Point-of-contact interactive record linkage between demographic surveillance and health facilities to measure patterns of HIV service utilisation in Tanzania

CHRISTOPHER T. RENTSCH

Thesis submitted in accordance with the requirements for the degree of
Doctor of Philosophy
of the
University of London

AUGUST 2018

Department of Population Health
Faculty of Epidemiology & Population Health
LONDON SCHOOL OF HYGIENE & TROPICAL MEDICINE

Funded by: UK Economic and Social Research Council (ESRC),
Bill & Melinda Gates Foundation grants to the ALPHA Network, and
the Global Fund to fight AIDS, TB, and Malaria grants to the TAZAMA Project

Research group affiliations: Population Studies Group (LSHTM), and
TAZAMA Project (National Institute of Medical Research, Mwanza, Tanzania)

Declaration

I, Christopher T. Rentsch, confirm that the work presented in this thesis is my own. Where information has been derived from other sources, I confirm that this has been indicated in the thesis.

Signature: [redacted]

Abstract

As significant investments and efforts have been made to strengthen HIV prevention and care service provisions throughout sub-Saharan Africa, approaches to monitoring uptake of these services have grown in importance. Global HIV/AIDS organisations use routinely updated estimates of the UNAIDS 90-90-90 targets, which state by 2020, 90% of all people living with HIV (PLHIV) should be diagnosed, 90% of diagnosed PLHIV should be receiving treatment, and 90% of PLHIV receiving treatment should achieve viral suppression. Currently, estimates of these targets in sub-Saharan Africa use population-based demographic and HIV serological surveillance systems, which comprehensively measure vital events and HIV status but rely on self-reports of health service use. In contrast, most analyses of health service use are limited to patients already diagnosed and enrolled into clinical care and lack a population perspective.

This thesis aims to augment existing computer software towards a novel approach to record linkage – termed point-of-contact interactive record linkage (PIRL) – and produce an infrastructure of linked surveillance data and medical records from clinics located within a surveillance area in northwest Tanzania. The linked data are then used to investigate methodological and substantive research questions.

Paper A details the PIRL software that was used to collect the data for this thesis. Paper B reviews the data created by PIRL and reports record linkage statistics, including match percentages and attributes associated with (un)successful linkage. A subset of personal identifiers was found to drive the success of the probabilistic linkage algorithm, and PIRL was shown to outperform a fully automated linkage approach. Paper C provides original evidence measuring bias and precision in analyses of linked data with substantial linkage errors. Paper D critiques the estimation of the first 90-90-90 target and shows that current guidelines may underestimate the percentage diagnosed by a relative factor of between 10% and 20%. Finally, Paper E determines that while HIV serological surveillance has increased testing coverage, PLHIV who were diagnosed for HIV in a facility-based clinic were statistically significantly more likely to register for HIV care than those diagnosed at village-level temporary clinics during a surveillance round. Once individuals were in care, there was no evidence of any further delays to treatment initiation by testing modality.

The collective findings of this thesis demonstrate the feasibility of PIRL to link community and medical records and use the linked data to measure patterns of HIV service use in a population.

In loving memory of Basia

Who imparted a great deal of wisdom on
science, Man U, Force-based malaria cures, and wit

Acknowledgements

I will forever be indebted to my Rebel Leader, Basia Żaba, for taking me under her wing and guiding both my professional and personal development over the course of this PhD. This thesis is dedicated to her. I am grateful to Georges Reniers who picked up the baton and saw this thesis over the finish line after Basia's sudden departure to the stars. A big thanks to Ally Wringe, Milly Marston, and Jim Todd who graciously offered to be extra pairs of eyes on this thesis. I want to especially thank Rob Stephenson, who falls somewhere between Basia and me on a research supervisory family tree, for continuing to support my development and providing valuable feedback on this research.

This thesis would not have been possible to complete within an appropriate time frame without the technical support of my personal geek squad, Jason Catlett and David Beckles. I will fondly remember the intensive, sleepless week of coding to prepare the software for the field in a hotel lobby. Countless real-time, screen-sharing sessions to devise appropriate import/export procedures and identify sources of bugs (read: features) during roll-out will be repaid with Old Speckled Hen the next time our three paths cross. And for the last time: yes, I have tried turning it off and on again.

I would like to thank Mark Urassa and the National Institute for Medical Research in Mwanza for allowing me to work alongside their established research group. An enormous debt of gratitude to my fieldwork team, including data manager Richard Machemba, and fieldworkers Faustine Kibazira, Lameck Bikengela, Winny Swai, Rebecca Mushi, Liberatha Isack, Magreth Shoo, Moses Mahenda, and Jacqueline Hezron, for making this PhD research possible. I truly hope record linkage activities in Kisesa find further support to employ these dedicated, hardworking individuals. I also want to thank Frank Katunzi and the entire HDSS team for allowing the PIRL squad to crowd into their daily transport to Kisesa and for ensuring Masanja (as I was nicknamed) had a most authentic experience in Tanzania, from lunches at Mama Suzy's and 'one for the road' at Mama George's to a very special time at the Bugando Festival. "Yombaga kinehe!" Also a big thanks to Sue Kelly for providing a roof over my head and several candlelight suppers, for which she is locally famous.

I would like to give a most sincere "Ta!" to the Brits who have made my experience as an international student seem less foreign. To Chris Jarvis for allowing his spirit animal to most gleefully play with mine in the form of dancefloor takeovers, jails on Tatooine, and competitive wig nights. To 'Step-wedge' Jenny Thompson for embodying the example PhD student Jarvis and I attempted (and failed) to emulate. To my 'Fab Four' counterparts, Ashley, David, and Paul, for the laughs, counsel, and shared silence, respectively. I am particularly grateful to my partner Yomi who has helped me over many hurdles during the production of this thesis. And for, let's say, coincidentally getting assigned to work in Nashville for the final year of my PhD, which provided me an empty flat to create a distraction-free workplace.

Williams, Zimmer, Bowie, and Florence provided the soundtrack to this PhD.

"I may not have gone where I intended to go,
but I think I have ended up where I needed to be."
Douglas Adams

Table of contents

List of tables

List of figures

List of abbreviations

AIDS	acquired immunodeficiency syndrome
ANC	antenatal clinic
AQM	Advanced Quantitative Methods
ART	HIV antiretroviral therapy
BMC	Bugando Medical Centre
BSPS	British Society for Population Studies
CD4	cluster of differentiation 4
CRO	central registration office
CRVS	Civil Registration and Vital Statistics
CTC	HIV care and treatment centre
DHS	demographic and health survey
DMO	District Medical Officer
HDSS	health and demographic surveillance system
HTC	HIV testing and counselling
HIV	human immunodeficiency virus
ID	identifier
IPDLN	International Population Data Linkage Network
IUSSP	International Union for the Scientific Study of Population
IWHOD	International Workshop on HIV Observational Databases
LSHTM	London School of Hygiene and Tropical Medicine
MoHSW	Tanzanian Ministry of Health and Social Welfare
MRCC	Tanzanian Medical Research Coordinating Committee
NIMR	Tanzanian National Institute of Medical Research
PAA	Population Association of America
PEPFAR	United States President's Emergency Plan for AIDS Relief
PGDip	Postgraduate Diploma in Research Methods
PLHIV	people living with HIV
PMTCT	prevention of mother-to-child transmission
PSG	Population Studies Group
PPV	positive predictive value
TAZAMA	Tanzania AIDS Monitoring Activities
TCL	ten-cell leader (locally known as “balozi”)
UAPS	Union for African Population Studies
UI	user interface
UNAIDS	the Joint United Nations Programme on HIV/AIDS
WHO	World Health Organization

1 Introduction

1.1 Background

1.1.1 Uptake of HIV services

Over the past two decades, significant investments and efforts have been made to strengthen HIV prevention and care service provisions throughout sub-Saharan Africa, including universal access to HIV testing and antiretroviral treatment (ART). Whether and how individuals engage with HIV services is critical to achieving the Joint United Nations Programme on HIV/AIDS (UNAIDS) 90-90-90 targets,[1] which state by 2020, 90% of all people living with HIV (PLHIV) should be diagnosed (the 'first 90'), 90% of people diagnosed with HIV should be receiving ART (the 'second 90'), and 90% of people receiving ART should achieve sustained viral suppression (the 'third 90'). Removing the conditional nature of these targets would imply that 81% of all PLHIV should be receiving ART and 73% of all PLHIV should achieve sustained viral suppression. Notably, the 90-90-90 targets can be used to evaluate the performance of HIV programs within health systems and identify areas for interventions to improve health at the population level.[2, 3] While the 90-90-90 targets are ambitious, studies have shown numerous benefits emerging in populations that achieve them, including reductions in HIV incidence, morbidity, and mortality.[4-7]

Of the estimated 37 million PLHIV globally in 2016, nearly half were residents in eastern and southern Africa.[8, 9] Among them, it was estimated that 76% knew their HIV status, resulting in a gap of 2.7 million individuals requiring diagnostic testing to reach the 'first 90'. Further, among all PLHIV, 60% were on ART and only 50% were virally suppressed, which was short of the 73% target.[8] While overall improvements in access to HIV services remain necessary to achieve the targets, those who were aware of their HIV status had relatively more success in initiating and adhering to ART suggesting that a lack of knowledge of HIV status remains a key barrier to being linked to appropriate care in the region. Among all diagnosed PLHIV in eastern and southern Africa, 79% were on ART and 83% of those were virally suppressed.[8]

Estimates of uptake of HIV services in large geographical regions are likely to mask regional variations. A comparative study of community cohorts in several eastern and southern African countries found that the proportion of individuals who knew their HIV status ranged between 37% in a Tanzanian cohort to 93% in one in Malawi.[10] The study also demonstrated that the proportion of diagnosed individuals who were screened for ART eligibility within two years of diagnosis ranged from 14% in the Tanzanian cohort to

84% in one in Uganda. Variation between sub-national regions is likely due to community-level stigma, how and when HIV services are delivered, and other social and structural barriers specific to each area.[10-13]

As HIV services continue to expand throughout the region, approaches to monitoring uptake of these services have grown in importance. Most analyses of HIV service use are limited to patients already linked to care and lack a population perspective. In contrast, population-based health and demographic surveillance systems (HDSS) and HIV serological surveys (sero-surveys) comprehensively measure vital events and HIV status but rely on self-reports of HIV service use. Such reports usually lack detail and accuracy about an individual's clinical events and services received, and their retrospective nature means that they quickly become dated. Linking demographic and serological surveillance data with medical records from HIV service clinics located within the surveillance area would produce a nascent research infrastructure for generating directly observed data on access to and utilization of these services at the sub-national level.[14] The linked clinical data could also be used to validate or substitute the self-reported health status and HIV service use data collected in the surveys.

1.1.2 Record linkage

A recent Wellcome Trust report detailed how record linkage – the matching of an individual's records between two or more data sources – adds to the value of medical research in low- and middle-income as well as high-income countries.[15] Broadly, record linkage can increase the range of questions that can be asked, provide a historical perspective necessary for some studies, improve the statistical properties of analyses, and make better use of resources.

In the United Kingdom where unique patient identifiers are available, researchers have used record linkage to merge the Clinical Practice Research Datalink – one of the largest databases of longitudinal medical records from primary care in the world – to a variety of other existing data sources that hold information on cardiovascular and cancer events, hospitalisation, and mortality.[16] Publications using this data infrastructure have covered a vast range of topics, including high-impact studies showing the absence of an association between measles, mumps, and rubella (MMR) vaccine and autism,[17] increased cardiovascular risk after acute infection,[18] and the association between body mass index and cancer.[19] Other studies have used record linkage to compare and validate the information across multiple data sources.[20, 21]

Few record linkage studies are conducted in sub-Saharan African settings, primarily due to the need for data to be in an electronic format. Many HDSS sites, for example, are situated in rural areas, where clinics likely rely on paper records in the absence of computers. Time and resources would be required to digitise paper records for the purposes of record linkage. Second, record linkage requires a common set of matching identifiers in all data sources. It is possible that clinic databases do not share enough variables with community data in these settings, and that common identifiers that do exist may be of relatively poor data quality.

