Internal Migration Reconciliation for the Kintampo HDSS Core Datasets using Probabilistic Record Linkage Techniques

ROBERT ADDA AWIAH

STUDENT NUMBER: 888714

A Research Report submitted to the Faculty of Health Science in partial fulfillment of the requirements for the degree of *Master of Science (MSc)* in Epidemiology - Research Data Management

March, 2018

Robert Adda Awiah . *2017.*
Internal Migration Reconciliation for the Kintampo HDSS Core Datasets using Probabilistic Record Linkage Techniques.

SUPERVISORS:
Dr. Gideon Nimako, University of The Witwatersrand

SUPPORTED BY:
INDEPTH Network
University of the Witwatersrand
Kintampo HDSS

DECLARATION

I hereby declare that this research work is project work carried out by me under the guidance of Dr. Gideon Nimako. I have taken care in all respect to honour the intellectual property right and have acknowledged the contribution of others for using them in this work and further declare that the work reported in this project has not been submitted and will not be submitted, either in part or in full, for the award of any other degree or diploma in this institute or any other institute or university.

Parktown, Johannesburg, March, 2018

Robert Adda Awiah

DEDICATION

I dedicate this work to my loving mother whose principles of *hard-work-pays* has been my working tool. Also to my lovely wife Irene, for all the support and encouragement you gave to me throughout this academic work. Special dedication also goes to all the women who have affected my life; Madam Kanlaki my mother, Doris my sister, Irene my wife and Joselyn my daughter. You have been my best cheerleader.

ABSTRACT

Internal migration reconciliation involves the tracking of internal migrants to link their places of origin and destination for each movement within a given Health and Demographic Surveillance Area (HDSS) site. This involves several related activities adopted to account for the time the person moves from one location to the other but, remaining under the HDSS surveillance. This poses a major challenge for longitudinal studies particularly Kintampo HDSS site, where manual data capture modality is still being used. For most HDSS sites, all data related operations rely much on an individual unique personal identifier which is issued to identify a resident member during registration. The identifier enables the data system to keep track of resident members over an extended period of time to enable an accurate estimation of the population under surveillance. Movement of resident members within the geographical boundaries of the surveillance areas must be tracked based on these personal identifiers. However, residents' records may not be linked to each other in the event of multiple movements when the personal identifier cannot be recorded or is wrongly recorded. The effort in reconciling such cases of inconsistencies resulting from internal migration involves the printing of mismatched records for field supervisors to trace back to original locations to ascertain the identity of migrants. This process is very expensive as a full-time field supervisor is required and the reconciliation process is time demanding. In this project, we explored alternative automatic and cheaper method of reconciling all categories of discrepancies that exist in the internal migration datasets using the probabilistic record linking techniques.A theoretical foundation for probabilistic record linkage technology was provided and the sequential order was followed.The EM algorithm was used in the estimation of parameters. This research report demonstrate clearly that the probabilistic record linkage frame work by Fellegi and sunter is useful for HDSS internal migration reconciliation. The EM algorithm showed an improved performance in terms of linked records compare to the probabilistic framework. However, more work needs to done to explore other parameter estimation algorithm such as the Frequency-based EM algorithm. Such results can be compared to the results in this report.

ACKNOWLEDGMENTS

I would like to acknowledge with gratitude, INDEPTH Network, for the scholarship opportunity they provided me to pursue this Master's programme. I also acknowledge Dr. Seth Owusu-Agyei, the Director of the Kintampo Health Research Centre for the many opportunities he created in the career development of junior members of the institution such as me.

Many thanks go to Dr. Gideon Nimako my supervisor for his time in supervising this research and his advice during the write up of the research report. I am also grateful to Mrs. Busi Ngoyi, Mrs. Zodwa Ndlovu and all lecturers within the division of Epidemiology and Biostatistics at the Wits School of Public Health for their support and constant reminders of deadlines.

I am also grateful to my three course mates and recipients of the same scholarship from INDEPTH Network, Djibril Dion from Senegal, Kouliga Kombassere from Burkina Faso and Admas Abaerei from Ethiopia. Their company and friendship contributed to the completion of this work. To my wife Irene and children, thank you very much for your prayers and encouragement and the sacrifice you made for me throughout the period.

CONTENTS

ACRONYMS

HDSS Health and Demographic Surveillance System

KHDSS Kintampo Health and Demographic Surveillance System

KHRC Kintampo Health Research Centre

INDEPTH International Network for the Demographic Evaluation of Populations and Their Health

ICPD International Conference on Population and Development

DHS Demographic and Health Survey

PERMID Permanent Identification

CRF Case Report Form

PK Primary Key

CKI Community Key Informant

HRB Household Registration Book

SQL Structural Query Language

DSS Demographic Surveillance System

FEBRL Freely Extensible Biomedical Record Linkage

TAILOR Record Linkage Toolbox

HRS Household Registration System

LIST OF FIGURES

LIST OF TABLES

1

INTRODUCTION

In this chapter, we introduce the Health and Demographic Surveillance System (HDSS) and the data it generates for research purposes. The chapter also provides a background of record linking and the methodology used. We also give motivation and the objectives for this research project.

1.1 OPERATIONS OF HDSS

Health and Demographic surveillance Systems are usually established in areas or places where routine vital registration systems are non-existent or poorly developed, making them an alternative source that produces timely and reliable vital-registration datasets at the district or primary level [19]. They conduct research which monitors new health challenges in the surrounding communities for an extended period of time. The Kintampo Health and Demographic Surveillance System (KHDSS) is one such HDSSs, operating within the purview of the Kintampo Health Research Centre (KHRC). KHRC is one of three research centers established under the Ministry of Health of Ghana. The research centre was established at the country's major ecological zones in the early 1990s. It provides regular updates on vital events such as pregnancies, births, deaths, and migration and covers the whole of the Kintampo North Municipality and the Kintampo South district located in the Brong-Ahafo Region of Ghana [16]. The operations of a HDSS begin with a baseline census, followed by core activities involving monitoring of all entries into the dynamic resident population through births and in-migrations and all exits through deaths and out-migrations.

The contribution of HDSS to policy and practice has been very crucial in decision making [19]. International Network for the Demographic Evaluation of Populations and Their Health (INDEPTH) Network is an international network for the Demographic Surveillance System (DSS) sites in developing countries. The vision of INDEPTH is to harness the collective potential of the world's community-based longitudinal demographic surveillance initiatives in resource constrained countries [3]. This collective work provides a better, empirical understanding of health and social issues, and enables scientists to apply this understanding to alleviate the most severe health and social challenges. For the past decade, scientific research has been the focal activity of INDEPTH member HDSSs and this is expected to continue. Recently, INDEPTH Members Sites have strengthened their commitment to making their data available for comparative studies and cross site analysis. This has been demonstrated by a majority of the HDSSs making comparable individual-level data on mortality and cause-specific mortality available for analysis. This datasets are mainly from the populations under routine surveillance. However, weak vital registration systems remain a major challenge for HDSS sites in keeping track of the dynamics of this population under surveillance [3].

The operations of the HDSS constitute a set of field procedures that exist to handle the longitudinal follow-up of the well-defined population of its primary subjects which consist of individuals, households and residential structures and all related demographic and health outcomes within a clearly circumscribed geographic area [19]. HDSS exist to provide health information that more accurately reflect the prevailing disease burden of the population under surveillance. The existence of HDSSs also assists in the monitoring and tracking of new health threats within the geographical area. HDSSs operate to serve as new platforms for the study of new action-oriented research that will enable the testing and evaluation of health interventions.

The registration of migrations in the HDSS remains a key challenge since such events are considered as recurring. For the HDSS to continue to provide health information that accurately reflect its population, it is required that the population dynamics posed by mi-

grations in general and internal migration in particular, is resolved. Several efforts have been instituted in all HDSS sites to ensure that data collected regarding migration is of high quality and meets the standard required to accurately determine its projected population under surveillance. The United Nations International Conference on Population and Development (ICPD) in Cairo 1994 recognized the important role that HDSS plays in socio-economic development by recommending the full integration of population factors into development strategies [11]. It is therefore imperatively important that the populations under the surveillance of HDSSs are reported accurately.

Every HDSS was started with a baseline census followed by a well-defined regular update of key demographic events (birth, death and migration) and other health events. The figure below represents the operational concepts in the HDSS setup. Most HDSSs may also

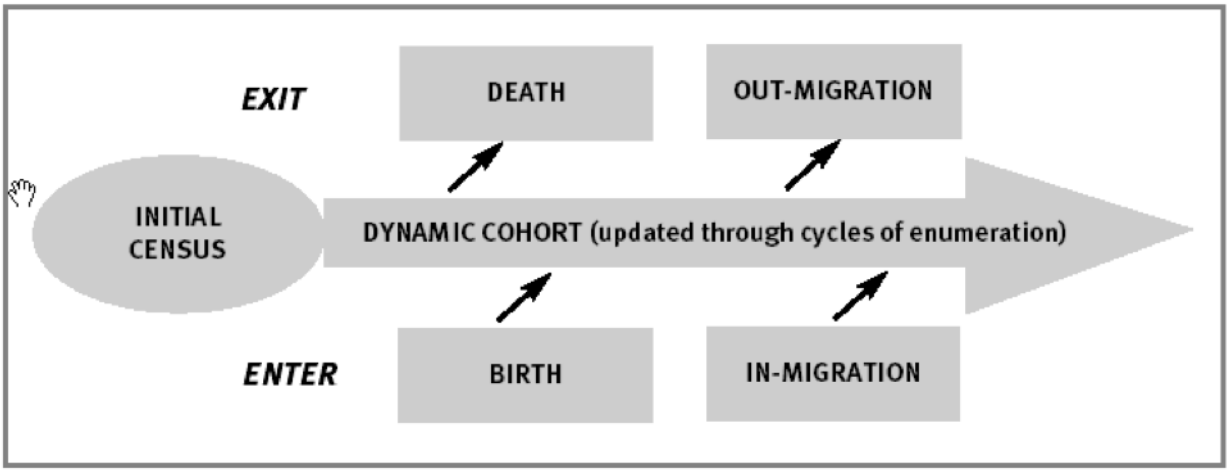


Figure 1: The Structure of HDSS Operations

include registration of marriages, divorces, changes in status and household relationships and fertility estimates. Such additional information is critical for a better understanding of the health and demographic dynamics of the population under observation. The operations of HDSS sites served to provide complementary and intermediate data to the other well-known methods of data collection such as national censuses, and Demographic and Health Survey (DHS)s.

