes and data quality *Journal of Management Information Systems*, 20(3):13-39

Lori B, John S and David N (1999) The Futures Group International in collaboration with: Research Triangle Institute (RTI) The Centre for Development and Population Activities (CEDPA)

Loshin D (2006) Monitoring Data Quality Performance Using Data Quality Metrics *Informatica White Paper*, 1-19

Lunani M, Nair V and Wasserman S (1997) Graphical methods for robust design with dynamic characteristics, *Journal of Quality Technology*, 29, 327-38

Manjunath T, Ravindra S and Ravikumar G (2011) International Journal of Computer Science and Information Technologies Analysis of Data Quality Aspects in Data Warehouse Systems , Vol. 2 (1) , 477-485

McGilvray D (2008) *Executing data quality projects: Ten steps to quality data and trusted information*. Amsterdam: Morgan Kaufmann/Elsevier

Kim W, Choi B, Hong E, Kim S and Lee D (2003) A taxonomy of dirty data *Data Mining and Knowledge Discovery*, 7(1), 81-99

Ministry of Health Kenya (2005) AIDS in Kenya, 7th edition. National AIDS and STI Control Programme (NASCOP), Nairobi

Mugenda O, Mugenda G (1999) Research Methods. Quantitative and Qualitative Approaches, African Centre for Technology Studies (ACTS) Municipal Council, Nairobi

NACC and NASCOP (2012) The Kenya AIDS epidemic update, National AIDS Control Council. Nairobi, Kenya

Otwombe K, Wanyungu J, Kilonzo N and Taegtmeycr M (2007) Improving national data collection systems from voluntary counselling and testing centres in Kenya. *Bull World Health Organ*, 85(4): 315–318.

Pipino L, Lee Y and Wang R (2002) Data Quality Assessment. *Communications of the ACM*, 45(4): 211-218.

Tayi K and Ballou P. (1998). Examining data quality. *Communications of the ACM*, 41(2), 54-57.

UNAIDS (2010) Report on the global AIDS epidemics. UNAIDS Geneva

US EPA (2006) Data Quality Assessment: A Reviewer's Guide (EPA QA/G-9R). EPA/240/B-06/002. February 2006

WHO (2003) Improving data quality; a guide for developing countries

WHO (2005) The international journal of public health; Improvements to data systems for the health sector

APPENDICES

Appendix 1: Consent form for Health care providers

Health Care Providers Consent Form for Study Participation

Use of Data Quality Assessment to Evaluate the Level of HIV Data Completeness in Health Facilities

Introduction and Purpose:

You have been asked to be a part of a research study. We are doing this project in HIV care and treatment clinics. We want to assess the level of HIV data completeness in HIV patient record (blue card) as well as to learn more about the perceptions behind missing HIV data entries and how health care providers handle them to improve on reporting system. We want to know whether health care providers know how to conduct HIV Data Quality Assessment (DQA) for completeness and how they would handle missing HIV data entries. The study is being conducted in lower Eastern province of Kenya in 42 Ministry of Health facilities.

Procedures:

If you choose to participate in the study, we will let you fill a questionnaire in regard to what you feel are the major causes of data incompleteness in your HIV care and treatment clinic and whether you know how to conduct a HIV DQA for completeness. Whether you agree to participate in the study or not will not affect your career in any way.

Risk/Discomforts:

There are no known risks to you for participating in this study. When completing the questionnaire, you may ask any question which you feel you don't understand.

Benefits:

There are no direct benefits to you for participating in this study. What we learn from this study will help us improve on handling of HIV data collection tools hence strengthening of reporting systems, HIV prevention, care and treatment feedbacks and services.

Confidentiality:

If you agree to participate in this study, we will not write down your name or personal information. Findings from the questionnaires and data charts will be analyzed together with others. The questionnaires will be secured in a locker at all the times. Only the principal investigator, research assistants and supervisors will be allowed to look at them. Your name or anything that might identify you will not appear in any write ups.

Right to Refuse:

It is your choice to participate in this study.

Contacts

If you have any question in relation to the study you can contact the principal investigator:

Kirimi Peter Mutugi, MSC Research Methods, JKUAT, Tel- 0787 176 445, [Email- gatuguti@yahoo.com](mailto:gatuguti@yahoo.com), mutugip@gmail.com

If you agree to participate please sign below.