Traditional approaches

Two popular methods of record linkage have been established, deterministic and probabilistic, to combine data sources holding different information on the same individual. More modern methods, including machine learning techniques, exist,[22, 23] but these typically require a sizeable "training" dataset of gold standard linked records that is representative of the actual data to be matched, which is not available in most HDSS sites.

Deterministic record linkage is a rule-based approach that typically requires exact matching on a set of identifiers existing in all data sources.[24] For example, a unique national identification system, such as a National Insurance Number in the United Kingdom or Social Security Number in the United States, can be utilized as the key to merge two databases if each holds such information. Deterministic linkage could also be employed without the availability of a unique identification number. An algorithm based on exact matching between personal identifiers such as name, age, and address would also be considered deterministic.

However, record linkage often relies on a set of personal identifiers (e.g., names, date of birth, address) that are reported with error or are dynamic (e.g., name or residence changes). Probabilistic record linkage is a statistical approach that allows for such variation between records.[25-27] The statistical framework for probabilistic record linkage was largely developed in the 1950s[28] and 1960s.[29] This approach, which uses an algorithm to assign weights based on the (dis)similarity of identifiers used for linkage, has been shown to be superior to purely deterministic approaches in many settings.[30-33] Additionally, whereas deterministic approaches ignore the fact that some identifiers may contribute more than others to discriminate between true matches and true non-matches, probabilistic methods do not. For example, a rare surname that matches between two records would increase the likelihood the record-pair is a true match more than a match on sex, assuming there is a similar proportion of males and females in the population.

Typically, probabilistic linkage is largely an automated process once a set of matching identifiers and their agreement conditions are defined. After feeding the dataset(s) and linkage algorithm into computer software, match weights are calculated and summed for each record-pair. This summary measure, known as the match score, is classified into one of three categories: match, non-match, or a potential match (Figure 1.1). The thresholds chosen to classify a match score into one of these three categories are often trial and error and sensitivity analyses are suggested.[27, 34] Each record-pair that cannot be classified as either a match or non-match requires manual review to determine the match status, which can involves subjective decisions by the reviewer and usually necessitates a large effort especially in large, population-level datasets.[35, 36]

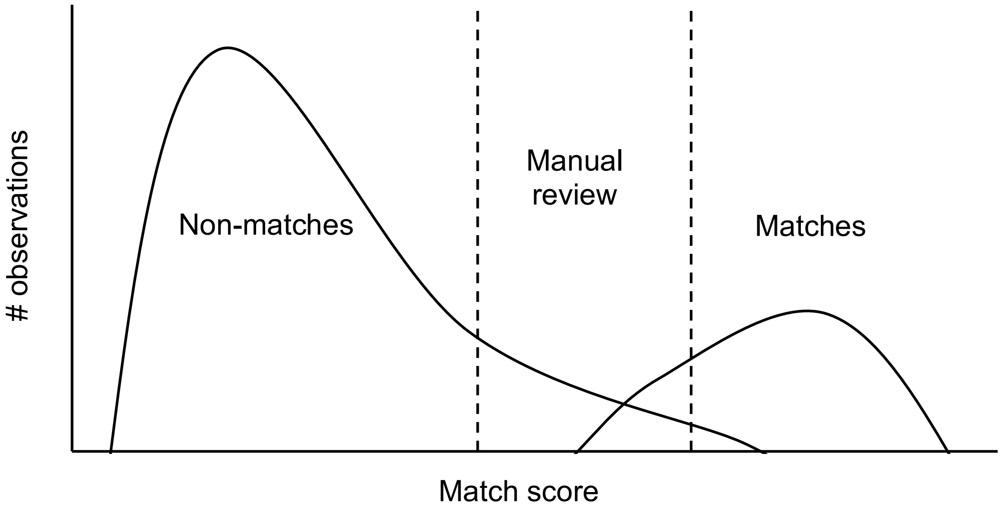


Figure 1.1: Categorisation of match scores (adapted from Jaro 1995)

The few studies that have performed record linkage in sub-Saharan Africa provided some evidence that linkage in resource-constrained settings was feasible. In Namibia, three databases – clinical, pharmaceutical, and laboratory – were retrospectively linked using patient name, sex, date of birth, and facility name; however, substantial missing data limited the success of the linkage to between 58% and 76% of records being matched.[37] In South Africa, a mix of deterministic (South Africa has a national identification number system) and probabilistic methods were employed to retrospectively link local health facility data to HDSS data with 88% of records being matched, suggesting linkage between these two data sources is achievable.[38]

There have been two previous attempts with varying levels of success to link clinic and community data using automated, probabilistic linkage in Kisesa, Tanzania. These attempts were possible because some residents of Kisesa have reported a clinic identifier during an HDSS or sero-survey round, thereby allowing for linkage between these data sources and a clinic database. These two previous studies used available clinic identifiers to build probabilistic algorithms to link individuals' records without available clinic identifiers. However, the limited number of individuals who reported clinic identifiers and poor data quality among those who had clinic identifiers available negatively impacted the success of these attempts.

In one study that linked HIV testing and counselling records to HDSS data, 388 true matches were identified using deterministic linkage of clinic identifiers captured in the HDSS and clinic data.[11] Using these matches as the gold standard, a probabilistic algorithm incorporating name, sex, year of birth, and village and sub-village of residence was developed. Weights for each matching variable were created using the Solver tool in Microsoft Excel and tested over many iterations of trial-and-error followed by manual review of linkage results. After calculating a match score for each record-pair, only one HDSS record with the highest match score was kept and was considered a match. The dataset was then sorted in descending match-score order and manual review was employed on every 100th record-pair. Based on the judgement of the researchers, all records below a match score that was deemed to obtain unlikely matches were dropped from analysis. The final linked dataset represented a linkage rate of 37% with poor sensitivity (18%) and positive predictive value (69%).

The second previous study to link data in Kisesa was between the antenatal clinic and HDSS data.[39] Using a similar approach described in the previous paragraph, 788 clinic identifiers captured in HDSS data were first used to deterministically link the data. Using these matches as the gold standard, a probabilistic algorithm incorporating two names, year of birth, village of residence, number of pregnancies, dates of pregnancies/births, and dates of residency in Kisesa was developed. Adjusted logistic regression models were used to identify weights for each matching variable, which more closely resembled the approach taken in the South African linkage study.[38, 39] After calculating a match score for each record-pair, only one HDSS record with the highest match score was kept and was considered a match. A match score threshold was identified by balancing sensitivity and specificity. The final linked dataset represented a linkage rate of 75% with moderate sensitivity (70%) and excellent positive predictive value (98%).

There are several aspects of the previous record linkage studies in Kisesa on which could be improved. First, both approaches relied on automated linkage approaches and multiple stages of manual review, which may be subject to bias. Second, the linkage algorithms were limited to variables that were common and of relatively high quality in both the clinic and HDSS data. Third, weights for each matching variable are typically created using an expectation maximisation algorithm on a large set of gold standard links. The approaches to create weights varied between the two Kisesa studies, but only used data on 388 and 788 record-pairs, respectively. Fourth, the HDSS record with the highest match score was automatically selected as the most likely match. Given these limitations, an approach that mitigates the need for subjective decisions when identifying matches, does not rely on limited sets of identifiers found in the clinic data, and includes a brief interaction with clinic attendees to adjudicate which HDSS records are theirs may be preferred.

Point-of-contact interactive record linkage

Most eastern and southern African countries do not benefit from having national identifiers, so linkage relies solely on other variables common in both data sources (e.g., name, age, or address). Additionally, data in these settings often suffer from data quality issues, such as incomplete records, spelling errors, and name and residence changes, all of which complicate both deterministic and probabilistic approaches to record linkage when applied retrospectively using automated processes.

In these settings, a semi-automatic record linkage process that incorporates manual, prospective inspection of potential matches, such as interactive record linkage,[40, 41] is preferred. In this thesis' implementation of interactive record linkage, termed point-of-contact interactive record linkage (PIRL), the manual review of potential matches identified by a probabilistic linkage algorithm is conducted in the presence of the individual whose records are being linked. This approach to record linkage has the advantage that any uncertainty surrounding a participant's identity can be resolved during a brief interview, whereby extraneous information (e.g., household membership) can be invoked as an additional criterion to adjudicate between multiple potential matches.

In addition, ethical and privacy concerns are properly addressed with PIRL as it offers an opportunity to seek informed consent, so individuals can be made fully aware of how their data are being used. Since record linkage algorithms often require personally identifiable information, such as names, birth dates, and residence details, protection of data in these studies has been a large concern for ethical reviewers.[42-44] Pursuing informed consent to perform retrospective linkage would be impractical for large-scale datasets and impossible for those individuals who are unavailable to give consent (those who died or migrated out of the area). These issues are non-existent in PIRL since individuals must give informed consent prior to having their records linked. More details on PIRL are provided in Chapter 2.

Of note, this novel approach to record linkage was originally termed “real-time record linkage” at the beginning of this PhD research. However, I received independent and constructive feedback from audience members at research conferences and reviewers of my first publication that “real-time” did not provide the best description of this approach. “Real-time record linkage” was initially used because the linkage was done in the presence of the individual whose records were being matched, who was able to steer the process, rejecting superficially plausible matches, and crucially was able to confirm if they had been a member of more than one household, which contrasted with the more conventional approach whereby record linkage was done retrospectively and focused on identifying a single “best” match. Upon further reflection, “real-time” is used in the software industry to signify processes that amend databases instantly (i.e. in real-time), whereas the system used in this PhD research did not depend on internet connectivity for links to be made immediately available on all machines. The updating of linked records was done as a batch process at the location where the central database was held when interviews were completed in a given day. What distinguished this approach was the interaction with the individual who provided critical input as to the veracity of the links identified using the probabilistic linkage algorithm embedded in the PIRL software. The approach was therefore renamed “point-of-contact interactive record linkage”, which better reflects the nuance and inter-personal nature of this methodology. Some conference presentations given prior to this switch will have used the term “real-time record linkage” (Appendix 10.1).

1.2 RATIONALE

The overall rationale for this thesis follows from the need for improved monitoring of HIV service uptake. While the majority of all PLHIV reside in eastern and southern Africa, and more broadly in sub-Saharan Africa, there remains a lack of directly-observed, population-based data needed to monitor progression toward the UNAIDS 90-90-90 targets across the region. Developing a locally-relevant approach to linking demographic and serological surveillance data with medical records from clinics located within the surveillance area would enable researchers to better track a population's progress towards meeting the 90-90-90 targets.

The data infrastructure produced by PIRL has the potential to become an invaluable resource for monitoring patterns of HIV service uptake in the community in which this thesis was conducted, and in similar settings if replicated. However, PIRL requires time and resources that may not be readily available in other HDSS sites. Using the links made by PIRL as the reference standard, it will be possible to compare the PIRL approach against an automated approach to record linkage that typically requires substantially less resources. In addition, secondary analyses using imperfectly matched data have been shown to be affected by errors made during linkage (e.g., false or missed matches).[45, 46] However, most of these analyses were conducted in settings with very low rates of linkage errors, such as North America and Europe, which often have unique national identifiers and excellent data quality.[47] The linked database created by PIRL will allow for the first known attempt to evaluate the impact of linkage errors on subsequent analyses in a setting with substantial linkage errors. The findings from these methodological analyses will provide an essential first step toward informing other researchers in similar settings desiring to perform record linkage whether to allocate resources toward automated linkage or invest in a PIRL system.