To maintain the longitudinal integrity of data concerning individuals under routine surveillance, HDSSs established whether an individual moving into the HDSS study area

has previously been registered into the HDSS. Registration of migration has therefore been part of the routine operations of every HDSS. Migration is defined as the movement of people in a given time period across a specified boundary usually for the purpose of establishing a new residence [7]. The movement could be permanent, temporary or circular. The migration defining boundaries could also national borders or administrative boundaries within a country which is also known as internal migration whiles those that crosses national borders are classified as external migration. The Kintampo Health and Demographic Surveillance System (KHDSS) registers two forms of migrations, external migration and internal migration. In all HDSS sites, the boundaries of the surveillance areas are used as the primary migration defining boundaries. External migration is the case where an individual changed residence between two residential units one located within the HDSS and one outside the HDSS area. Internal migration on the other hand is a case where there is a residential change between two units within the HDSS area. Movement from one household to another within the study site has become very common among populations. At the HDSS level, an individual is registered in only one household at a point in time within the HDSS database system. On moving to another household, an internal migrant would be assigned a new identification number and hence would appear in the database twice.

To overcome this, internal migration reconciliation is implemented to ensure that every in-migration within the study area is matched with its counterpart out-migration. This involves the tracking of internal migrants to link the household of origin and destination for each episode of movement made, thereby ensuring a single unique identifier for each individual, and removing the potential of one person being concurrently registered in two households. Individuals retain this same unique identifier when they move and can therefore be readily followed up, strengthening studies of household and individual migration behavior, and improving follow-up of subjects enrolled in cohort studies or clinical trials.

Internal migration reconciliation constitute a set of procedures adopted by most HDSS sites to resolve discrepancies that existed in the internal migration datasets when the

unique permanent identifier of the migrants are not recorded or are misreported. This is done after the database failed to uniquely identify multiple episodes of internal migration records of a person. Several revisits are therefore required to be made by a migration field supervisor to collect more information on individual migrants to determine a possible matching of multiple datasets with discrepancies.

Internal migration reconciliation remained a major challenge to most HDSSs sites especially where manual data collection is still being practiced. The effort involved in reconciling such cases of inconsistencies resulting from internal migration involved the printing of mismatched record list for field supervisors to trace back to the original location to ascertain the identity of such migrants. The process is very expensive and time consuming as a full-time field supervisor is required and manual processes are followed.

1.2 AN OVERVIEW OF RECORD LINKAGE METHODOLOGIES

Record linkage can be thought of as the process of bringing information from two distinct sources together. It also has a number of other uses including building longitudinal profiles, de-duplication of individual records within a single database of records and case re-identification in capture-recapture studies. Probabilistic record linkage is a family of record linkage techniques that assigns similarity scores to pairs of records and treats all pairs that score above a certain thresholds as matches, non-matches and possible matches [11]. Many database programs often face the task of linking together records that do not share a common identifier, but nevertheless refer to the same entity.

In general, there are two main types of record linkage methods; the deterministic method and the probabilistic method. Deterministic record linkage is the process of linking information by a uniquely shared key(s). In this case, records are matched if the unique key field agrees or unmatched if the unique key field disagrees. This method has been used in most HDSS sites due to the availability of the unique or non-key field called

the Permanent Identification (PERMID). Probabilistic record linkage on the other hand attempts to link two pieces of information together using multiple, possibly non-unique, key fields. Despite the seemly simplicity of its description, the process has always been complicated especially by errors in the linkage key(s) or an apparent lack of unique key(s) linking both pieces of information.

In this work, we discussed record linkage in the context of linking data between two data files, although similar methods can be used to link more than two data files. Though accuracy has long been viewed as the cornerstone of any successful database, deciding which method to use to ensure precise automated data matching can be difficult [11]. In the HDSS data structure, *matching* refers to the process of establishing a link when two records seem to belong to the same individual. Probabilistic record linkage has been used to link HDSS surveillance datasets to external databases such as medical and hospital records [3]. This research however seeks to apply record linking technique to the internal migration datasets that have inherent inconsistencies with the aim of finding possible matches for records without permanent IDs. We utilize computational probabilistic algorithms to match data using quasi-identifiers in the datasets other than one personal unique identifier. Quasi-identifiers do not uniquely identify by themselves but, may in combination with other fields, uniquely identify an entity. For instance, name plus gender alone may not uniquely identify an individual due to individuals with the same name having the greater chance of having the same gender agreement. The addition of a date of birth or an address when combined with name and gender, will usually give a greater probability of identification.

1.3 PROBLEM STATEMENT

Discrepancies in HDSS datasets and the procedure involved in resolving those cases caused delays in the data processing cycle of HDSSs. Internal migration, which is one of

the major causes of these discrepancies, was tracked using either internal in-migration or internal out-migration Case Report Form (CRF) during regular visits. Internal in-migration is the situation where a study participant is reported to have been seen in the new place of residence and a census CRF is completed. In the case of an internal out-migration, the person is reported to have left the last residence to a new residence and a CRF is also completed to capture that event. Referential integrity constraint in relational database design theory is able to reconcile the two records of this individual using the unique identifier. If the database is unable to do the reconciliation, it implies the two records have different unique identifier or the identifier is missing in one case or both. The challenge lies in the effort involved in reconciling discrepancies in internal in-migration and internal out-migration as this task is time and resource consuming. This is an eminent data management challenge across almost all HDSS sites. Currently, it is estimated that over 3000 records have unresolved discrepancies due to internal migration at the Kintampo HDSS site and over 20000 records have one or two discrepancies as reported in the Agincourt HDSS datasets [11].

1.4 MOTIVATION

To resolve the challenge internal migration reconciliation poses to the quality of data generated by most HDSS sites for analysis and research activities, we apply automated techniques to the problem by employing probabilistic record linking techniques on the unresolved discrepancies recordsets. Record linking technique has been chosen because, it has been justified to provide a high successful link rate and tend to minimize the uncertainties in the matches between two datasets [3]. The technique has been used in many instances to link health records to other external databases for patient care and research purposes[11]. By employing such computational techniques/tools to resolve the discrepancies in the HDSS surveillance datasets, HDSS sites will be able to have the capacity

of providing near accurate census datasets that forms the basis for most cohort studies within HDSS sites.

1.5 CONTRIBUTIONS

The major contribution of this research report is the application of probabilistic record linkage algorithms on Kintampo HDSS dataset in an effort to reconcile discrepancies caused by internal migration. The main results being reported here include:

1. The implementation of an automated method of reconciling the discrepancies that exists in the HDSS internal migration datasets using the probabilistic record linking technique.

2. The evaluation of the probabilistic record linking technique used in the reconciliation of the discrepancies identified in the HDSS internal migration dataset and ascertaining the sensitivities of the matches.

3. The provision of recommendations and guidelines that allows the incorporation of automatic linkage algorithms in the surveillance data collection cycles of the Kintampo HDSS, particularly in cases of electronic data collection modality is utilized. This will allow reconciliation during data capture instead of post capture.

1.6 OUTLINE OF RESEARCH REPORT

The remainder of this research report is organized as follows. Chapter 2 presents a review of literature on HDSS operations and Record linkage. Chapter 3 presents an over-view of the HDSS Study area and the data management challenges posed by migration. In chapter 4, we give an introduction to probabilistic record linkage algorithms. Chapter 5 presents the experiments and discussion of results. We also discuss some challenges and

the importance of integrating automatic linkage algorithms in the surveillance data collection cycles. We summarize the work of this report in Chapter 6 and give some directions for future work.

2

BACKGROUND AND RELATED WORK

In this chapter, we give a review of existing literature relating to the operations of HDSS sites and the data they generate. We also give background to related work in record linkage methods and techniques.

2.1 HDSS SITES AND POPULATION RESEARCH

HDSS sites play key roles in monitoring populations under surveillance and has remained the basis for quantifying and understanding the complex demographic and health transitions in low- and middle-income country settings [7]. This is expected to continue if the Demographic and National Survey departments do not redesign their operations to tackle the challenges of population dynamics. The World Health Organization (WHO) has advocated that efforts should be made to strengthen HDSS sites to collect accurate and meaningful research datasets. The core mandate of any HDSS site is to register and monitor its dynamic population through routine collection and processing of information on births, deaths and migration. HDSS sites also operate complementary systems that allows for the collection of other datasets used to provide information on social and economic correlates of population and health. In other cases, HDSS sites provide health information that more accurately reflect the prevailing disease burden of the community. They in a way monitor and track new health threats. The aggregation of HDSS sites served as a platform for action oriented research to test and evaluate health interventions. This is illustrated in the increasing number of emerging HDSS sites in Africa and Asia and the

Oceania. Currently, there are about 52 INDEPTH HDSS Sites located across Africa, Asia, Central America and the Oceania [7].

The most challenging aspect of HDSS Sites is the ability to generation information that accurately reflects its population under routine surveillance. At the Agincourt HDSS site, rigorous and extensive analytic possibilities including application of automated measurement techniques are employed to analyze cause-of-death estimation by verbal autopsy and full reconciliation of in- and out-migrations, follow-up of migrants who depart the study area and recording of extra-household social connections [17].

2.2 HDSS DATA MANAGEMENT CHALLENGES

Every HDSS is an intensive longitudinal data collection and processing machine. Like any large population data source, HDSS is susceptible to errors and biases as it is common with most population data sources, especially those involved in the collection and compilation of large volumes of data. These errors and biases are likely to be substantial if internal migration reconciliation is ignored. Experience has shown that in a longitudinal surveillance data collection setting, small errors when not detected and corrected multiply over time [12].