Yes, I agree_____

Signature

Appendix 2: KNH/UON ERC Approval Letter



UNIVERSITY OF NAIROBI
COLLEGE OF HEALTH SCIENCES
P O BOX 19676 Code 00202
Telegrams: varsity
(254-020) 2726300 Ext 44355
Ref: KNH-ERC/A/271

KNH/UON-ERC
Email: uonknh_erc@uonbi.ac.ke
Website: www.uonbi.ac.ke
Link: www.uonbi.ac.ke/activities/KNHUoN



KENYATTA NATIONAL HOSPITAL
P O BOX 20723 Code 00202
Tel: 726300-9
Fax: 725272
Telegrams: MEDSUP, Nairobi
20th September 2012

Kirimi Peter Mutugi
AG332/1822/2010
JKUAT

Dear Mr. Kirimi

Research proposal: "Use of Data quality assessment to evaluate the magnitude of HIV Data Completeness in health facilities: A case study of Eastern province-Kenya" (P320/06/2012)

This is to inform you that the KNH/UoN-Ethics & Research Committee (KNH/UoN-ERC) has reviewed and **approved** your above revised proposal. The approval periods are 20th September 2012 to 19th September 2013.

This approval is subject to compliance with the following requirements:

- a) Only approved documents (informed consents, study instruments, advertising materials etc) will be used.
- b) All changes (amendments, deviations, violations etc) are submitted for review and approval by KNH/UoN ERC before implementation.
- c) Death and life threatening problems and severe adverse events (SAEs) or unexpected adverse events whether related or unrelated to the study must be reported to the KNH/UoN ERC within 72 hours of notification.
- d) Any changes, anticipated or otherwise that may increase the risks or affect safety or welfare of study participants and others or affect the integrity of the research must be reported to KNH/UoN ERC within 72 hours.
- e) Submission of a request for renewal of approval at least 60 days prior to expiry of the approval period. (*Attach a comprehensive progress report to support the renewal*).
- f) Clearance for export of biological specimens must be obtained from KNH/UoN-Ethics & Research Committee for each batch of shipment.
- g) Submission of an *executive summary* report within 90 days upon completion of the study
This information will form part of the data base that will be consulted in future when processing related research studies so as to minimize chances of study duplication and/or plagiarism.

For more details consult the KNH/UoN ERC website www.uonbi.ac.ke/activities/KNHUoN

"Protect to Discover"

Yours sincerely



PROF. A.N. GUANTAI
SECRETARY, KNH/UON-ERC

c.c.

The Deputy Director CS, KNH
The Principal, College of Health Sciences, UoN
The HOD, Records, KNH
Supervisors: Dr. Sandra Mudhune, Dr. Edward Mamati

"Protect to Discover"

Appendix 3: Questionnaire on HIV DQA for completeness competence and perceptions behind missing HIV data entries.

Number: _____

This questionnaire is meant to collect information that would be useful in HIV data quality improvement for quality prevention, care and treatment feedback in health facilities. You were arrived at through random sampling. Any information provided by you will be held private and confidential and would not be disclosed to any person. Please do not write your name or your employment number on this questionnaire. Analysis of the questions from the questionnaire will be done together with others to ensure anonymity. You are required to answer all the questions to the best of your knowledge. Thank you for willing to participate.

SECTION A

Personal Details

1. Sex: Male Female [Tick one]

2. Department of duty

[OPD]

[CCC]

[MCH]

[Tick all that apply]

3. Duration worked in the department (in months) _____

4. Have you ever trained in Health information Management?

[Yes]

[No]

[Tick one]

5. Which of the following issues contribute most to HIV data incompleteness in the Comprehensive Care Clinic (CCC)?

(a) Too many clients scheduled on a certain clinic day

(b) Too many unscheduled clients attending a clinic on a certain clinic day

(c) Too many data collection tools for documentation

(d) Don't know how to use HIV data collection tools

(e) Few data handling personnel's in the CCC

(f) Other(s) Specify

[Tick one]

6. Do you know how to conduct a HIV DQA for completeness?

a) Yes

b) No

Thank you for your time and participation!

Appendix 4: Chi square and logistic model syntax

GET

FILE='C:\Users\Reuben.ngumo\Desktop\peter\data.sav'.

value lab Gender 1'male' 2'female'.

value lab TRAININGIN 0'No' 1'Yes'.

value lab patientseen 1'One' 2'More than one'.

value lab Nifain2abo 0'No'1'Yes'.

value lab NDoyouknow 0'No'1'Yes'.