The linked data infrastructure produced by PIRL will also allow for more substantive research on patterns of HIV service uptake. Following an HIV-positive diagnosis, individuals are advised to register for care and initiate ART in an HIV care and treatment centre. Both late presentation into care and delayed ART initiation are associated with HIV-related morbidity and mortality[48-59] and increase the risk of further HIV transmission.[60-64] In order to expand access to HIV testing and increase linkage with care and treatment services, the World Health Organisation (WHO) recommends community-based HIV testing and counselling (HTC) with facilitated linkage to care services (for example, a lay-counsellor follow-up to encourage a clinic visit) in addition to traditional, facility-based HTC.[65] Community-based HTC includes services that are delivered using mobile and home-based approaches thus removing structural, logistical, and social barriers to HTC.[66] While population-based HIV sero-surveys – a unique form of community-based HTC – have increased testing coverage in high HIV burden areas like sub-Saharan Africa,[12] it is not yet clear whether this form of HTC results in higher or faster rates of linkage to care than other forms of community- or facility-based HTC. The proposed linked data infrastructure between serological surveillance and HIV testing and treatment clinic records will facilitate a comparison of time from diagnosis to treatment by testing modality. Identifying characteristics of individuals associated with better linkage to care will help inform future interventions that aim to reduce the attrition between diagnosis and treatment and to meet the second of the 90-90-90 targets.

Global HIV/AIDS programmes and organisations, including UNAIDS, WHO, and the United States President's Emergency Plan for AIDS Relief (PEPFAR), use routinely updated estimates of the 90-90-90 targets to measure progress of HIV service uptake around the world. Any bias in these estimates has the potential to mislead organisations on where gaps exist in HIV care and treatment programmes. Currently, UNAIDS rely on self-reported measures from sero-surveys to estimate the proportion of PLHIV who are diagnosed (the ‘first 90’). However, the ‘first 90’ may be underestimated due to respondents not disclosing their HIV testing history during the survey. The proposed linked data infrastructure will allow for the validation of self-reported HIV testing history against directly-observed HIV testing records from a local HTC facility. Furthermore, understanding characteristics of individuals who misreport their HIV testing history will serve as a critical first step in guiding UNAIDS and other stakeholders to better estimate a widely used target that assists programmes and organisations to track progress and prioritise further programme implementation.[67]

1.3 AIMS AND OBJECTIVES

The primary aim of this PhD research is to augment existing computer software towards a novel approach to record linkage in a rural community in northwest Tanzania (Kisesa) and use the emerging data infrastructure to measure patterns of HIV service utilisation. This thesis also aims to provide evidence toward whether the linked data source allows for analyses that would promote the continuation and expansion of PIRL within Kisesa, other HDSS sites, and beyond.

The objectives are:

1. To implement a locally-relevant approach to link community cohort data with medical records from three separate health facilities offering HIV services.

2. To identify individual characteristics associated with successful linkage using PIRL and compare PIRL with automated probabilistic record linkage.

3. To measure patterns of HIV service utilisation using the linked data infrastructure created by PIRL.

Objectives 1 and 2 form the methodological component and Objective 3 forms the substantive component of this PhD thesis. The methodological analyses will describe PIRL, compare PIRL to less resource-intensive alternatives, and ultimately recommend whether researchers should allocate resources toward automated linkage or invest in a PIRL system. The substantive analyses will demonstrate the utility of the linked data infrastructure created by PIRL by investigating how PLHIV progress through HIV services from diagnosis to care. This multifaceted approach, allowing for the synthesis of findings from both methodological and substantive research, will be essential to provide recommendations for improving measurement of health services use in this community and other rural African settings more broadly.

1.4 RESEARCH SETTING

The network for analysing longitudinal population-based data on HIV in Africa (ALPHA) consists of ten population-based HDSS sites in six eastern and southern African countries. This PhD research was conducted at the ALPHA Network site in Kisesa, Tanzania.

1.4.1 Study area

Since 1994, the TAZAMA (Tanzania AIDS Monitoring Activities) project based at the National Institute of Medical Research (NIMR) in Mwanza, Tanzania has been monitoring the HIV epidemic in a population of approximately 35,000 residents in Kisesa, a rural ward located in the Magu district of Mwanza region in northwest Tanzania (Figure 1.2). The study population is the Kisesa open cohort study – the longest running HIV community-cohort study in Tanzania, and one of the oldest in sub-Saharan Africa – which includes multiple rounds of HDSS and sero-surveys (described further in Section 1.4.2). The study area is situated 20 kilometres east of Mwanza city along the main road leading to Kenya. It comprises seven villages, one in which has a roadside trading centre and the others that are located in more rural areas.

Approximately half (51%) of Kisesa residents in 2016 were female. The age structure of the surveillance population suggests high fertility and high mortality.[68] About half (48%) of Kisesa residents in 2016 were aged ≤15 years, 23% were aged 15-29 years, 19% between 30-49 years, and the remaining 10% were aged ≥50 years. Among residents aged ≥15 years in 2016, 36% reported having never been married, 45% reported having one marriage, 5% reported being in a subsequent marriage, and the remaining 13% reported being separated or widowed. There are 15 primary schools and three secondary schools in the area. Among residents aged between 5 and 25 years in 2016, 58%

reported being currently in school; of these, 67% reported being in primary school and 17% reported being in secondary school.

Farming and trading were the main sources of income (97%) and commonly grown crops are cotton, cassava, paddy, maize, and sweet potatoes.[68] The majority of Kisesa residents are from the Sukuma tribe. The dominant religion is Christianity with a smaller proportion belonging to traditional religions (10%) and Islam (2%). GBP per capita in Mwanza region doubled from $500 USD in 2010 to around $1000 in 2016,[69] but is likely to be lower in Kisesa as Mwanza region includes Mwanza city. Data on in- or out-migration rates in Kisesa ward are limited. Using data from the 2016 HDSS survey, 12% of Kisesa residents were new to the surveillance area at that round. Previous studies have shown the primary reasons for migration were marriage, household migration, employment, returning home, and schooling.[68]

A government-run health centre is located within the trading centre in Kisesa village, offering a wide spectrum of services, including an HIV care and treatment centre (CTC, since 2008) that provides ART free of charge to all PLHIV,[70] an HIV testing and counselling clinic (HTC, operating since 2005), and an antenatal clinic (ANC) that offers opt-out HIV testing and prevention of mother-to-child transmission services (PMTCT, full package of services since 2008) as part of its routine care.

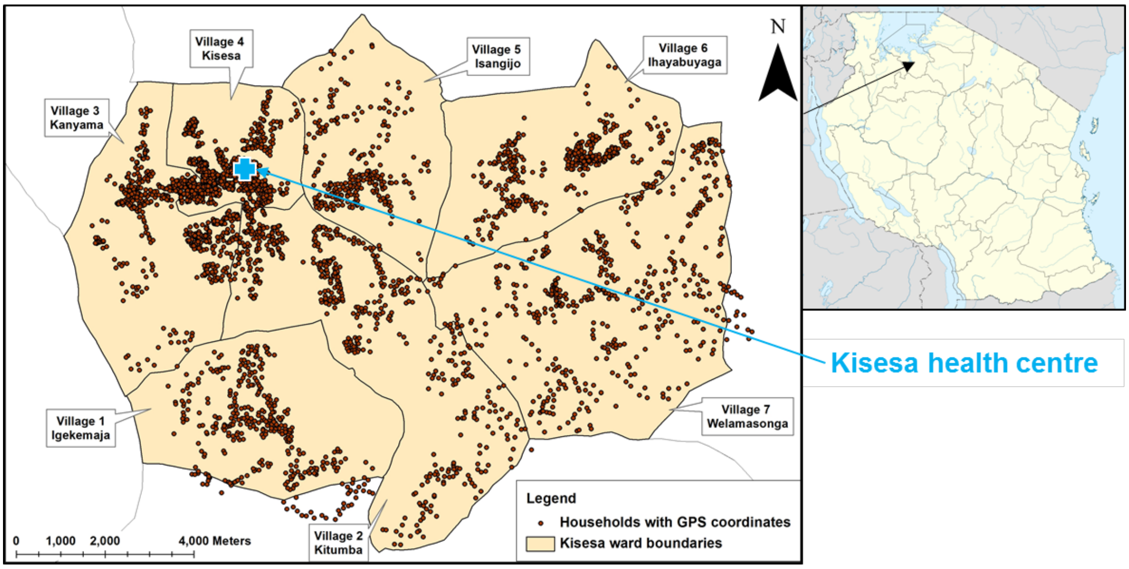


Figure 1.2: Map of study area (adapted from Jocelyn Popinchalk)

Each of these three clinics service different patient populations. Among those who received care between June 2015 and May 2017, the proportion of clinic attendees who were female was 65% in the CTC, 58% in the HTC, and 100% in the ANC. Median age was 37 years (interquartile range 29-46 years) in the CTC, 32 years (IQR 24-42) in the HTC, and 24 years (IQR 20-29) in the ANC.

In the CTC, 55% of patients reported being married and 39% reported being single. The majority of CTC patients (55%) were referred to care from a voluntary counselling and testing centre while another 17% were referred from an outpatient clinic. Two-thirds of CTC patients reported residence in Kisesa while the remaining 33% reported living outside the HDSS surveillance area. Among those who had data available for their first HIV-positive diagnosis, median duration since first HIV-positive diagnosis was 2.2 years (IQR 0.7-5.5 years). Most CTC patients (94%) were on ART.

In the HTC, 57% of patients reported being married and 24% reported being single. A majority (66%) of HTC patients reported education status at the primary level while another 16% reported secondary or higher educational attainment and 19% reported no education. Approximately 80% of HTC patients reported their occupation as farmer or trader. The majority of HTC patients (87%) reported their principal reason for visiting the HTC was to obtain knowledge of their HIV sero-status while another 8% reported feeling unwell or a recent illness. Nearly two-thirds (65%) of HTC patients reported residence in Kisesa while the remaining 35% reported living outside the HDSS surveillance area. Of all tests conducted between June 2015 and May 2017, 15% (n=593) were HIV-positive.

Among all ANC clinic attendees, median gestation at first visit was 22 weeks (IQR 20-26 weeks). About one-third of ANC patients reported no previous births (i.e., parity=0), 22% reported one previous birth, 16% reported two previous births, 12% reported three previous births, and the remaining 20% reported four or more previous births. Most ANC patients (86%) reported residence in the HDSS surveillance area; however, one-third of patients reported moving into the area during their current pregnancy. Approximately 7% of ANC patients either reported a previous HIV-positive diagnostic test or were diagnosed with HIV during ANC care.

The Kisesa study site provided an opportunity to link demographic and serological surveillance data with medical records from clinics offering HIV services located within the surveillance area. Prior to the research conducted for this thesis, Kisesa already had the capacity and necessary features, such as the electronic storage of data collected through surveillance activities and in some clinics, to implement a record linkage system like PIRL. This thesis sought to link individuals' medical records from the CTC, HTC, and ANC with the existing community databases created by the surveillance activities (Figure 1.3).

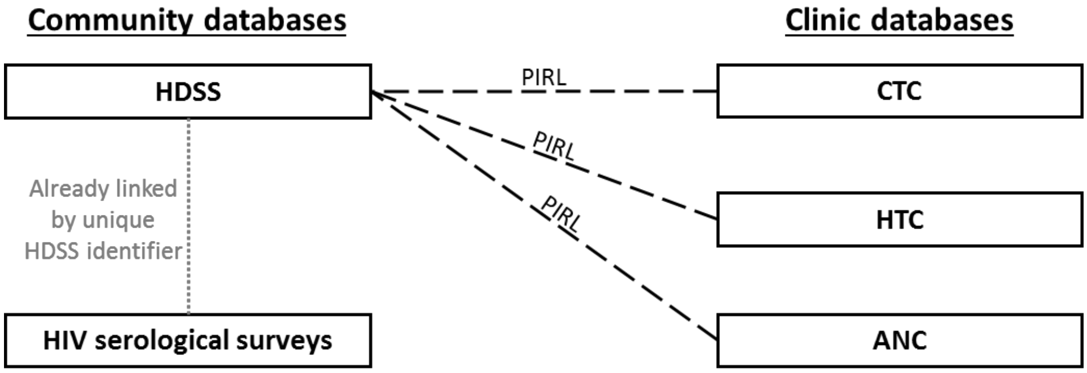


Figure 1.3: Proposed record linkage activities in Kisesa

Previously in Kisesa, a paper-based tracking system was developed to link patients testing HIV-positive in the HTC to the CTC; however, this system fell into disuse. A more robust form of linkage between the community cohort data and clinic data, as proposed by PIRL, would enable improved quantification of the uptake of and engagement with HIV services among all PLHIV in this community.