The major challenge posed by internal migration reconciliation in population estimation has afforded the chance to exploit ways of finding opportunities to resolve it. Migration reconciliation has been a challenge for most HDSS sites especially where manual data collection is still practice. The opportunities provided by record linking techniques can also be exploited to link the records from multiple datasets to make it become useful in making meaning out of multiple collection of datasets [9]. The benefits of combining data from multiple sources has been explicitly demonstrated in earlier studies by Schumacher [18]. It has also been shown that the utilization of HDSS data sets would be more enhanced if linked to other data sources [3].

The concept of the HDSS data model was built around the relational database model and relied on the existence of unique personal identifier. The unique identifier issued to an individual resident serves as the Primary Key (PK). This makes it difficult for relations to be established with a second table if the unique personal identifier is not known. Since the introduction of record linking techniques, it has made it feasible and efficient to link large public health databases in statistically justifiable manner, making its importance in the field of health and research well documented [1]. It is shown in other studies that probabilistic record linking allowed the creation of a high quality linked database from crude registry data [13].

This research therefore translates the benefit of probabilistic record linking to one HDSS site where it has not been used to derive maximum benefits of the technique. The central theme of this research is the application of probabilistic linkage techniques to records that the relational model failed to match due to the unavailability of unique identifiers within the HDSS database. The existence of similar but dynamic data sets within the HDSS system provided the opportunity to exploit the potential of probabilistic record linkage techniques used in previous studies [1]. The deterministic record linking technique has been used in many instances within the HDSS databases where personal identifiers exist. However, in cases where the personal identifiers could not be found, it remains difficult for the system to link records between data sets. This makes most HDSS sites to rely on the clerical method where a field supervisor run around to collect further information to enable a possibility of searching within the database for a possible match of the records.

2.3 RECORD LINKAGE

The emergence of big data analytics as a way to generate new information and make better decisions has given rise to the use of modern methodologies in research data management. Continuous monitoring in HDSS generates longitudinal demographic data which

are health related and socio-economic indicators of each site's population. This result in a collection of multiple episodes of events stored within different data structures. Records within these structures are built on the relational model and are linked to each other by a primary key [15].

Record linkage has been used as a procedure of bringing together information from two or more records that are believed to belong to the same entity. In health research, big data represents a challenging problem due to the poor quality of data in some circumstances and the constant need to retrieve, aggregate, and process a huge amount of data from disparate databases in repeated successions [12].

Record linking is not new and has been used extensively in many areas especially health and research. The basic idea is to use a set of common attributes present in records from different data source in order to classify records. The linking of pairs of records without identifier is normally based on attributes both records have in common[12]. A record linking framework classifies pairs of records as links, non-links or possible links based on a comparison of the attributes found in both records.

Most recent applications developed for the purposes of record linkage are the Freely Extensible Biomedical Record Linkage (FEBRL) [5] and Record Linkage Toolbox (TAILOR) [8]. FEBRL implements the state-of-the art algorithms for record linkage and is one of the most current applications developed in medical research. It offers both researchers and programmers a fully flexible programing interface written in Python. TAILOR is also a record linkage toolbox intended to provide users an extensible and abstract framework to both develop and test record matching algorithms [5].

In this work, we looked at the approach to internal migration data processing and probabilistic record linkage of such datasets with discrepancies in order to produce very accurate data matching of exact individuals in the two datasets. These datasets are expected to be used in the calculations of rates in relation to the overall HDSS quarterly and yearly outputs.

This research used datasets that was unable to be analyzed due to inconsistencies caused by missing primary key field values. We have applied probabilistic record linkage technique to such record which normal would have pass through the manual migration reconciliation process. In an ideal scenario, if the migrant is identified in the two datasets by a personal identifier, the record is linked deterministically. In most cases, further effort was required to find the personal identifier to enable the establishment of a match. The utilization of probabilistic record linkage methodology in this research was therefore to serve as a second method of establishing a link in the two data sets when the personal identifier is missing in any one of the case.

3

HDSS MIGRATION AND SURVEILLANCE DATA QUALITY

In this chapter, we provide an introduction to the study site, its operations and how migration impacts the quality of research datasets.

3.1 THE KINTAMPO HDSS SITE

The Kintampo Health and Demographic Surveillance Site comprise the Kintampo North Municipality and Kintampo South District of the Brong-Ahafo Region of Ghana. It has a surface land area of $7,162$ km^2, which is 18.1% of the total land area of the region. It has a total population of approximately 152,000. The entire households in the study area have been geo-referenced, using Geographic Information System (GIS) and systematically allocated address codes on our data bases and painted on the walls of all compounds. Its strategic location makes it the geographical center of Ghana [16].

Kintampo Health Research Centre (KHRC) is the custodian of the KHDSS and it is one of INDEPTH Network affiliate HDSS sites located in the middle belt of Ghana, West Africa. HDSS sites are a network of research sites located in over 20 low and middle income countries (Africa, Asia and Oceania). Majority of these field sites are located in rural areas. Only Nairobi HDSS and Ouagadougou HDSS are urban sites. KHRC operated a partial HDSS since 1999 before moving to a full HDSS in 2003. Since then, it has established a geographically defined population under continuous demographic monitoring, with timely production of data on all births, deaths, and migrations. The monitoring system has also provided a platform for assessing a wide range of health-system, social and economic interventions, which are all closely associated with current research priorities.

3.2 HDSS FIELD OPERATIONS AND SURVEILLANCE

KHDSS covers the population residents in the Kintampo North municipality and Kintampo South district of the Brong Ahafo Region of Ghana. As part of its field operations, the KHDSS collects data manually and routinely updates the health and demographic information on the population and helps in selecting study participants and following them up in the community in the course of other cohort studies. It serves as the backbone and main research activity of work at the Kintampo Health Research Centre (KHRC). The resident population as at the last count in 2015 was 151,898. Currently, the KHDSS has 19 field staff: 11 fieldworkers, 6 field supervisors and 2 Research Officers. The number of KHDSS update rounds was reduced from three to two rounds per year from January 2014 and further reduced to one round a year from 2016. The current round of updates is round 26 and it covered the period from January to December 2016, with updates for pregnancies, births, deaths, and migration and data collection on verbal autopsies and marriages. Annual updates on education, household assets and socio-economic indicators such as employment were done in the year 2016.

The primary instrument used in collecting data within the KHDSS is a comprehensive annual update of resident status and vital events that involved every resident member who continued to be resident with the KHDSS [10]. A team of fieldworkers and supervisors make regular and schedule visits to each household following verbal consent and interview the most knowledgeable respondent. Every visit gives the fieldworker the opportunity to verify existing data and subsequently record any new event experienced by each household member, pregnancy outcomes, deaths and in-migrations and out-migrations. Field workers are supported by community leaders and members who act as Community Key Informant (CKI)s, recording all pregnancies, births and deaths on daily basis.

3.3 HDSS MIGRATION

Migration is defined in general as a change of residence. The KHDSS registers two forms of migrations, external migration, where the residence changes between a residential unit in the HDSS and one outside its geographic boundaries and internal migration, where residence changes from one residential unit to another within the HDSS geographic area. Recording internal migration is very important as it ensures the accuracy and validity of the KHDSS population datasets. Migrations influence the registration of births and deaths. A death, for example would not be recorded for an individual who out-migrated before his or her death. In- and out-migrations differ from internal migrations in the sense that the former refers to in and out movement of the HDSS area whilst the latter refers to moves within the HDSS area.

Migrations are considered as recurring events since an individual may make several migrations over time, both internally and externally. To maintain longitudinal integrity of data concerning individuals, KHDSS establishes whether an individual moving into the HDSS area from outside has previously been registered into the HDSS. The individual's current and previous records are matched so that he or she is not treated as a new individual in the system but as an individual under observation for several periods. New household occupants identified on the day of the staff visit are asked for the date of their arrival into the compound. Individuals who had not yet been resident for 3 calendar months were provisionally recorded on the back of the Household Registration Book (HRB) for follow-up at the next visit by which time they will be qualified for registration as new in-migrants and given permanent identification numbers. Previously registered residents who were reported as no longer residing in the house on the day of the visit were provisionally recorded as being out of the area, but not considered to have migrated out until they had been away for at least three consecutive calendar months.

3.4 INTERNAL MIGRATIONS RECONCILIATION

The issue of internal migration reconciliation deals only with resident members who move from one location to the other but, within the geographical area of the HDSS surveillance site. Movement of resident members outside the boundaries of HDSS sites is not considered in this case since the operations of HDSS sites is limited to the geographical boundaries of the surveillance area. Each movement of an internal migrant initiates two pair of records within the database. The movement of the migrant out of the old residence initiates a record called Internal Out-Migration record whiles the movement of the same person into a new residence initiates a record also called Internal In-Migration record. Internal is applied here because, the in and out movement is within the surveillance area. The migration database has been built on the expectations that, every internal out-migration is followed by an internal in-migration.

Internal migration reconciliation is therefore a data management procedure adopted by most HDSS sites especially where manual data collection is still being use for the tracking of internal migrants to link up their places of origin and destination for each movement within the surveillance site (geographical area). This process involves checking that the person who left the place of origin arrived at the place of destination using the same identification in the system. The motivation for this activity is primarily for data quality and assurance since the activity eliminates the possibility of having the same person registered in two places at the same time. This increases the accuracy of migration datasets because people can be followed when they move, which also improves the study of households' and individuals' migration behavior. It also facilitates accurate longitudinal studies, such as cohort studies or clinical trials, because it enables the follow up of individuals who move within the study site [3].