RECODE Months (0 thru 24=1) (25 thru 48=2) (49 thru Highest=3) INTO monthexperience.

VARIABLE LABELS monthexperience 'Months of experience'.

EXECUTE.

val lab monthexperience 1'1 Year'2'2 Year'3'3years and above'.

save outfile='C:\Users\Reuben.ngumo\Desktop\peter\data_edit.sav'.

CROSSTABS

/TABLES=Gender TRAININGIN monthexperience BY NDoyouknow

/FORMAT=AVALUE TABLES

/STATISTICS=CHISQ

/CELLS=COUNT

/COUNT ROUND CELL.

LOGISTIC REGRESSION VARIABLES NDoyouknow

/METHOD=ENTER Gender TRAININGIN monthexperience

/CONTRAST (Gender)=Indicator

/CONTRAST (TRAININGIN)=Indicator

/CONTRAST (monthexperience)=Indicator

/CRITERIA=PIN(.05) POUT(.10) ITERATE(20) CUT(.5).

Appendix 5: Participants Master list template for data entry

Serial No.	Questionnaire No.	Date Received	Date entered in data base	Comments
1				
2				
3				
4				
5				
6				
7				
8				
9				
10				
11				
12				
13				
14				
15				
16				
17				
18				
19				
20				
21				
22				
23				
24				
25				
26				
27				
28				
29				
30				
31				
32				
33				
34				
35				
36				
37				

Appendix 6: Ministry of Health facilities where HIV data completeness was evaluated

Facility	Total Complete entries				Total Evaluated entries in all patients visits			
	Weight	Pregnancy	TB Status	New OI	Weight	Pregnancy	TB Status	New OI
Masinga HC	36	36	36	36	36	36	36	36
D Comboni	273	187	270	254	275	231	275	275
Ekalakala HC	486	350	505	472	519	391	519	519
Ikombe	7	8	8	7	8	8	8	8
Kamuwongo	83	65	84	84	84	65	84	84
Kanyangi	23	13	21	14	24	14	24	20
Kathonzweni	68	64	67	67	68	68	68	68
Katulani	63	35	63	63	65	36	65	65
Kauwi	40	16	39	41	41	16	41	41
Kilala	35	35	31	39	41	41	41	41
Kisasi HC	48	26	51	44	53	39	53	53
Kithyoko HC	15	16	16	16	16	16	16	16
Kitise	110	112	112	111	112	112	112	112
Kitui DH	312	182	307	254	378	295	378	378
Kivaa HC	42	32	40	33	44	39	44	44
Kwan Vonza	24	4	26	26	26	4	26	26
Kyambeke	18	19	18	19	19	19	19	19
Kyuso DH	23	22	25	25	25	25	25	25
Makindu HC	79	57	73	69	81	81	81	81
Makueni DH	83	83	79	82	84	84	84	84
Masongaleni	26	31	31	29	36	36	36	36
Mathuki HC	240	241	235	237	242	242	242	242
Matiliku	90	95	95	95	92	95	95	95
Matinyani	111	66	107	95	112	68	112	112
Matuu DH	65	48	46	85	92	81	92	92
Mavindini	18	18	18	18	19	19	19	19
Mbitini HC	30	22	30	30	30	22	30	30
Miambani	84	45	71	81	93	46	93	93
Migwani	209	133	209	195	215	141	215	215
Mtito andei	117	103	119	119	118	103	119	119
Mukuyuni	32	32	32	32	32	32	32	32
Mtito wa ndo	39	26	45	34	45	26	45	45
Mwingi DH	146	88	141	134	154	95	154	154
Ngomeni HC	135	105	133	133	150	108	150	150

Nguni HC	64	41	69	66	69	41	69	69
Ngwata HC	148	157	137	138	159	159	159	159
Nuu SDH	9	12	12	9	16	16	16	16
Sultan Ham	43	43	45	43	48	44	48	48
Waita HC	32	32	29	30	38	35	38	38
Wingemi	29	26	30	24	35	26	35	35
Yatta HC	30	9	30	30	31	10	31	31
Average								
Total	87.0	66.7	86.2	83.2	93.3	74.8	93.4	93.2