1.4.2 Community databases

The Kisesa HDSS includes 32 rounds (29 at the time fieldwork for this PhD was conducted) of household-based surveys that collect self-reported information on births, pregnancies, deaths, in- and out-migration, and spousal and parent-child relationships. In an HDSS round, an enumerator visits each household in the surveillance area. At each household, only one respondent reports information of all residents in the household. The respondents are normally heads of household though, on some occasions, the respondent is another adult household member who is well informed about the household.[68] Participation rates during HDSS rounds are very high (>98%), which is due to repeated attempts to survey a household if no connection is made during the initial visit and the good relationship between the TAZAMA team, leaders, and community members of Kisesa ward.

One major weakness of the Kisesa HDSS data is the limited reconciling of individuals records who move households within the HDSS area. Therefore, some individuals may have multiple HDSS records (and therefore, identifiers) if they resided in more than one household in the HDSS area since the start of the HDSS in 1994. Although not the primary purpose of this PhD research, migration reconciliation, the identification of a set of HDSS identifiers attributed to a single individual, will be a by-product of PIRL that could potentially be focused on and scaled up in future research.

Alongside HDSS rounds, there have been eight rounds of HIV sero-surveys conducted every three years, including HIV testing and a detailed questionnaire on sexual behaviour and partnership factors, fertility outcomes, HIV-related knowledge, and use of health services. Individuals who participate in an HIV surveillance round are given a unique identifier that links their records across multiple sero-survey rounds, and their current household-based identification from the HDSS is cross-referenced on their record.

1.4.3 Clinic databases

Three clinics offering HIV services located in Kisesa health centre were approached to participate in this study, all of which operate according to national prevention, care, and treatment guidelines and protocols.[71] The CTC databases have been fully digitised, and data clerks regularly update and run quality checks on these data. For the HTC and ANC, I developed electronic databases and trained a team to digitise the paper-based logbooks using a double-entry system whereby two different fieldworkers independently capture each book, and any discrepancy between fields are reconciled in subsequent cleaning stages.

Clinic identifiers are assigned and recorded differently in each clinic. During an initial visit to the CTC, individuals are issued a clinic ID card and receive a unique CTC number according to national guidelines that follows them throughout their care. CTC attendees are asked to bring their CTC ID cards with them on each return visit. If an individual does not return with their CTC ID card, a full-time data clerk who works in the CTC locates the individual's file and issues another CTC ID card with the same CTC ID.

Clinic logbooks in the HTC have been developed by the TAZAMA project and therefore I had greater control over the process of assigning and collecting these identifiers compared to those in the CTC and ANC. As individuals arrived to the HTC, PIRL fieldworkers assigned HTC IDs from a pre-determined, TAZAMA-created list of available HTC IDs. Fieldworkers collected these HTC IDs in the PIRL software and conducted PIRL prior to the individual's session with the HTC counsellor. Before the individual is tested, the HTC counsellor verified the HTC ID with those that are found in a pre-printed logbook. The HTC counsellor writes all personal information and test results on the line associated with the HTC ID. Importantly, individuals obtain an HTC number that identifies their test number. Attendees to the HTC collect their unique HTC ID number on a piece of cardstock paper to bring back on subsequent visits in order to link multiple tests for the same individual. If an individual does not return with their HTC ID card, they are issued a new HTC ID.

In the ANC, women receive a unique ANC number on an ANC ID card that identifies a particular pregnancy. If an individual does not return with their ANC ID card, an ANC nurse locates the individual's file and issues another ANC ID card with the same ANC ID. For any subsequent pregnancy a woman may have, she receives a new ANC number that is unlinked to her previous care in the clinic. The collection of clinic identifiers is described further in Section 2.4.1.

1.5 Structure of the Thesis

This thesis is presented in research paper format, including five published or submitted academic papers (A-E), and three additional chapters including this introductory chapter. A brief introduction is provided before each paper outlining the rationale for the paper and linking it to findings from preceding chapters.

Chapters 2, 3, 4, and 5 form the methodological portion of this thesis. Chapter 2 describes PIRL in detail, including the interview process and software used in the field, which has been published in an open-source repository using an MIT license.[72] This license allows others to download, edit, and use the software in any way as long as they provide attribution to the license holders. Paper A (Chapter 3), a Software Tool article, published in *Gates Open Research*,[73] provides further details on the tailored PIRL software that was used to collect the data for this thesis. This paper also presents anonymised versions of the data infrastructure created by PIRL using multiple case studies based on interactions I had in the field.

Paper B (Chapter 4), published in the *International Journal for Population Data Science*,[74] overviews the data created by PIRL and reports record linkage statistics, including match percentages and attributes associated with (un)successful linkage. This paper also includes a head-to-head comparison between PIRL and a fully automated linkage approach. Paper C (Chapter 5), under review at *BMC Medical Research Methodology*, measures the impact of linkage errors on bias and precision of analyses using data with high rates of linkage errors, as was the case when using automated linkage in Kisesa. Using an exemplar research question in the field of HIV epidemiology, this paper provides original evidence that analyses using linked data are impacted by substantial linkage errors similarly to how they are impacted by more negligible linkage errors as found in Europe and North America.

Chapters 6 and 7 form the substantive portion of this thesis. Using the novel data infrastructure created by PIRL, Paper D (Chapter 6), under review at *AIDS*, measures the extent of undisclosed HIV testing history among sero-survey participants who had attended a previous sero-survey or had previously registered for HIV care. Associations with non-disclosure of HIV testing history are identified and stratified by HIV test result. This paper quantifies the discrepancy between an estimate of the 'first 90' using sero-survey data per current UNAIDS guidelines and an augmented estimate using linked HIV testing history and medical records across three sero-survey rounds.

Paper E (Chapter 7), under review at *Tropical Medicine and International Health*, investigates linkage to care and ART initiation rates among individuals newly diagnosed with HIV by testing modality (i.e., where they received their first HIV-positive diagnosis). Systematic reviews on linkage to care in sub-Saharan Africa do not include sero-surveys as a testing modality nor a distinction between newly diagnosed individuals or repeat testers. This paper is the first to compare linkage to care and ART initiation rates between newly diagnosed individuals in a sero-survey with those diagnosed using voluntary or provider-initiated HTC in a stationary clinic.

Finally, Chapter 8 synthesises findings across the objectives, provides recommendations for programmes, policy, and future research, discusses strengths and limitations of this PhD research, and lists efforts to disseminate findings.

Appendices include select conference presentations, ethical clearances, consent forms, certificates for completed trainings undertaken during the PhD, an example of a field report that was circulated monthly to the linkage team, an agenda for training I provided to the field team, evidence of retention of copyright for the research papers in this thesis, and supplementary material published alongside the research papers.

The original contributions of this thesis are: the augmentation of existing computer software toward a novel approach of record linkage, a nascent data infrastructure of high quality links between community cohort and health facility data that enabled this PhD research and provides legacy for future research, the measurement of bias and precision in analyses using data with substantial linkage errors, the quantification of bias in the 'first 90' across multiple rounds of population-based surveys at a sub-national level, and the measurement of linkage to care and ART initiation rates among newly diagnosed individuals across three testing modalities including sero-surveys.

1.6 ROLE OF THE CANDIDATE

1.6.1 Overall design and planning

I contributed to the overall concept and led the framing of the research questions and design for this study. I was successful in securing additional funding from the Economic and Social Research Council (ESRC) to supplement fieldwork support previously awarded to the ALPHA Network. I prepared all applications to ethical review boards associated with PIRL.

1.6.2 Software

A computer software package that prospectively links records between community and clinic data was originally conceived and built for use in another ALPHA Network site (Agincourt HDSS in South Africa), which has led to multiple manuscripts showing the promise of record linkage in resource-constrained settings.[38, 75] The Agincourt programmers built the structure of a software package tailored to Kisesa during a two-day trip to Kisesa in summer 2014 (prior to my PhD start). Included in their package was a linkage algorithm, match-probabilities, and agreement conditions that were created from and tested on Agincourt data.

I met with one of the Agincourt programmers after I started my PhD in late 2014, at which time I was transferred the software that was written that summer. I took multiple face-to-face computer programming classes at City University London (Appendix 10.4) and online courses on Microsoft Pluralsight to obtain the skills necessary to adapt the software that was provided to me into the PIRL package used to collect data for this PhD research. I worked in-person and online with Jason Catlett, a Data Architect from Atlanta, Georgia, USA, throughout early 2015 to reconcile the PIRL package, including the database management system, with the original conceptualisation of the software to be used in Kisesa.

Before implementing the software in the field, I added clinic ID fields that were not included in the Agincourt-created version of the software. I also coded several data integrity checks, including double entry and check-digits on all clinic ID fields (described further in Section 2.4.1). I amended several fields from free-text to a drop-down list, including villages and sub-villages, and ensured these fields were standardised in the HDSS data sources. I developed import and export scripts for the TAZAMA data manager to run at the end of each working day. I also ran monthly checks to ensure there was consistency between Agincourt-derived and Kisesa-derived match-probabilities included in the linkage algorithm; there was never any indication that suggested an adjustment to be made for Kisesa data. I also tested six alternative linkage algorithms by adding and removing variables, adjusting match-probabilities, amending agreement conditions; none of which performed differently from the original algorithm.

I solely led the implementation of the software in Kisesa and directed a field team of up to seven members for three years. I performed validation of the data collected through the PIRL software and disseminated updates through monthly calls with the field team (Appendix 10.5). I documented and uploaded the software to GitHub[72] – an open-source repository – and published a Software Tool article describing the PIRL package,[73] as presented in Chapter 3.

1.6.3 Fieldwork operations

Alongside my NIMR colleagues, I conducted interviews and hired an original team of four fieldworkers for this PhD research. I created and led a two-day training of the field team, including the data manager, on how to approach potential study participants, obtain informed written consent, conduct brief interviews, and use the PIRL software (Appendix 10.6).

During the course of this PhD, I travelled to Tanzania on three occasions for a total of approximately four months (Table 1.1). On 1 June 2015, I began to roll out record linkage operations in each of the three clinics in Kisesa health centre.

Trip number	Date of arrival	Duration of stay	Primary achievements
1	26/11/2014	1 week	· Introductions made with the study site and clinic staff · Conducted pilot study in Kisesa health centre to test software and work out study protocol · Visited two health posts to assess the feasibility of introducing record linkage in them · Obtained data clearances for TAZAMA data · Received a copy of the record linkage software
2	27/03/2015	1 month	· Interviewed, hired, and trained four fieldworkers on research ethics and record linkage operations · Created electronic data infrastructures to digitally capture paper logbooks used in the HTC and ANC · Re-designed HTC logbook to minimise previously found discrepancies when linking personal identifiers with HIV test results · Met with the District Medical Officer to give overview of record linkage project as well as discuss long-term visions of a unified registration office at Kisesa health centre
3	18/05/2015	2.5 months	· Introduced record linkage operations in the CTC, ANC, and HTC · Monitored the record linkage software and the database management system for quality and reliability · Held daily trainings with field team to share experiences, ask questions, and gain knowledge surrounding daily operations in each clinic · With the assistance of local researchers, translated most of the record linkage software into Kiswahili so the user is presented with both English and Kiswahili instructions

Table 1.1: Fieldwork trips

1.6.4 Secondary data

This PhD research also relied on secondary data collected through the TAZAMA Project, including the HDSS and sero-survey data. Although I was not responsible for the design or management of the surveillance data, I cleaned and transformed the HDSS data that was embedded within the PIRL software. I have also given feedback to the TAZAMA Project principal investigators on how to enhance future record linkage work in Kisesa.