3.5 DATA MANAGEMENT

The Household Registration System (HRS) is a software system that implements the HDSS demographic core data and maintained a consistent record of significant demographic events that occur to the population in the surveillance area. It has the ability to generate HRBs which are used by field workers. It can also compute basic demographic rates such as birth rate, mortality rate, fertility rates migration rates and many more. The principal objective of the KHDSS is to compile a longitudinal database that will serve as a resource to assess the demographic and health impact of various health interventions. The KHDSS data management system is adopted from the data model used by all INDEPTH sites and it is held in the form of HRS2 (a second version of the old HRS) and was first hosted in Microsoft Visual Foxpro version 6.0 and later converted to Microsoft Visual Foxpro version 9.0. The attempts made over the years to upgrade the data management system to Microsoft Structural Query Language (SQL) Server have since remained unsuccessful. Overall, the data management system is still maintained in a form that ensured a more robust and stable environment for the rapid increase in both data volume and complexity.

The data reference model presented in this report is the general data model adopted by all HDSS sites. All data tables relates to each other by a PK. The primary table which is also referred to as the individual table contained records for every individual who resides or has ever resided in the study area. The physical location table records information on where individuals can stay. The position of a physical location can be described through the use of a co-ordinate or a series of co-ordinates, such as latitudes and longitudes. A Membership table records information on entry and exit from a particular household and there is also a table for each vital event category. The data stored in the various tables are connected logically by a common field name. Every data table has a primary key which uniquely identified each record in a given table. Every HDSS data is modeled to reference the manner in which individuals enter and exit the study area through a set of

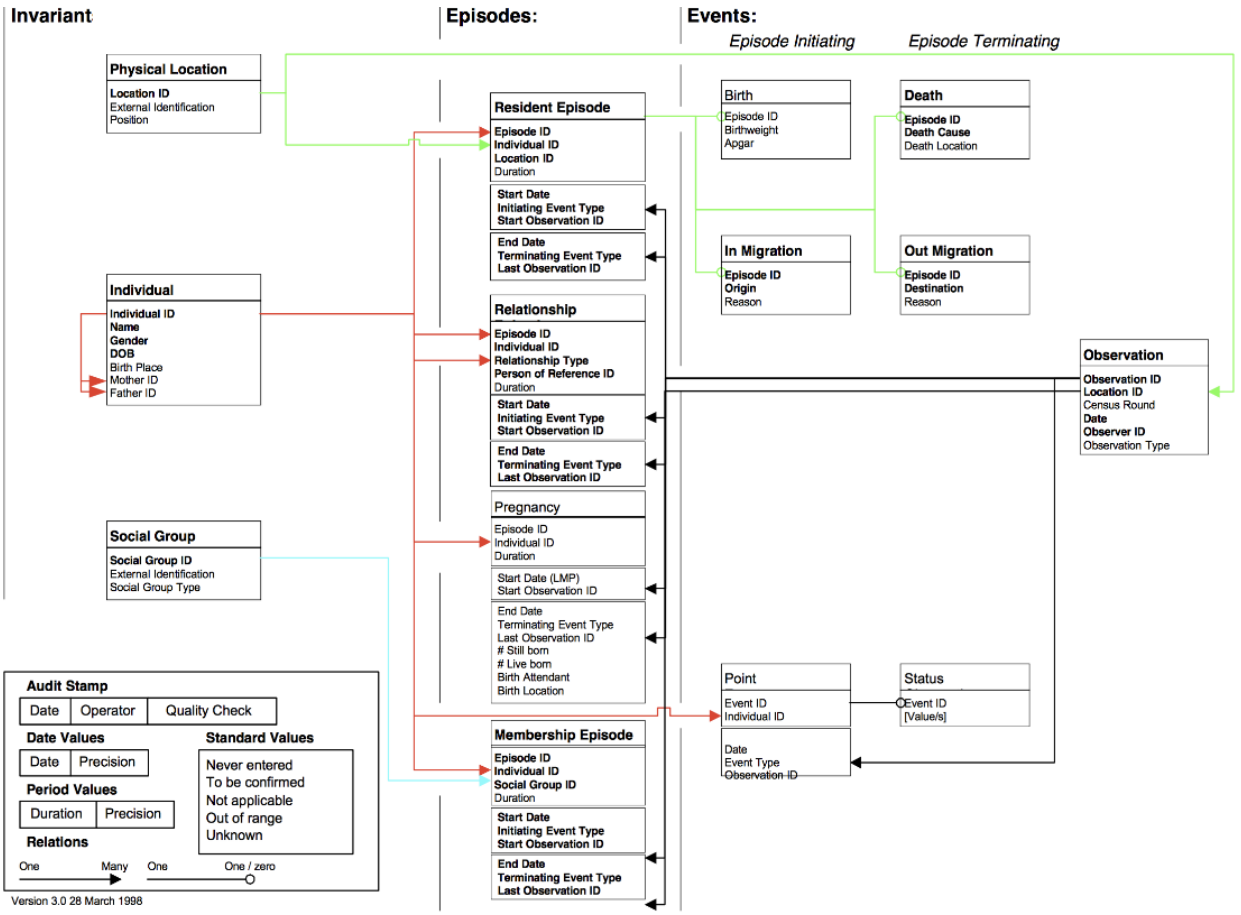


Figure 2: Reference HDSS Data Model (Extracted from [2])

well-defined set of rules and procedures. The data stored within the KHDSS data system are handled by a custom-designed relational database modeled to reflect the Reference data model designed specifically for longitudinal population data. It also encompasses a strong focus on the temporality nature of the data and facilitates linkage between episodes (periods of time in a given state) and events (an occurrence which opens or closes an episode). The core entity in the KHDSS data model is the individual, and each individual record is computer-assigned a unique identity number also referred to as a PERMID when first registered in the field by a visiting fieldworker. This provides the link between the Individuals table and the information that describes them. Each individual is thus linked to events and episodes, stored in other tables that together comprise their life history while under surveillance. Manipulation of this database allows calculation of person-time

at risk (exposure) and counts of various events; together this allows the calculation of probabilities and rates, and supports the range of other analyses.

3.5.1 *Data Quality Control*

Several examples exist to show how useful HDSS data, obtained routinely from the continuous monitoring of population and its registers is. It must also be emphasized, however, that effective analytical use of registration datasets depends, in large extend, the completeness and accuracy of the registers and the data so generated. The mechanism of continuous registration and monitoring means that errors will not only be cumulative but can involve serious biases. It is, therefore, desirable to institute measures to mitigate and reduce such problems. Quality of data collected from HDSS sites can influence various aspects of the healthcare and socio-economic system, hence the need for adequate attention to data quality assurance.

Quality control is therefore a mitigation measure, carried across all aspects of both field operations and data management system. It is being conducted at different levels with forms checked at three field levels – fieldworkers themselves on a daily basis, cross-checks by fellow team members on a weekly basis, and randomly checked by team supervisors. Errors are corrected in field offices otherwise a household revisit is conducted. CRFs are then submitted weekly to the main field office where a senior field coordinator carries out a final review with errors recorded and, where indicated, forms returned to the field for correction. Thereafter, all forms are log in to the filling office and received by a filling clerk indication form type and number of forms which is onwards passed on to the data room for data entry. At data entry level, programmed computer checks identify invalid codes, missing values, and inconsistent or duplicate entries; data that do not pass the pre-determined validation rules are blocked and an error message produced, resulting in the form being manually re-checked and returned to the field if unable to resolve.

A random sample of 5% work of each fieldworker area is generated and undertaken by a quality control supervisor. These allows for constructive feedback and an assessment of error rates. Information on quality is fed back to the site manager and team supervisors at weekly meetings. This work has been informed by the many challenges of data quality especially those imposed by internal migration reconciliation. It is expected to add to the efforts of HDSS data quality control by introducing automated reconciliation measures. The techniques and procedures adopted in this research work has been documented and explained in the succeeding chapter.

4

EXPECTATION-MAXIMIZATION ALGORITHM FOR UNSUPERVISED RECORD LINKAGE CLASSIFICATION

In this chapter, we provide the theoretical foundation for probabilistic record linkage. These theories feed into the simulations and experiments conducted in Chapter 5

4.1 RECORD LINKAGE WORKFLOW

Record linkage process usually follows a sequential order of steps. This is usually presented as a workflow [6]. Figure 3 provides the steps of this workflow. This workflow only illustrates a record linkage operation for matching two datasets. The flow can also be used for *deduplication* of a dataset. In such a case, the two datasets will represent the same dataset. The first step for a record linkage operation is to prepare the two datasets so that they are clean, standardised and in a comparable format. The next step is to pair all records using *Cartesian Product*. This implies paring each record in one dataset with every record in the other dataset. The resultant dataset can be very large. The aim of the indexing step is to exclude pairs of records that are not likely to correspond with the same entity. This reduces the number comparisons that would have been done on the raw results of the Cartesian product.

The next step is to compare the attributes of the record pairs. These comparison is based on the type of information contained in the variables being compared. The comparison can be done strictly (i.e., the they are either identical or not) or partially where the output results in partial agreement. These comparisons produce *comparison vectors* that serve as input for the classification process. The classification step is used to decide if a pair of

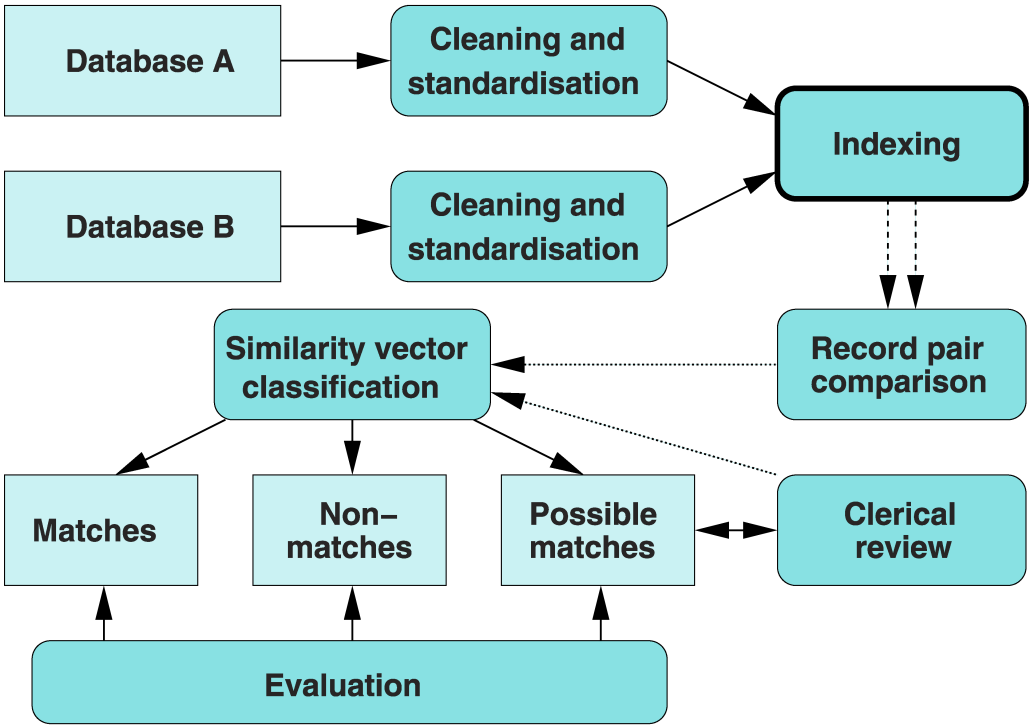


Figure 3: Record Linkage Workflow [6]

records belongs to the same entity or not. The results of this step can be any of these three sets; the set of links, the set of non-links and the set of possible links. Possible links are record pairs for which it is not clear if they belong to the same entity. This requires clerical review.

4.2 INTRODUCTION OF PROBABILISTIC RECORD LINKING THEORY

Suppose that we have two statistical populations $\mathcal{A}$ and $\mathcal{B}$, the probabilistic record linkage theory introduced by Fellegi and Sunter [21] assumes that both population are drawn from a simple random sample. The data files that correspond these two population are designated as A and B respectively.

A record $a \in A$ can represent the same entity as a record $b \in B$. In the comparison step, we pair a record $a \in A$ with a record $b \in B$ and decide if they belong to the same entity.

This results is set of record pairs $A \times B$. Each pair contains one record out of A and one record out of B. This is represented mathematically as:

$$(a, b) \in A \times B \tag{1}$$

The record pairs $A \times B$ is classified into two distinct sets of record pairs, a subset of $A \times B$ with pairs that represents the same entity and those that represents different entities. The first subset is called the *linked set* and the second subset is called the *non-linked set*. The record pairs in $A \times B$ are compared on selected of attributes/fields that are common to both files. Let $K \in \mathbb{Z}^+$ be the number of fields used for comparison. The comparisons are done by using a comparison function denoted by:

$$s : A \times B \to \Gamma \tag{2}$$

This maps the record pairs in $A \times B$ into the *comparison space* Γ. Every element $\mathbf{y} \in \Gamma$ is a K vector define by:

$$\mathbf{y} = (y^1, ..., y^K) \tag{3}$$

This is refereed to as *comparison vector* or the *agreement pattern*. It was assumed in [21] that the comparison vector $\mathbf{y} \in \Gamma$ is the realisation of a random variable

$$\mathbf{Y} = (Y^1, ..., Y^K) \tag{4}$$

We denote m as the probability of finding $\mathbf{y} \in \Gamma$ given that it is a true link. This is determined by the mass conditional probability function:

$$m(\mathbf{y}) = P(\mathbf{Y} = \mathbf{y} | M = 1) \tag{5}$$

where M is the link status. Similarly, we denote u as the probability of finding $\mathbf{y} \in \Gamma$ given that it is a non-link. This determined by the mass conditional probability function:

$$u(\mathbf{y}) = P(\mathbf{Y} = \mathbf{y}|M = 0) \tag{6}$$

4.2.1 *Conditional Independence Assumption*

One way to simplify the theories introduced in [21] is to assume that the elements of realization vector $\mathbf{Y}$ are mutually conditional independent given the true link status. This assumption is called the *conditional independence assumption*. This can be broken into two separate assumptions. The first is the that the elements of realization vector are mutually independent if a pair is a true link. The second is that, if a pair is a true non-link, the components of the comparison vector are mutually independent. Under these assumptions, the m-probability mass function in terms of marginal probabilities functions is:

$$m(\mathbf{y}) = m_1(y^1) \cdot m_2(y^2) \cdot ... \cdot m_K(y^K) \tag{7}$$

A similar notation can be deduce for u.

4.2.2 *Binary Assumption*

Comparing fields of record pairs can result in a vector of agreement, disagreement or partial agreement. Record linkage techniques make use of multiple levels of agreement. In many applications, only two level are used, namely *agreement* and *disagreement*, denoted as 1 and 0 respectively. This implies the elements comparison vector $\mathbf{y} = (y^1, ..., y^K)$ are restricted to $y^i \in 0, 1$ for $i \in 1, ..., K$. This results in a binary vector of length K. This is referred to as the *binary assumption*.

4.2.3 *Computing Weights*

The weight of record pairs denoted by $w(\mathbf{y})$ is computed using:

$$w(\mathbf{y}) = \log\left(\frac{m(\mathbf{y})}{u(\mathbf{y})}\right) = \log m(\mathbf{y}) - \log u(\mathbf{y}) \tag{8}$$

4.3 EXPECTATION-MAXIMIZATION ALGORITHM FOR PROBABILISTIC RECORD LINKAGE

The Expectation-Maximization algorithm starts with a set of observed comparison vectors $\mathbf{y}^1, ..., \mathbf{y}^K$. These are realisations of random variables $\mathbf{Y}^1, ..., \mathbf{Y}^K$. Every random variable $\mathbf{Y}_j$ corresponds a random variable M_j. M_j is the true link status of the record pair. The goal is to estimate parameters given the $\mathbf{y}^1, ..., \mathbf{y}^K$. These parameters are

$$\theta = (m, u, \pi) \tag{9}$$

The EM-algorithm is used to estimate these parameters. The algorithm is an iterative algorithm for which each iteration consists of two steps; an *Expectation* step and a *Maximization* step. In the Expectation step, the expected value of the log-likelihood is calculated based on the current estimates of parameters $\theta^{(t)}$ given the observed datasets $\mathbf{y}^1, ..., \mathbf{y}^K$. The superscript $t \in \mathbb{N}_0$ is an integer indicating the iteration number. The complete data likelihood for the Fellegi and Sunter model is given by:

$$\mathcal{L}(\theta; g_1, ..., g_N, \mathbf{y}_1, ..., \mathbf{y}_N,) = \prod_{j=1}^{N} P(\mathbf{Y}_j = \mathbf{y}_j, M_j = g_j) \tag{10}$$

Then, the complete data log- likelihood rewritten in summation is given by

$$\mathcal{L}(\theta; g_1, ..., g_N, \mathbf{y}_1, ..., \mathbf{y}_N,) = \sum_{j=1}^{N} g_j \log(\pi \cdot m(\mathbf{y}_j)) + (1 - g_j) \log((1 - \pi) \cdot u(\mathbf{y}_j)) \tag{11}$$

The derivation of Equation 11 from Equation 10 can be found in [14]. The complete data log-likelihood plays a vital role in the EM-algorithm. In the Maximization step, the conditional expectation of the complete data log likelihood, is maximized with respect to parameters θ. The conditional expectation of the complete data log likelihood, $Q(\theta|\theta^{(t)})$, is maximized with respect to parameters θ. The step is given by

$$\theta^{(t+1)} = \arg\max_{\theta} Q(\theta|\theta^{(t)}) \tag{12}$$

For each component θ_n in θ, the maximized parameter is given by

$$\theta_n^{(t+1)} = (\arg\max_{\theta} Q(\theta|\theta^{(t)}))_n \tag{13}$$

These steps are repeated until convergence. The algorithms returns the estimates of the m-probabilities, u-probabilities and π. The full details of this algorithm can be found in [14] and we have given an implementation codes in Python programming language at this Dropbox link[1].

1 https://www.dropbox.com/s/h4m9zc4r83vd9n8/estimation.py?dl=0

5

INTERNAL MIGRATION RECONCILIATION USING EXPECTATION-MAXIMIZATION LINKAGE

In this chapter, we provide simulation results and as well as the results of the Expectation-Maximisation algorithms applied to our datasets.

5.1 SIMULATIONS

The aim of these experiments is to explore the behaviour of the Expectation-Maximisation algorithms methods. Such experiments is needed as the *true* linkage status is not known for most record linkage applications. There are usually no complete data available to train the classifiers and to validate results outputs. The simulations are only for analysis and evaluation purposes and does not apply to our internal migration reconciliation analysis.

The datasets used in this simulation experiment was generated using the FEBRL [5] package and codes are also found at the FEBRL website. The types of datasets and the different comparison vectors are:

- **Good Data Quality**

  In this category, the comparison vectors represent the comparison of two good quality datasets. Good quality means the number of errors in the records is relatively small in both datasets. The parameters $m_1(1), ..., m_K(1)$ for such case are realisations of the uniform distribution $U(0.85, 0.99)$ whiles $u_1(1), ..., u_K(1)$ are realisations of the distribution $U(0.02, 0.5)$.

- **Low Data Quality**

  This is simulated in the same way as the simulation of the *good* dataset except the data is of low quality. This set of comparison vectors represents two datasets with a lot of errors. In this case, the $m_i(1)$-probabilities are realisations of $U(0.7, 0.85)$ and the $u_i(1)$-probabilities are realisations of $U(0.02, 0.5)$.