Routinely collected clinic data from the government-run health facilities in Kisesa health centre were also used. With the assistance of Denna Michael, a NIMR researcher, I liaised with the Kisesa health centre physician, counsellor, and nurses in each of the three clinics and obtained their approval for enrolling clinic attendees into the PIRL study and collecting clinic data. The CTC databases were already in electronic form and were routinely updated and cleaned by government data entry clerks. In the HTC, previous paper logbooks were entirely handwritten in blank notebooks and managed by the HTC counsellor, with individual identifiers appearing in one notebook and test results in another. Incorporating feedback provided by previous PhD students who worked with HTC data, the NIMR data manager (Richard Machemba) and I designed new paper logbooks and clinic ID cards for the HTC. The updated books were created electronically and included all HTC IDs pre-printed on every row to minimise the potential for errors arising from incorrectly written or transcribed identifiers. The ANC uses an abundance of paper logbooks to record care received, including separate logbooks for mothers, children, residents, non-residents, PMTCT, family planning, labour and delivery services, and many more. I perused all logbooks and made the decision on which logbooks to digitally capture for the purposes of this PhD research. During weekly calls with the fieldwork team, I supervised the progression of digitising the paper logbooks in the HTC and ANC.

1.6.5 Analyses

I designed and executed the statistical analyses presented in this thesis in consultation with London School of Hygiene and Tropical Medicine (LSHTM) and NIMR statisticians. I attended the 2015 ALPHA Network workshop in Entebbe, Uganda, which exposed me to the data and methods used to analyse surveillance data from Kisesa.

Unique to a four-year PhD studentship from the ESRC, a Postgraduate Diploma in Research Methods (PGDip) must be obtained alongside the PhD research. A PGDip requires taking and passing assessments on eight modules or approximately 120 credit hours within the PhD student’s registered department. I took the following modules offered at LSHTM: Demographic Methods, Population Studies, Population Dynamics and Projections, Analysing Survey and Population Data, Analysis of Hierarchical and other Dependent Data, Advanced Statistical Methods in Epidemiology, Spatial Epidemiology in Public Health, and Advanced Statistical Modelling. My overall Award GPA for these modules have placed my PGDip under consideration for distinction.

In addition, I took two short courses offered at LSHTM: Causal Inference in Epidemiology: Recent Methodological Developments, and Introduction to GIS. These modules and additional trainings were selected to give me analytic skills that would be useful for this PhD research and beyond.

1.6.6 Dissemination

For all five academic papers presented in subsequent chapters, I led the conception of the investigation, designed the study, created the analysis plan, extracted and cleaned data, conducted analysis, drafted the manuscript, collated feedback from co-authors, submitted the manuscript, and for those that have passed the review stage, liaised with journal editors. I wrote all additional chapters in this PhD, incorporating feedback from my PhD supervisors. I attended domestic and international conferences for poster and oral presentations on my findings and also disseminated presentations and academic papers resulting from this PhD research to my NIMR colleagues. Conference presentations can be found in Appendix 10.1. Further dissemination efforts are detailed in Section 8.6.

1.7 Ethical clearance

Ethical approval for this PhD research was obtained by LSHTM (Project ID #8852) and the National Institute for Medical Research, Tanzania (ref MR/53/100/314 and MR/53/100/450) (Appendix 10.2). Informed written consent to link records across data sources was obtained from all participants (Appendix 10.3).

1.8 Funding

This PhD research was funded by a four-year PhD studentship through the ESRC, which covered my tuition and annual stipend. The ESRC studentship was supplemented by an Advanced Quantitative Methods (AQM) enhanced stipend, which was to encourage further training in AQM and apply this to this PhD research and beyond. I was also awarded a Collaborative Development Grant to support my travel to Kisesa and online computer programming training for the NIMR data manager who oversaw PIRL activities on a daily basis. Leftover funds from this award were paid out to the field team as bonuses after two years of the PIRL study.

Travel and fieldwork costs, such as laptops and fieldworker salaries, were supported by the Bill and Melinda Gates Foundation (OPP1082114 and OPP1120138). Other fieldwork costs have been supported through alternative ESRC funding streams, including the Research Training Support Grant and the Overseas Fieldwork Grant, and the Measurement & Surveillance of HIV Epidemics (MeSH) Consortium. The ongoing TAZAMA Project research activities were funded by the Global Fund to fight AIDS, TB, and Malaria. Of note, the core funding for PIRL activities in Kisesa, as anticipated, were depleted by 31 May 2017, exactly two years after PIRL was introduced in the field. The original intention was for further linkage activities to be included in the larger funding package obtained for repeated HDSS rounds; however, such support was not obtained for the remainder of this PhD research. Thus, primary data collection for this PhD extended from 1 June 2015 through 31 May 2017.

2 Field methodology

In this section, I will detail the field methods implemented in Kisesa to collect data used for analysis in this thesis.

2.1 Field team

Record linkage fieldwork started in Kisesa health centre on 1 June 2015. At the beginning of the study, the team was comprised of one data manager and four fieldworkers, one of whom had previous experience with management of health facility and HDSS data and three others who had experience with HDSS data only. Before the initial rollout of the software, I provided formative training to all fieldworkers and the data manager (Appendix 10.6). The training session included instructions on how to obtain informed consent, conduct brief interviews, and several demonstrations of the PIRL software. Fieldworkers who were hired after the initial rollout of the software were trained by the data manager and existing fieldworkers through shadowing and close oversight for at least one month before working on their own.

During the first four months, fieldworkers were assigned to a single clinic. Beginning in October 2015, the fieldworkers rotated between clinics bi-weekly to mitigate any potential of bias by a fieldworker's level of experience on that clinic's linkage statistics. At any time over the study period, a fourth fieldworker would substitute for any of the three primary fieldworkers in case of any absences. Each fieldworker functioned as both a linkage interviewer and switched duties to digitise paper logbooks when all clinic attendees had been interviewed on a given day. Each was equipped with a password-protected laptop that ran the PIRL software. The fieldworkers were fluent in English, KiSwahili, and the local language, KiSukuma.

2.2 Interview process

All individuals who attended any of the three clinics offering HIV services in Kisesa health centre were invited to participate in this research. No invitations or advertisements were used to invite individuals to participate in this research. There were no restrictions based on age; if a patient was less than 18 years of age, they were required to have a parent or legal guardian present. Informed written consent (for adults) and assent (for those <18 years) were obtained from all individuals who participated in this project. Figure 2.1 overviews the PIRL process used for this PhD research.

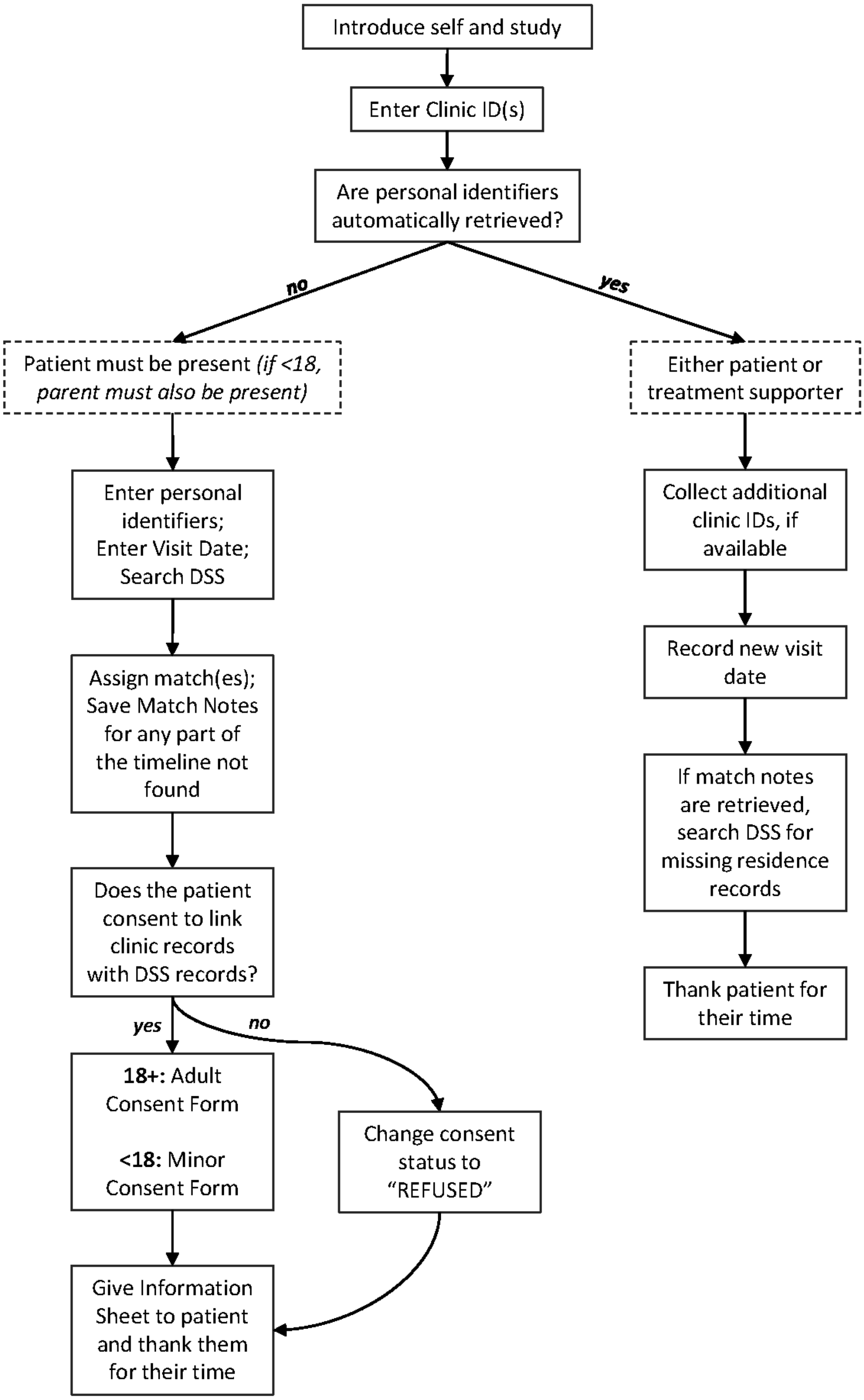

Introduce self and study
Enter Clinic ID(s)
Are personal identifiers automatically retrieved?
no
yes
Patient must be present *(if <18, parent must also be present)*
Either patient or treatment supporter
Enter personal identifiers; Enter Visit Date; Search DSS
Collect additional clinic IDs, if available
Assign match(es); Save Match Notes for any part of the timeline not found
Record new visit date
If match notes are retrieved, search DSS for missing residence records
Does the patient consent to link clinic records with DSS records?
Thank patient for their time
yes
no
18+: Adult Consent Form
<18: Minor Consent Form
Change consent status to "REFUSED"
Give Information Sheet to patient and thank them for their time

Figure 2.1: PIRL process

As individuals arrived at the clinics, a fieldworker introduced him/herself and then described the study. Since the CTC and ANC have high patient loads who often arrive to clinic before opening, the fieldworkers, upon their arrival and at regular intervals throughout the day, introduced themselves and provided information about the project to all patients in the clinic's waiting room as a group. In the HTC, individuals tended to arrive steadily throughout the morning, which enabled the fieldworker to introduce him/herself and the project on an individual basis as the patients arrived. In each clinic, the fieldworker handed out number cards (e.g., 1-20) to clinic attendees with the intention to be in the order in which patients arrived at the clinic.