- **Poor Data Quality**

  The $m_i(1)$-probabilities are realisations of $U(0.6, 0.7)$ and the $u_i(1)$–probabilities are realisations of $U(0.02, 0.5)$

- **Skewed Data Quality**

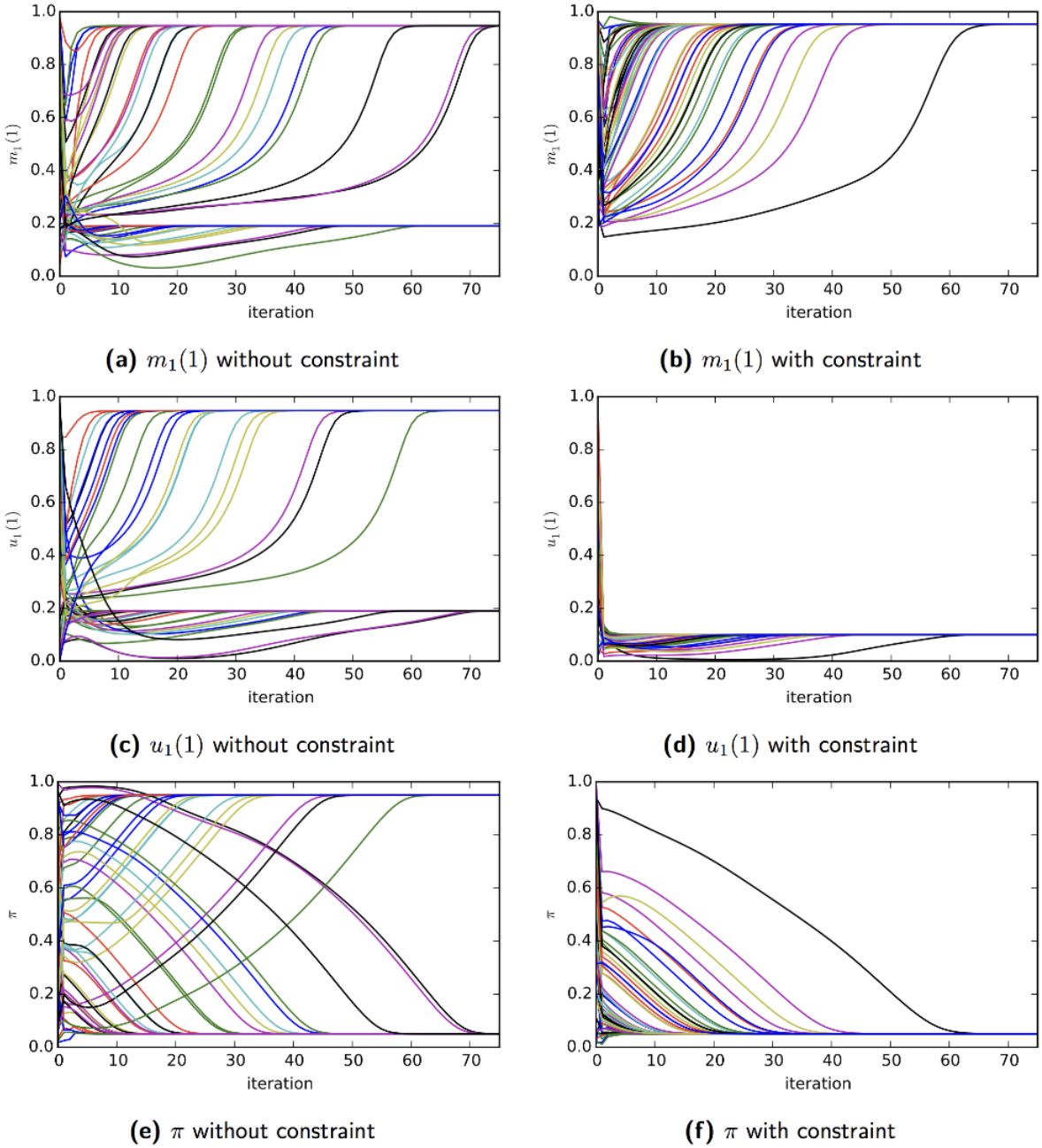
  The data in this category now has two comparison variables with high and low quality. Comparison variables Y^1 and Y^2 for a true link are realisations for which $m_i(1)$ is the realisation of $U(0.9, 0.99)$. The other vectors are simulated with m-probabilities drawn from $U(0.6, 0.8)$. For the true non-links, the $u_i(1)$-probabilities are realisations of $U(0.02, 0.5)$.

5.1.1 *Estimation Methods Simulation*

The EM-algorithm is used to estimate parameters of interest in the record linkage model when the data is considered to be conditional independent given the true link status (see Chapter 4). The algorithm converges to a stationary point. We first discuss the choice of starting parameters.

Suppose the simulated dataset of $N = 10^6$ comparison vectors for which 500 comparison vectors represent an identical recordset. Every vector is of length 8. This means that there are 17 parameters of interest in this model; $m_1(1), ..., m_i(8), u_i(1), ..., u_8(1)$ and the link prevalence π. They all needs a starting value to start the EM-algorithm. For this simulation, the EM-algorithm is applied 100 times to the same dataset using

variant starting values. The starting values are chosen randomly between 0 and 1. Figure 4 illustrates the parameters $m_1(1)$, $u_1(1)$ and π for each iteration and 100 random sets of starting points.



(a) $m_1(1)$ without constraint

(b) $m_1(1)$ with constraint

(c) $u_1(1)$ without constraint

(d) $u_1(1)$ with constraint

(e) π without constraint

(f) π with constraint

Figure 4: The Convergence Behaviour of the EM-Algorithm

The EM-algorithm converges in up to a maximum of 75 iterations. There are two stationary points to which each of the estimates of $m_1(1)$, $u_1(1)$ and π converges. The aim is to choose the starting values so that the algorithm converges to the desired parameter

estimates. As such the m-marginal probability mass functions for agreement should be large propensities. For true non-link set, there is relatively much less disagreement. The u-probabilities mass functions for agreement should be low probabilities. This indicates that an arbitrary choice of the starting values is not recommended.

Constraining the starting m- and u-marginal probabilities to $m > u$ prevents that estimates converge to wrong values. Using the same dataset in Figure 4, but now with constraints to the starting points, estimates converge to the same stationary point. Usually, a good starting value for the $m_i(1)$-probabilities is 0.9 and for the $m_i(0)$-probabilities 0.1.

Figure 4 indicates that the EM-algorithm convergences to a stationary point. We then studied the accuracy of the EM-algorithm. We considered 1000 datasets of type *good* with $N = 106$ comparison vectors and $K = 8$ comparison variables. We applied starting values of 0.9 for agreeing m-marginal probability mass functions, 0.1 for agreeing u-marginal probability mass functions and 0.01 for the link prevalence. The EM-algorithm is applied until the iteration converged. The results of the estimations and classifications are presented in Table 1. The average number of estimated links N_M, is 13728.704. This estimation is by far not close to the 500 true links in the set of comparison vectors. The 25% percentile gives reasonable results, but there are the algorithm did not work well in some situation. We repeat the same process for sets of $N = 104$ and $N = 105$ vectors with 500 true links. In the case of $N = 105$, the number of times the algorithm worked well is much larger. The 95% percentile is still a good estimate. The average error level λ is not accurate for $N = 105$ and $N = 106$ vectors. The m-probability mass functions are not correctly estimated hence the error level estimates of λ are incorrect. This effect of incorrect estimations fare described in [22].

N	Stats	I	II	III	N_M	μ_{exact}	μ_{sim}	λ_{exact}	λ_{sim}
10^6	mean	13439.95	566.52	985993.53	13728.70	$1.30e-02$	$1.22e-02$	$4.10e-02$	$5.08e-01$
	std	19479.48	915.35	20204.19	19855.86	$1.95e-02$	$1.83e-02$	$4.50e-02$	$3.40e-01$
	min	422.00	0.00	862162.00	460.00	$5.00e-06$	$9.85e-06$	$0.00e+00$	$1.71e-02$
	5%	477.00	2.00	944379.75	488.00	$2.50e-05$	$2.51e-05$	$0.00e+00$	$5.46e-02$
	25%	503.00	17.00	979553.25	510.00	$5.40e-05$	$5.13e-05$	$06.00e-03$	$1.07e-01$
	50%	4322.50	187.00	995433.50	4384.00	$3.83e-03$	$3.66e-03$	$1.60e-02$	$6.78e-01$
	75%	19715.50	767.00	999482.00	20311.25	$31.92e-02$	$1.83e-02$	$7.40e-02$	$8.30e-01$
	95%	52780.20	2293.00	999508.05	54003.75	$5.23e-02$	$4.93e-02$	$1.34e-01$	$8.67e-01$
	max	136577.00	8438.00	999532.00	136718.00	$1.36e-01$	$1.26e-01$	$1.74e-01$	$8.87e-01$
10^5	mean	616.89	18.18	99364.94	626.31	$1.60e-03$	$1.46e-03$	$7.50e-02$	$9.33e-02$
	std	853.17	32.20	874.85	862.30	$8.53e-03$	$7.43e-03$	$4.31e-02$	$1.02e-01$
	min	392.00	0.00	88016.00	446.00	$3.02e-05$	$2.36e-05$	$0.00e+00$	$7.13e-03$
	5%	464.90	0.00	99397.95	481.00	$1.01e-04$	$1.16e-04$	$1.60e-02$	$2.15e-02$
	25%	488.00	3.00	99484.00	494.00	$2.11e-04$	$2.19e-04$	$4.20e-02$	$4.42e-02$
	50%	9.00	99495.00	501.00	4384.00	$3.42e-04$	$3.47e-04$	$6.80e-02$	$6.95e-02$
	75%	505.00	17.00	99502.00	509.00	$5.33e-04$	$5.15e-04$	$1.00e-01$	$1.03e-01$
	95%	535.05	80.00	99514.00	560.20	$1.20e-03$	$1.14e-03$	$1.62e-01$	$2.13e-01$
	max	11696.00	288.00	99554.00	11885.00	$1.13e-01$	$9.65e-02$	$2.40e-01$	$7.01e-01$