The fieldworker then invited an attendee by number to a desk located within the clinic but out of the way of normal clinic operations to conduct the brief record linkage interview. We tried to situate the desks in a private area, but given the limited space, some were off to the side of the waiting room. The interview only involved asking for demographic information, such as name, sex, date of birth, and residence details, and did not ask for any medical information.

The primary goals of the interview were to identify the true HDSS record(s) and to confirm residence histories of all participants using the PIRL software. I trained the fieldworkers to use interview tools and ask probing questions such as, "How long have you lived in your current residence?" As a patient gave details of their residence history, the fieldworkers were trained to construct a residency timeline on a notepad (Figure 2.2). Since the first HDSS survey was conducted in 1994, the fieldworker probed about residence history from 1994 through to the most current HDSS survey, which was through 2014, inclusive, at the time of fieldwork for the PhD research. The history of the attendee's residency assisted the fieldworker in searching for potential matches by knowing how many HDSS records they were expected to find including their time period and location.

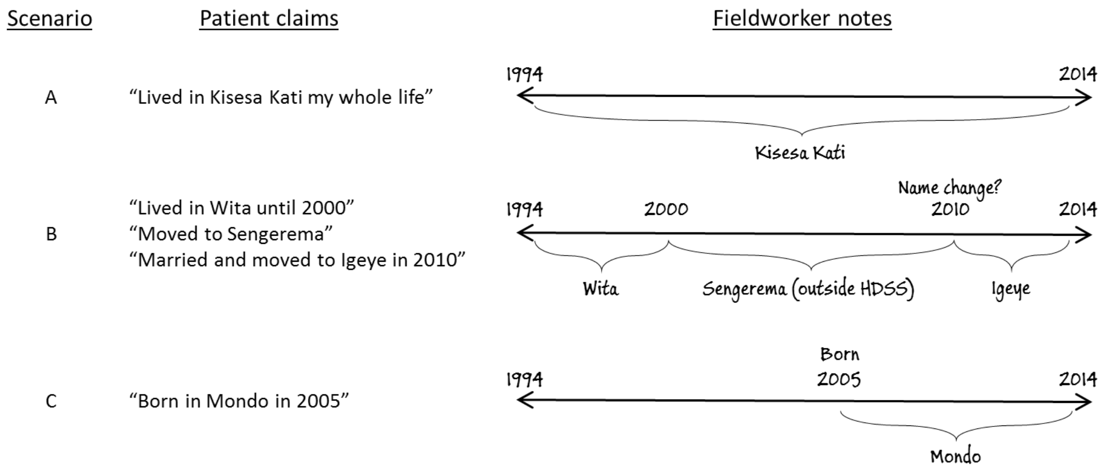
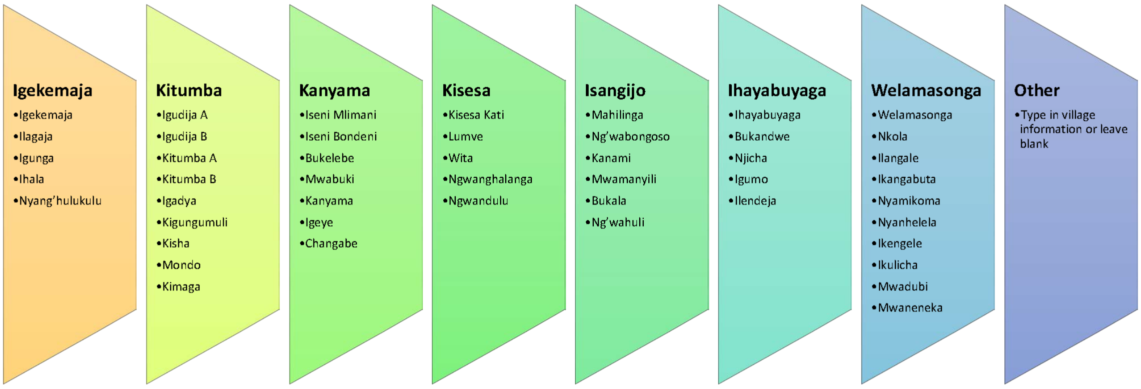


Figure 2.2: Examples of residency timelines

Shortly after PIRL was launched in the clinics, I realised there was confusion over the term "Kisesa", which not only refers to the ward (i.e., entire HDSS surveillance area), but also one of the seven villages within Kisesa ward, and a sub-village within that village wherein the health facility is located. This realisation made it conceivable that some clinic attendees may have reported not living in "Kisesa" because they interpreted the question to mean village or sub-village rather than ward/surveillance area. As soon as I became aware of this potential issue, we ceased from asking participants if they "lived in Kisesa" and instead asked a more open-ended question, "Where do you live?". To assist fieldworkers and participants, I created a complete list of village and sub-villages located within the surveillance area (Figure 2.3) and ensured this list corresponded to the drop-down lists in the PIRL software. Further, I made this list part of each fieldworker's laptops desktop background, along with other useful tools to have during the linkage interviews.



Igekemaja
- Igekemaja
- Ilagaja
- Igunga
- Ihala
- Nyang'hulukulu

Kitumba
- Igudija A
- Igudija B
- Kitumba A
- Kitumba B
- Igadya
- Kigungumuli
- Kisha
- Mondo
- Kimaga

Kanyama
- Iseni Mlimani
- Iseni Bondeni
- Bukelebe
- Mwabuki
- Kanyama
- Igeye
- Changabe

Kisesa
- Kisesa Kati
- Lumve
- Wita
- Ngwanghalanga
- Ngwandulu

Isangijo
- Mahilinga
- Ng'wabongoso
- Kanami
- Mwamanyili
- Bukala
- Ng'wahuli

Ihayabuyaga
- Ihayabuyaga
- Bukandwe
- Njicha
- Igumo
- Ilendeja

Welamasonga
- Welamasonga
- Nkola
- Ilangale
- Ikangabuta
- Nyamikoma
- Nyanhelela
- Ikengele
- Ikulicha
- Mwadubi
- Mwaneneka

Other
- Type in village information or leave blank

Figure 2.3: List of villages and sub-villages in Kisesa HDSS

The software uses demographic and residence details that a participant has shared to search through the HDSS database and output the top 20 most likely matches. Once the potential matches were on screen, the fieldworker began with the highest ranked potential match (based on match score – see Section 2.4.2) and asked the participant if s/he knew any of the other individuals listed in the household. Household membership was used as an extra step to adjudicate whether the HDSS record in question was indeed a true match. The fieldworker was instructed to assess each record in a stepwise fashion until all matches were found.

If a fieldworker's first search did not result in identifying all HDSS records as expected from the timeline constructed during the brief interview, participants were asked if they went by any other names or had moved residence since 1994. By design, the HDSS data are collected during household-based surveys in which one household representative reports on behalf of the entire household. Therefore, the name collected during an HDSS round may not be the same as an individual reported in a health facility. Fieldworkers, who all had experience working with HDSS data, were trained to probe for identifying information that would be on an individual's HDSS record, update the information in the software, and repeat the search attempt.

Once all matches were made, the fieldworker ended the session in the PIRL software, at which point all collected data was deleted from the screen. If a match was not made, an open-text field in the software was available for the fieldworker to input comments from the interview that may have caused not finding a match (e.g. an individual moved into the HDSS area only two weeks prior). These notes were saved in the software for each interview session and were retrieved by the software during subsequent visits to guide the fieldworkers' future searches.

When a clinic attendee was approached a second time and thereafter, and if they returned with their clinic ID card(s), the fieldworker input the unique clinic identifier to automatically retrieve all information saved during previous visits. At this point, the fieldworker logged the date of the new visit and checked the match status and/or match notes from the previous session(s), which enabled the fieldworker to quickly reconstruct the patient's residency history on a timeline and determine if an HDSS record was yet to be found. If all HDSS records had been found for the patient, no further searching was required. However, if there were any HDSS records remaining to be found, these repeat visits offered an additional opportunity to link the participants' records.

2.3 INFORMED CONSENT/ASSENT

All patients who participated in the PIRL project were offered informed written consent (Appendix 10.3). For clinic attendees under the age of 18 years, both a parental consent and minor assent form were required. Participants could sign the written consent with a pen or by thumbprint, as per local guidelines. All study documents for participants were made available in English and KiSwahili.

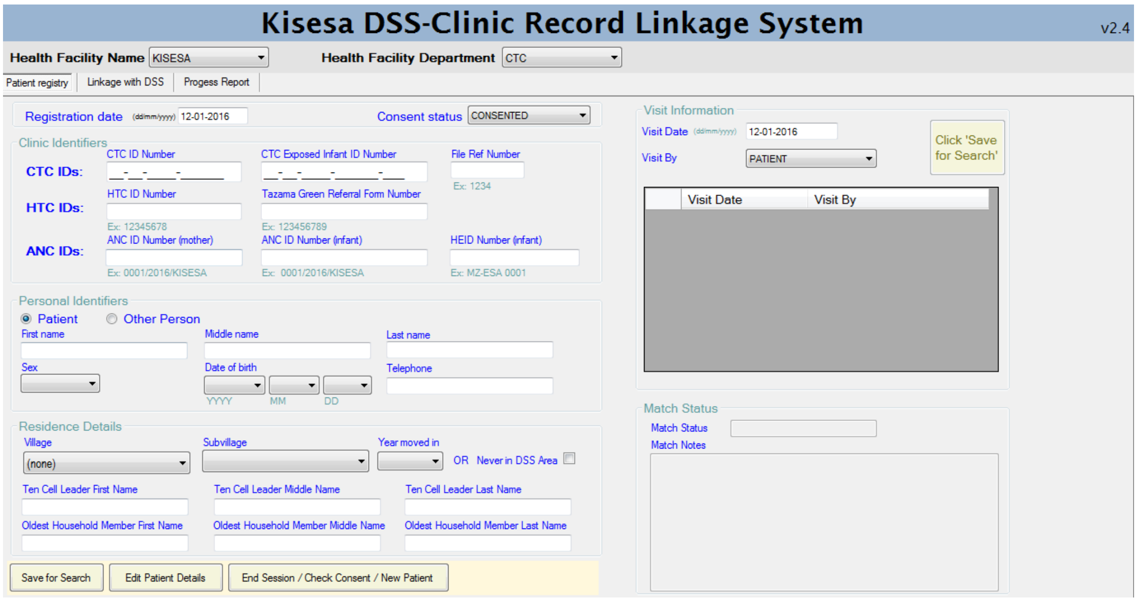
All participants were offered a patient information sheet that included details on the study as well as contact information in case of enquiries or a desire to be removed from the study (Appendix 10.3.3). If any patient who had been previously interviewed and linked expressed that they no longer wanted their medical records linked with their community data, the fieldworker asked the patient for their unique clinic identifiers and passed this information along to the field manager. It was the responsibility of the data manager under my supervision to retrieve the individual's linked information and delete the link between the clinical and demographic records.

All completed consent forms were kept by each fieldworker in a binder throughout the day and were combined and placed in a locked safe in a locked room in the CTC at the end of each day. A total count of consent forms was sent to me daily, which I verified in the data. At regular intervals, the data manager collected the forms and transported them to the main TAZAMA office in Mwanza where they were stored in a locked cabinet.

2.4 SOFTWARE

Chapter 3 describes the rationale, implementation, operation, and system requirements for the PIRL software in detail. In addition, a full user guide, including how to install and use the PIRL software, is available online in an open-source repository.[72]

Briefly, all relevant clinic and personal identifiers and residence details used in the probabilistic linkage algorithm were entered on the patient registry page (Figure 2.4).

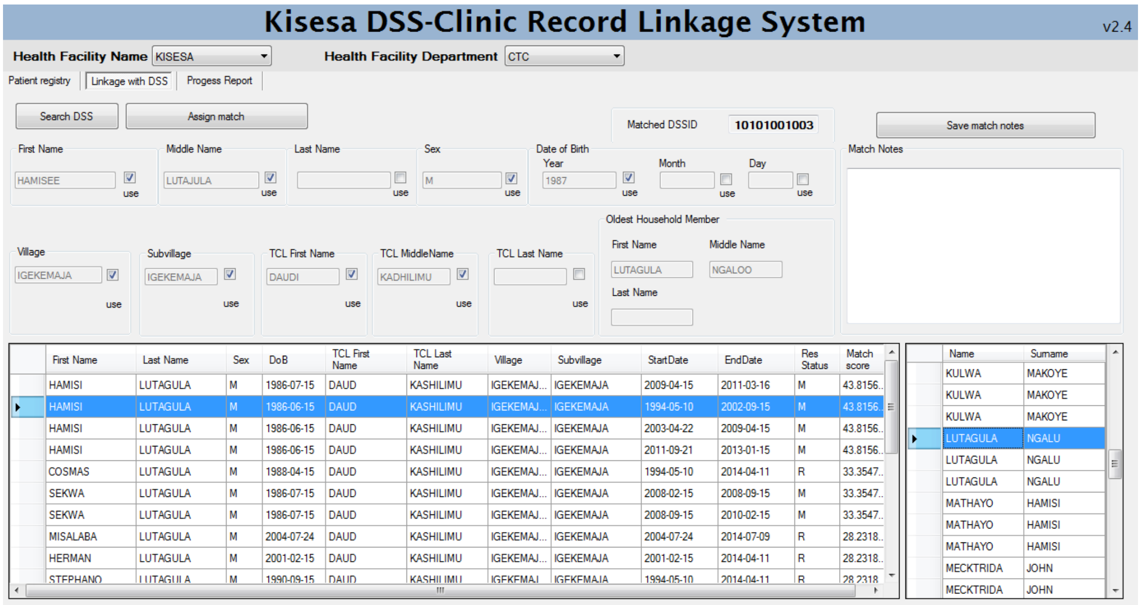


Health Facility Name: KISESA
Health Facility Department: CTC
Registration date (dd/mm/yyyy): 12-01-2016
Consent status: CONSENTED
● Patient ○ Other Person
Village: (none)
Visit Date (dd/mm/yyyy): 12-01-2016
Visit By: PATIENT

Figure 2.4: Screen shot of the patient registry page in the PIRL software

Once all information was collected on the patient registry page, the fieldworker moved to the “Linkage with DSS” tab in the software and clicked “Search DSS” (Figure 2.5). The software invoked the linkage algorithm and output the most likely matches based on the calculated match score – a weighted value denoting the (dis)similarity of every record-pair (described further in Section 2.4.2). On this tab, the fieldworker was able to view the list of potential matches, and the entire household profile for each record.

The version of the software I received at the start of my PhD took approximately 90 seconds, on average, to perform a search on the HDSS database. One of the key enhancements I, along with Jason Catlett, made to the software was streamlining the code that performed the search. The version of the software used in the field for this PhD research took approximately 10-15 seconds, on average, to perform a single search.



Kisesa DSS-Clinic Record Linkage System v2.4

Health Facility Name: KISESA

Health Facility Department: CTC

Matched DSSID: 10101001003

First Name: HAMISEE ☑ use

Middle Name: LUTAJULA ☑ use

Last Name: ☐ use

Sex: M ☑ use

Date of Birth: Year: 1987 ☑ use; Month: ☐ use; Day: ☐ use

Match Notes:

Village: IGEKEMAJA ☑ use

Subvillage: IGEKEMAJA ☑ use

TCL First Name: DAUDI ☑ use

TCL MiddleName: KADHILIMU ☑ use

TCL Last Name: ☐ use

Oldest Household Member

First Name: LUTAGULA

Middle Name: NGALOO

Last Name:

	First Name	Last Name	Sex	DoB	TCL First Name	TCL Last Name	Village	Subvillage	StartDate	EndDate	Res Status	Match score
	HAMISI	LUTAGULA	M	1986-07-15	DAUD	KASHILIMU	IGEKEMAJ...	IGEKEMAJA	2009-04-15	2011-03-16	M	43.8156..
▶	HAMISI	LUTAGULA	M	1986-06-15	DAUD	KASHILIMU	IGEKEMAJ...	IGEKEMAJA	1994-05-10	2002-09-15	M	43.8156..
	HAMISI	LUTAGULA	M	1986-06-15	DAUD	KASHILIMU	IGEKEMAJ...	IGEKEMAJA	2003-04-22	2009-04-15	M	43.8156..
	HAMISI	LUTAGULA	M	1986-06-15	DAUD	KASHILIMU	IGEKEMAJ...	IGEKEMAJA	2011-09-21	2013-01-15	M	43.8156..
	COSMAS	LUTAGULA	M	1988-04-15	DAUD	KASHILIMU	IGEKEMAJ...	IGEKEMAJA	1994-05-10	2014-04-11	R	33.3547..
	SEKWA	LUTAGULA	M	1986-07-15	DAUD	KASHILIMU	IGEKEMAJ...	IGEKEMAJA	2008-02-15	2008-09-15	M	33.3547..
	SEKWA	LUTAGULA	M	1986-07-15	DAUD	KASHILIMU	IGEKEMAJ...	IGEKEMAJA	2008-09-15	2010-02-15	M	33.3547..
	MISALABA	LUTAGULA	M	2004-07-24	DAUD	KASHILIMU	IGEKEMAJ...	IGEKEMAJA	2004-07-24	2014-07-09	R	28.2318..
	HERMAN	LUTAGULA	M	2001-02-15	DAUD	KASHILIMU	IGEKEMAJ...	IGEKEMAJA	2001-02-15	2014-04-11	R	28.2318..
	STEPHANO	LUTAGULA	M	1990-09-15	DAUD	KASHILIMU	IGEKEMAJ	IGEKEMAJA	1994-05-10	2014-04-11	R	28.2318

	Name	Surname
	KULWA	MAKOYE
	KULWA	MAKOYE
	KULWA	MAKOYE
▶	LUTAGULA	NGALU
	LUTAGULA	NGALU
	LUTAGULA	NGALU
	MATHAYO	HAMISI
	MATHAYO	HAMISI
	MATHAYO	HAMISI
	MECKTRIDA	JOHN
	MECKTRIDA	JOHN

Figure 2.5: Screen shot of the linkage tab in the PIRL software

2.4.1 Clinic identifiers

Each of the clinics have their own personal identification process that is outside the control of the researchers, except for the HTC (as described in Section 1.4.3). The number of different clinic identifiers varied by clinic (Figure 2.6).

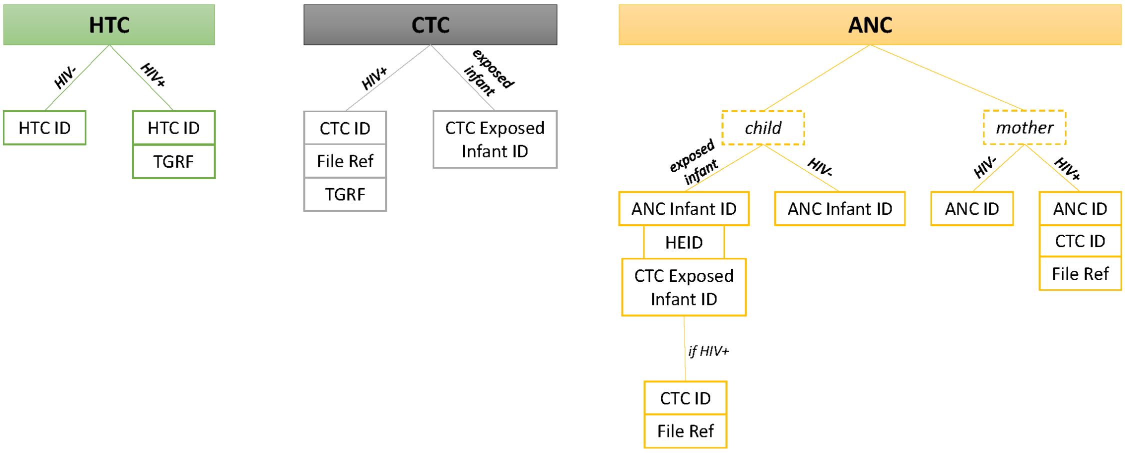


Figure 2.6: List of available clinic identifiers by clinic.
NB. TAZAMA Green Referral Form (TGRF) refers to a paper-based tracking system to link HTC and CTC records, but this system fell out of use

After a clinic attendee agreed to participate in this research, the fieldworkers asked participants for their clinic identification cards, knowing that some individuals will have more than one clinic identifier, especially if they had received care from multiple clinics. Clinic identifiers were the crucial piece of information that not only linked the medical records to the community data but were also used to link records across subsequent visits to the same or different record linkage clinic. Since the data between all three fieldworkers' machines were synced daily, a patient's clinic identifiers retrieved previous linkage sessions regardless of the clinic they attended. Thus, as patients attended multiple clinics, we had the opportunity to add all clinic identifiers that a patient has ever used to their profile. However, since attendees, particularly those in the HTC, were less likely to bring their ID cards with them to subsequent visits (reasons included misplacing the card or not wanting to keep a card that was associated with HIV testing), linkages to the same HDSS records over multiple visits were used to obtain a more accurate portrayal of visit patterns to the clinics.

Due to their importance, I coded several data integrity checks for clinic identifiers into the software. All clinic ID fields required double-entry, had internally coded checks that immediately warned a fieldworker if the format was incorrect, and listed examples of each clinic ID underneath each entry box in the PIRL software. All CTC IDs offered throughout Tanzania follow the same 14-digit format, and the software automatically places the entered digits into this format. However, some individuals presented to care with outdated 11-digit CTC numbers or those from health facilities outside Kisesa HDSS area, including the Magu District health centre or Bugando Medical Centre (BMC) in Mwanza. National guidelines were developed to convert these outdated CTC numbers into the current 14-digit format prior to this PhD research. However, I created a colour-coded conversion chart that was included on the desktop background of the fieldworkers' machine to minimise the risk of conversion errors (Figure 2.7).

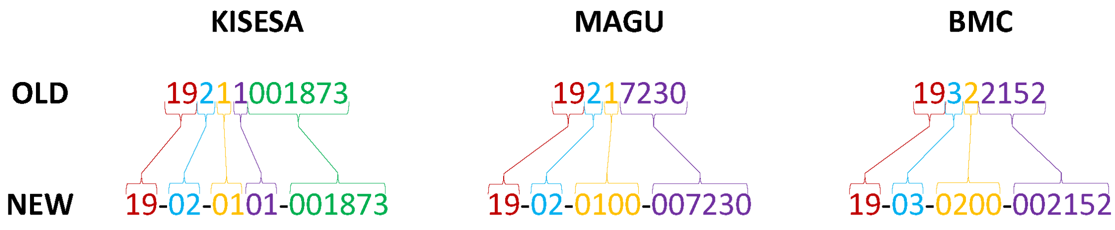


Figure 2.7: CTC ID conversion chart

The HTC IDs included check digits based on a modulus 97 algorithm – a method to ensure validity of the HTC ID entered into the software. The modulus 97 algorithm dictates that any number divided by 97 should result in a remainder of 1. Here, I provide an example if an HTC ID was 1941. Dividing 1941 by 97 results in 20 with a remainder of 1. The next HTC ID in this sequence could be 2038, since 2038 divided by 97 results in 21 with a remainder of 1.

The uniqueness of ANC IDs, however, was complicated by several factors. ANC IDs started at "0001" and incremented by one for each new clinic attendee in each ANC facility and calendar year. Therefore, there were multiple individuals in Tanzania with an ANC ID of "0001" in each calendar year, and it was possible they obtained care at multiple ANC clinics. I devised a numbering system to be used in this PhD research that amalgamated an individual's ANC ID with the year and village name in which they initiated care. For example, for the first pregnant mother to initiate care in the ANC at Kisesa health centre in 2015, the fieldworker recorded, "0001/2015/KISESA." As mentioned, ANC IDs did not carry over to subsequent pregnancies. Similar to individuals not presenting with ID cards in the HTC at each visit, linkages to the same HDSS records were used to link a single mother's entire ANC visit history over multiple pregnancies.