Table 1: EM Classifications and Error Levels

5.2 INTERNAL MIGRATION RECONCILIATION

5.2.1 *Dataset*

In this section, we utilized the EM-algorithm to automate reconcile internal migration inconsistiencis within the Kintampo HDSS core DSS datasets. The records are linked for the years 2006 to 2014. We utilized the record linkage algorithm described in Chapter 4 for this reconciliation. Two datasets are used. The first dataset were extracted from the main DSS data repository. It has 62958 records. The second dataset are those records that have not being linked to the main recordset due internal migration mismatches. This dataset has 3079. As the underlying true link status between the two datasets is unknown, the unsupervised learning methods described in this research are needed for the estimation of important parameters. We first give descriptions and explanation of the various variables in these two datasets:

- **INDIVIDID** – A system generated individual ID based on village code and location ID
- **REASON** – The reason for which the individual is making a migration. This is a question asked by the interviewer
- **SOCIALGPID** - This is determined by the number of social groups in the household, it is issued by the interviewer to encompass the village code, compound number and the number of the household.
- **VILLAGE** - Every village within the study area has a name and a code. This code is the first two digits of every individual ID in the system
- **VILLAGE_NAME** - The name of the village
- **DATE** - This refers to the date of migration

- **SECTION** - Villages are made up of sections. This allows for the interviewer to narrow down his/her search for a migrant
- **LOCATIONID** - This also refers to as a compound number. This number is also specially generated by the enumeration team and it is a composition of village code and the number of the compound within a specified village
- **NAME** - Referred to as full name of the individual
- **GENDER** - Gender of the individual
- **BIRTH_DATE** – Date of birth of the individual
- **MOTHERID** – Permanent ID of the individual's mother if it is known
- **FATHERID** - Permanent ID of the individual's father if it is known
- **OLD_NAME** - If the individual has changed of names during the migration period, it is stated in this column.
- **COMPOUND_N** - Name of the compound the individual is migrating from/to if it is known

5.2.2 *Indexing/Blocking*

Previous research [5, 8] has shown that the link prevalence of the record linkage between two recordset is low if blocking is not used. The main DSS data file contains about 62958 records and the unmatched file contains 3079 records. This gives a Cartesian product of 193847682 record pairs. The computational complexity of this product is too high for the EM-algorithm. We experimented with the following blocking key(s) namely `VILLAGE_NAME`, `GENDER`, `MOTHERID`, `FATHERID` and `COMPOUND_N`. The key that gave optimal results is `GENDER`.

5.2.3 *Estimating Reconciliation Parameters using EM-Algorithm*

The main DSS data file is linked with the unmatched file using the EM-algorithm. This implies that the binary assumption and the conditional independence assumption of the EM-algorithm are applied. We observed that the EM-algorithm can result in a good classification in section 5.1. We label agreement and disagreement in comparison vector as 1 and 0 respectively. We used 5 comparison variables used the reconciliation. These variables are `VILLAGE_NAME`, `NAME`, `BIRTH_DATE`, `SECTION` and `COMPOUND_N`. These variables correspond to $\mathbf{y}^{vname}$, $\mathbf{y}^{name}$, $\mathbf{y}^{dob}$, $\mathbf{y}^{sec}$ and $\mathbf{y}^{cname}$ in Table 2 respectively. The 5 variables indicates that there are 11 parameters of used in the model, namely 5 m-marginal probability mass functions, 5 u-marginal probability mass functions and the link prevalence. We chose the starting values for these parameters such that convergences to the correct parameters. The choice of starting parameters is based efforts of trying to reconcile these two files. We made the assumption that the number of true links is much less than the number of true non-links.

Table 2 shows the comparison space for the classification. The the size of this space is $2^5 = 32$. The number of links for this classification was estimated on 1248 using the `recordlinkage.ECMClassifier()` function in the FEBRL package within the Python [20] programming language. However to obtain the specific m and u marginal probabilities, we implemented our own Python Classes. These codes can be downloaded at the following Dropbox link[1].

The comparison vectors in the green rows are positive/true links. The orange comparison vectors are possible links and the red comparison vectors are positive non-links. The possible link comparison vector requires clerical review to determine the exact link prevalence. The comparison vectors in Table 2 are sorted based on the weight. The comparison vector with the highest weight agrees on all 5 comparison variables. The next comparison vector is the comparison vectors that agrees on all comparison variables, except the com-

1 https://www.dropbox.com/s/h4m9zc4r83vd9n8/estimation.py?dl=0

$\mathbf{y}^{vname}$	$\mathbf{y}^{name}$	$\mathbf{y}^{dob}$	$\mathbf{y}^{sec}$	$\mathbf{y}^{cname}$	$f(\mathbf{y})$	$w(\mathbf{y})$	$m(\mathbf{u})$	$u(\mathbf{y})$
1	1	1	1	1	5458	16.40	4.72E − 01	4.41E − 03
0	1	1	1	1	2361	13.81	1.98E − 01	4.41E − 03
1	1	1	1	0	234	12.21	8.97E − 03	4.41E − 03
0	1	1	1	0	21	11.20	1.89E − 02	4.41E − 03
1	0	1	1	1	3671	10.35	1.27E − 02	4.41E − 03
0	0	1	1	1	1183	9.44	5.80E − 03	4.41E − 03
1	0	1	1	0	34	8.41	9.90E − 03	4.41E − 03
0	0	1	1	0	148231	7.54	7.33E − 03	4.41E − 03
1	1	1	0	1	873806	6.84	3.75E − 03	4.41E − 03
0	1	1	0	1	34	5.97	3.68E − 03	4.41E − 03
1	1	1	0	0	15153	5.22	2.00E − 01	5.15E − 03
0	1	1	0	0	34067	3.56	3.87E − 03	4.41E − 03
1	0	1	0	1	592078	3.11	3.64E − 03	4.41E − 03
0	0	1	0	1	102	2.84	3.61E − 03	4.41E − 03
1	0	1	0	0	3592278	2.62	8.49E − 02	9.37E − 03
0	0	1	0	0	903707	2.13	3.69E − 03	4.43E − 03
1	1	0	1	1	302671	1.55	5.84E − 03	5.08E − 03
0	1	0	1	1	41	−6.40	3.57E − 03	4.41E − 03
1	1	0	1	0	23761	−8.06	9.99E − 03	9.72E − 03
0	1	0	1	0	22	−9.72	7.38E − 03	1.19E − 02
1	0	0	1	1	1828	−11.38	4.51E − 03	8.90E − 03
0	0	0	1	1	975	−13.04	3.57E − 03	4.43E − 03
1	0	0	1	0	92678	−14.70	6.23E − 03	4.01E − 02
0	0	0	1	0	112	−16.36	5.15E − 03	5.50E − 02
1	1	0	0	1	21735	−18.02	3.64E − 03	9.23E − 03
0	1	0	0	1	277034	−19.68	3.61E − 03	1.12E − 02
1	1	0	0	0	31	−21.34	3.70E − 03	5.87E − 02
0	1	0	0	0	22673	−23.00	3.60E − 03	3.68E − 02
1	0	0	0	1	314438	−24.66	3.59E − 03	5.03E − 02
0	0	0	0	1	53443287	−26.32	3.62E − 03	3.69E − 01
1	0	0	0	0	17400338	-27.98	3.57E − 03	5.36E − 02
0	0	0	0	0	21040834	−29.64	3.57E − 03	3.35E − 01

Table 2: Kintampo HDSS Internal Migration Reconciliation with the EM-Algorithm

Year	N_I	N_{II}	N_{III}	N_μ	N_λ
2006	6258	14729	19533721	84	1135
2007	5804	13187	18426214	77	901
2008	5741	13889	17124912	75	791
2009	5297	12692	14498678	61	521
2010	3417	7755	8011523	28	680
2011	1249	3353	3283533	5	191
2012	1111	2755	2470399	4	192
2013	1550	2923	4201542	16	603
2014	1286	2651	3451097	13	471

Table 3: Kintampo HDSS : Results of Classifications with the EM-Algorithm

parison of village name. The vector for which every comparison disagrees has the lowest weight. It can be noted in Table 2 that all the comparison vectors in the positive link set agree on the date of birth. This variable is very important for the classification. The village is the least informative variable.

In Table 3 we show the classification results from 2007 to 2013. The table shows the number of record pairs with the positive link action N_I , the number of record pairs with the possible link action N_{II} and the number of records with the positive non-link action N_{III}. The trend over the years in the nearly the same.

5.3 INTEGRATING RECORD LINKAGE IN DSS

The DSS and HDSS around the world collect data about various aspects of the population, culture, economy and the environment in their respective country. DSS datasets are vital for governments and WHO to plan the allocation of funding and resources.

Data matching has been recognised as an important tool for DSS. The techniques discussed here can be incorporated into the DSS data management system and cycle. This will help reduce the costs and efforts required to reconcile inconsistencies that results from activities such internal migration. It also helps to improve data quality and integrity, as matching data from different census collections can help detect and correct conflicting or

missing information, or improve estimates of population sizes through other techniques such as capture–recapture.

Probabilistic record linkage can also be used to generate longitudinal datasets, by matching DSS data that have been collected at different instants in time. This will serve as an important source of information about how the characteristics of a population change over time. This requires buy-in from law makers as countries have different laws and regulations that govern what kind of data matching can be done. Such restrictions make it very challenging to create longitudinal datasets, because the matching has to rely on personalized information such as age, gender, birthplace, religion etc.

6

CONCLUSION AND FUTURE DIRECTIONS

Record linkage is widely used for many practises where data needs to be linked between multiple sources. This research report demonstrates that the probabilistic record linkage framework by Fellegi and Sunter [21] is useful for HDSS internal migration reconciliation. There are several methods to estimate parameters of the framework. The research applies the EM-algorithm because it is one of the effective algorithms to estimate parameters.

6.1 PROBABILISTIC RECORD LINKAGE

The techniques disccussed in Chapter 4 are effective frameworks for classification for record linkage. The simulations in Chapter 5 showed that the EM-Algorithm is useful to link records between dataset files. The number of misclassification can be estimated accurately. The number of comparison variables and the quality of the data. Records of good quality data result in better classifications than records of poor quality. Also, the number of comparison variables is related to the accuracy of the classification. In general, more comparison variables lead to a more accurate classification. Probabilistic record linkage employs comparing record attributes which is often performed under the binary assumption.

6.2 PARAMETER ESTIMATION USING EM-ALGORITHM

The EM-algorithm is, according to the literature, a good method to estimate parameters. The simulation experiments in Chapter 5 showed that this algorithm is a good method

to estimate m-, u-probabilities and the prevalence rate, π. The algorithm was able to estimate the number of links in the dataset and the error levels very accurately. For most of the starting parameters, the algorithm converges to the desired estimates. Choosing extremely bad starting values, the algorithm classified the links as non-links and the non-links as links. This can easily be observed from the estimation results. EM-algorithm has properties that can lead to incorrect classifications. The simulation experiments shows that π has to be larger than 0.01 to result in good estimates. If the link prevalence is less than 0.01, the estimates has potential to result in error.

6.3 DSS INTERNAL MIGRATION RECONCILIATION

In Chapter 5, we linked the core DSS of Kintampo HDSS with unmatched records (that results due to internal migration) from 2007 to 2013. The record linkage was performed with the EM-algorithm. The classification was made for each year between 2007 and 2013. The number of links shows a similar trend as for the record linkage. The ordering of the comparison vectors was well justifiable on the basis of manual review and data knowledge. Each classification had the same configuration of the comparison vector as a possible link. We used `GENDER` as the blocking key. We obtained an estimated 1248 positive links. This shows that the technique can be incorporated into the DSS data management system and cycle. This will help reduce the costs and efforts required to reconcile inconsistencies that results from activities such internal migration.

6.4 FUTURE DIRECTIONS

The following activities have been identified for future work to embed this algorithms and implementations into HDSS data life cycle and also increase its usability:

1. More work need to be done to explore other parameter estimation algorithms such as Frequency-based EM algorithms. Such results can be compared to the results presented in this report.

2. There is a need for a Graphical User Interface (GUI) for data clerks to perform clerical review on the possible links. This work has been earmarked for future work.

3. In [4], it was demonstrated that machine learning approaches such Support Vector Machines (SVM) achieve more accuracy than EM algorithms. Such comparison need further work and this has also been earmark for future work.

BIBLIOGRAPHY

[1] Frank Baiden, Abraham Hodgson, and Fred N Binka. "Demographic surveillance sites and emerging challenges in international health." In: *Bulletin of the World Health Organization* 84.3 (2006), pp. 163–163.

[2] Justus Benzler, Kobus Herbst, and Bruce MacLeod. *A data model for demographic surveillance systems*. 1998.

[3] International Development Research Centre (Canada) and INDEPTH Network. *Population and Health in Developing Countries. Volume 1, Population, Health and Survival at INDEPTH Sites*. International Development Research Centre, 2002.

[4] Peter Christen. "Automatic record linkage using seeded nearest neighbour and support vector machine classification." In: *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM. 2008, pp. 151–159.

[5] Peter Christen. "Febrl: a freely available record linkage system with a graphical user interface." In: *Proceedings of the second Australasian workshop on Health data and knowledge management-Volume 80*. Australian Computer Society, Inc. 2008, pp. 17–25.

[6] Peter Christen. *Data matching: concepts and techniques for record linkage, entity resolution, and duplicate detection*. Springer Science & Business Media, 2012.

[7] Mark Collinson. *The Dynamics of Migration, Health, and Livelihoods: INDEPTH Network Perspectives*. Ashgate Publishing, Ltd., 2009.

[8] Mohamed G Elfeky, Vassilios S Verykios, and Ahmed K Elmagarmid. "TAILOR: A record linkage toolbox." In: *Data Engineering, 2002. Proceedings. 18th International Conference on*. IEEE. 2002, pp. 17–28.

[9] MPA Glenn Wright. "Probabilistic Record Linkage in SAS®." In: ().

[10] Kobus Herbst, Sanjay Juvekar, Tathagata Bhattacharjee, Martin Bangha, Nidhi Patharia, Titus Tei, Brendan Gilbert, and Osman Sankoh. "The INDEPTH Data Repository An International Resource for Longitudinal Population and Health Data From Health and Demographic Surveillance Systems." In: *Journal of Empirical Research on Human Research Ethics* 10.3 (2015), pp. 324–333.

[11] Chodziwadziwa W Kabudula, Benjamin D Clark, Francesc Xavier Gómez-Olivé, Stephen Tollman, Jane Menken, and Georges Reniers. "The promise of record linkage for assessing the uptake of health services in resource constrained settings: a pilot study from South Africa." In: *BMC medical research methodology* 14.1 (2014), p. 71.

[12] Kathleen Kahn, Mark A Collinson, F Xavier Gómez-Olivé, Obed Mokoena, Rhian Twine, Paul Mee, Sulaimon A Afolabi, Benjamin D Clark, Chodziwadziwa W Kabudula, Audrey Khosa, et al. "Profile: Agincourt health and socio-demographic surveillance system." In: *International journal of epidemiology* 41.4 (2012), pp. 988–1001.

[13] Bing Li, Hude Quan, Andrew Fong, and Mingshan Lu. "Assessing record linkage between health care and Vital Statistics databases using deterministic methods." In: *BMC health services research* 6.1 (2006), p. 48.

[14] Geoffrey McLachlan and Thriyambakam Krishnan. *The EM algorithm and extensions*. Vol. 382. John Wiley & Sons, 2007.

[15] Greg D Mears, Wayne D Rosamond, Chad Lohmeier, Carol Murphy, Emily O'Brien, Andrew W Asimos, and Jane H Brice. "A link to improve stroke patient care: a successful linkage between a statewide emergency medical services data system and a stroke registry." In: *Academic Emergency Medicine* 17.12 (2010), pp. 1398–1404.

[16] Seth Owusu-Agyei, Obed Ernest A Nettey, Charles Zandoh, Abubakari Sulemana, Robert Adda, Seeba Amenga-Etego, and Cheikh Mbacke. "Demographic patterns and trends in Central Ghana: baseline indicators from the Kintampo Health and Demographic Surveillance System." In: *Global health action* 5 (2012).

[17] Osman Sankoh and Peter Byass. *The INDEPTH Network: filling vital gaps in global epidemiology*. 2012.

[18] Scott Schumacher. "Probabilistic versus deterministic data matching: making an accurate decision." In: *DM Direct* (2007).

[19] Ali SiÃ, Valerie R Louis, Adjima Gbangou, Olaf MÃ1/4ller, Louis Niamba, Gabriele Stieglbauer, YÃ Maurice, Bocar KouyatÃ, Rainer Sauerborn, and Heiko Becher. "The health and demographic surveillance system (HDSS) in Nouna, Burkina Faso, 1993-2007." In: *Global health action* 3 (2010).

[20] Guido Van Rossum et al. "Python Programming Language." In: *USENIX Annual Technical Conference*. Vol. 41. 2007, p. 36.

[21] William E Winkler. "Frequency-based matching in Fellegi-Sunter model of record linkage." In: *Bureau of the Census Statistical Research Division* 14 (2000).

[22] William E Yancey. *Improving EM algorithm estimates for record linkage parameters*. 2002.

A

PLAGIARISM DECLARATION

UNIVERSITY OF THE WITWATERSRAND, JOHANNESBURG | FACULTY OF HEALTH SCIENCES

PLAGIARISM DECLARATION TO BE SIGNED BY ALL HIGHER DEGREE STUDENTS

SENATE PLAGIARISM POLICY: APPENDIX ONE

I **ADDA, ROBERT AWIAH** (Student number:**888714**) am a student

registered for the degree of **MSc in EPIDEMIOLOGY-RDM** in the academic year **2017**

I hereby declare the following:

- I am aware that plagiarism (the use of someone else's work without their permission and/or without acknowledging the original source) is wrong.
- I confirm that the work submitted for assessment for the above degree is my own unaided work except where I have explicitly indicated otherwise.
- I have followed the required conventions in referencing the thoughts and ideas of others.
- I understand that the University of the Witwatersrand may take disciplinary action against me if there is a belief that this is not my own unaided work or that I have failed to acknowledge the source of the ideas or words in my writing.

Signature: Date: 28/03/2017

26/04/2015
1

Figure 5: Plagiarism Declaration

B

ETHICS CLEARANCE CERTIFICATE

R14/49 Mr Adda Robert Awiah

HUMAN RESEARCH ETHICS COMMITTEE (MEDICAL)

CLEARANCE CERTIFICATE NO. M151028

NAME: (Principal Investigator)	Mr Adda Robert Awiah
DEPARTMENT:	Public Health Kintampo Health Research Centre, Brong Ahafo Region of Ghana
PROJECT TITLE:	Internal Migration Reconciliation Using Record Linking Technique for the Kintampo HDSS- Ghana, using Surveillance Datasets from 2006 - 2010
DATE CONSIDERED:	30/10/2015
DECISION:	Approved unconditionally
CONDITIONS:	
SUPERVISOR:	Gideon Nimako
APPROVED BY:	Professor P Cleaton-Jones, Chairperson, HREC (Medical)
DATE OF APPROVAL:	20/11/2015

This clearance certificate is valid for 5 years from date of approval. Extension may be applied for.

DECLARATION OF INVESTIGATORS

To be completed in duplicate and **ONE COPY** returned to the Secretary in Room 10004, 10th floor, Senate House, University.
I/we fully understand the conditions under which I am/we are authorized to carry out the above-mentioned research and I/we undertake to ensure compliance with these conditions. Should any departure be contemplated, from the research protocol as approved, I/we undertake to resubmit the application to the Committee. **I agree to submit a yearly progress report**.

Principal Investigator Signature Date

PLEASE QUOTE THE PROTOCOL NUMBER IN ALL ENQUIRIES

Figure 6: Ethics Clearance Certificate