2.4.2 Record linkage algorithm

The PIRL software utilises a probabilistic search algorithm to identify and rank potential matches in the HDSS database. The algorithm incorporated the following parameters or data fields: up to three names for the participant; sex; year, month, and day of birth; village and sub-village; up to three names of a household member; and up to three names for the ten-cell leader (TCL) of the participant. A ten-cell leader, locally known as a "balozi," is an individual who acts as a leader for a group of ten households and these positions have been relatively stable over time.

The algorithm used for searching possible matches and ranking them was based on a the Fellegi-Sunter record linkage model,[28, 29] with match probabilities (m_i) adopted from a prior linkage study in Agincourt HDSS.[38] Probabilistic record linkage has been well described.[27, 76-79] In brief, let M be a set of true matches and U be a set of true non-matched record pairs. Two individual agreement probabilities were defined for each field i in record pair j as follows:

$$\text{match probability: } m_i = \text{P(field } i \text{ agrees} \mid j \in M) \qquad (2.1)$$

$$\text{unmatch probability: } u_i = \text{P(field } i \text{ agrees} \mid j \in U) \qquad (2.2)$$

The higher the ratio m_i/u_i, the more useful a field was for matching purposes. For a given field with match probability m_i and unmatch probability u_i, matching weights were calculated as w_{ai} = log₂[m_i/u_i] for fields where both datasets agree, and w_{di} = log₂[(1-m_i) /(1-u_i)] where they disagree. Logarithms (base 2) are commonly used in probabilistic linkage as it enhances the interpretability of the match weights, so that a one-unit increase corresponds to a doubling in the ratio for a matched record-pair.[27] Match scores were computed by summing the weights across all fields with collected information.[29, 78] Incomplete fields did not add or subtract from the match score.

Spelling errors, the use of multiple names (including nicknames), and interchangeable name order complicate locating an exact match between names in these databases. Numerous methods have been developed to compare the (dis)similarity between two string variables (e.g., names), known as string comparators. Edit distance methods[80-84] calculate the number of operations (character deletions, insertions, and substitutions) required to turn one string into the other string. Alternatively, a *q*-gram based method[85-87] splits an input string into shorter sub-strings of length *q* characters and calculates the similarity between the sub-strings. Many other string comparator methods have been developed and used in record linkage studies.[83, 88-100] However, comparative studies between the various string comparator methods have suggested that the Jaro-Winkler method,[101] which is essentially a combination of the edit distance and *q*-gram methods described above, often outperforms the others in a variety of settings including in the Agincourt HDSS site in South Africa (Table 2.1).[38, 76, 81, 102-104] Thus, the PIRL software used in Kisesa incorporated the Jaro-Winkler string comparator approach to compare the name fields between two records allowing for all pairwise comparisons between reported names and names found in the HDSS. Informed by analyses conducted on Agincourt data, a name-pair resulting in a Jaro-Winkler score of ≥0.8 was considered an agreement.[38]

Table 2.1: Examples of Jaro-Winkler comparisons

Name1	Name2	Jaro-Winkler score
MANENO	MENANO	0.950
JULIANA	JULLIANNA	0.948
JANE	JAN	0.942
KABULA	KADULLA	0.879
YONAH	JONAH	0.867
LUCIA	RUCIA	0.867
YUNGILE	LONGILE	0.810
LUFAN	RUFANNE	0.790
ALLY	ALI	0.778
SHIJA	MASANJA	0.565

Agreement conditions varied for the other matching variables (Table 2.2). For year of birth to agree, the difference could differentiate up to two years. All other parameters (sex, month and day of birth, village, and sub-village) were required to agree exactly.

Table 2.2: Agreement conditions for the identifiers included in the linkage algorithm

Identifier	Agreement condition
First name	Jaro-Winkler ≥ 0.8
Second name	Jaro-Winkler ≥ 0.8
Third name	Jaro-Winkler ≥ 0.8
TCL first name	Jaro-Winkler ≥ 0.8
TCL second name	Jaro-Winkler ≥ 0.8
TCL third name	Jaro-Winkler ≥ 0.8
Sex	exact match
Year of birth	within two years
Month of birth	exact match
Day of birth	exact match
Village	exact match
Sub-village	exact match

2.4.3 Review of links

Matches selected in the field were assumed to be true matches. As an additional data integrity check, I performed periodic and manual, back-end inspection of the data to verify the matches made in the field. These checks flagged individuals who were matched to multiple HDSS records with large age differences (>10 years), of conflicting sex, within the same household, or with overlapping residency episodes in which one record’s start date occurred before another record’s end date. During this PhD research, eight (0.2%) matches were deemed unlikely and were deleted from the back-end database. Further details are provided in Section 3.4.3 on page 57.

During the pilot phase of the software in November 2014, I learned the most likely reasons for not finding a match were having no residence history in the HDSS surveillance area and migrating into the area or born after the last HDSS round. I adapted the software to flag these individuals in the data.

2.4.4 Data privacy and storage

All interactions with the PIRL software were logged and labelled with a unique username for each fieldworker. The data collected by the PIRL software included personal identifiers used by the linkage algorithm, clinic identifiers, and visit dates. No medical information was captured or stored in the PIRL database management system. Data were stored on password-protected laptops and in an encrypted form. Once a fieldworker ended a session with a participant, they could not access the unencrypted data. At the end of each working day, the data manager collated the data collected on each laptop and performed a backup of the database. At regular intervals, the data manager transferred the encrypted database to a password-protected server housed in the TAZAMA Project data room at NIMR campus in Mwanza, which is a guarded and gated facility. Only the PIRL project data manager (Richard Machemba), the TAZAMA Project principal investigator (Mark Urassa), and I had access to the unencrypted data. All data captured from the clinic logbooks, and the TAZAMA Project data (i.e., HDSS and HIV sero-surveys) were also stored in the TAZAMA Project data room.

3 Paper A. Point-of-contact Interactive Record Linkage (PIRL): A software tool to prospectively link demographic surveillance and health facility data

Christopher T. Rentsch[1], Chodziwadziwa Whiteson Kabudula[2], Jason Catlett[3], David Beckles[4], Richard Machemba[5], Baltazar Mtenga[5], Nkosinathi Masilela[2], Denna Michael[5], Redempta Natalis[6], Mark Urassa[5], Jim Todd[1,5], Basia Żaba[1], Georges Reniers[1,2]

[1]Department of Population Health, London School of Hygiene & Tropical Medicine, London, WC1E 7HT, UK
[2]MRC/Wits Rural Public Health and Health Transitions Research Unit (Agincourt), School of Public Health, Faculty of Health Sciences, University of the Witwatersrand, Johannesburg, 2193, South Africa
[3]Independent Consultant, SELECT Star, Atlanta, GA, 30309, USA
[4]Independent Consultant, London, UK
[5]The TAZAMA Project, National Institute for Medical Research, Mwanza, Tanzania
[6]District Medical Officer, Ministry of Health Tanzania, Magu District, Tanzania

London School of Hygiene & Tropical Medicine
Keppel Street, London WC1E 7HT
www.lshtm.ac.uk

Registry
T: +44(0)20 7299 4646
F: +44(0)20 7299 4656
E: registry@lshtm.ac.uk

LONDON
SCHOOL *of*
HYGIENE
&TROPICAL
MEDICINE

RESEARCH PAPER COVER SHEET

PLEASE NOTE THAT A COVER SHEET MUST BE COMPLETED FOR EACH RESEARCH PAPER INCLUDED IN A THESIS.

SECTION A – Student Details

Student	Christopher T. Rentsch
Principal Supervisor	Professor Basia Żaba and Dr. Georges Reniers
Thesis Title	Point-of-contact interactive record linkage between demographic surveillance and health facilities to measure patterns of HIV service utilisation in Tanzania

If the Research Paper has previously been published please complete Section B, if not please move to Section C

SECTION B – Paper already published

Where was the work published?	*Gates Open Research*		
When was the work published?	Jan 2018		
If the work was published prior to registration for your research degree, give a brief rationale for its inclusion			
Have you retained the copyright for the work?*	Yes	Was the work subject to academic peer review?	Yes

**If yes, please attach evidence of retention. If no, or if the work is being included in its published format, please attach evidence of permission from the copyright holder (publisher or other author) to include this work.*

SECTION C – Prepared for publication, but not yet published

Where is the work intended to be published?	
Please list the paper's authors in the intended authorship order:	
Stage of publication	

SECTION D – Multi-authored work

For multi-authored work, give full details of your role in the research included in the paper and in the preparation of the paper. (Attach a further sheet if necessary)	I contributed to the conception of the investigation, designed the study, created analysis plans, extracted and cleaned data, conducted the analysis, drafted the manuscript, collated feedback from co-authors, submitted manuscript, and liaised with journal editors

Student Signature: [redacted] ____________ **Date:** 8/8/2018

Supervisor Signature: [redacted] ____________ **Date:** 8/8/2018

3.1 Overview

In Section 2.4, the PIRL software used to collect the data for this PhD research was introduced. In this chapter, the rationale, implementation, operation, and system requirements for the PIRL software are described in further detail. In addition, a full user guide, including how to install and use the PIRL software, is available online in an open-source repository.[72] The software was published with an MIT license, which allows others to download, edit, and use the software in any way as long as they provide attribution to the license holders. This paper also presents anonymised versions of the data infrastructure created by PIRL using multiple case studies based on interactions I had in the field.

Objective 1. To implement a locally-relevant approach to link community cohort data with medical records from three facilities offering HIV services.

3.2 ABSTRACT

Linking a health and demographic surveillance system (HDSS) to data from a health facility that serves the HDSS population generates a research infrastructure for directly observed data on access to and utilization of health facility services. Many HDSS sites, however, are in areas that lack unique national identifiers or suffer from data quality issues, such as incomplete records, spelling errors, and name and residence changes, all of which complicate record linkage approaches when applied retrospectively. We developed Point-of-contact Interactive Record Linkage (PIRL) software that is used to prospectively link health records from a local health facility to an HDSS in rural Tanzania. This prospective approach to record linkage is carried out in the presence of the individual whose records are being linked, which has the advantage that any uncertainty surrounding their identity can be resolved during a brief interaction, whereby extraneous information (e.g., household membership) can be referred to as an additional criterion to adjudicate between multiple potential matches. Our software uses a probabilistic record linkage algorithm based on the Fellegi-Sunter model to search and rank potential matches in the HDSS data source. Key advantages of this software are its ability to perform multiple searches for the same individual and save patient-specific notes that are retrieved during subsequent clinic visits. A search on the HDSS database (n=110,000) takes less than 15 seconds to complete. Excluding time spent obtaining written consent, the median duration of time we spend with each patient is six minutes. In this setting, a purely automated retrospective approach to record linkage would have only correctly identified about half of the true matches and resulted in high linkage errors; therefore highlighting immediate benefit of conducting interactive record linkage using the PIRL software.

3.3 Introduction

The amount of collected data is ever-increasing in various sectors, including healthcare and government administration. While each individual data source holds value and was likely created for a specific purpose, researchers could study more complex relationships by combining data sources holding information on the same entity or individual. A recent Wellcome Trust report detailed how record linkage – the matching of an individual's records between two or more data sources – adds to the value of medical research in low- and middle-income as well as high-income countries.[15] Broadly, record linkage can increase the range of questions that could be asked, provide a historical perspective necessary for some studies, improve the statistical properties of analyses, and make better use of resources.

The statistical framework for record linkage was largely developed in the 1950s[28] and 1960s.[29] Two popular methods of record linkage have been used to combine data sources. Deterministic record linkage[24] is a rule-based approach that typically requires exact matching on a set of identifiers existing in all data sources. Probabilistic methods[25-27] can be employed to assign weights based on the (dis)similarity of identifiers (e.g., name, sex, and date of birth) between records.

In the United Kingdom, researchers use record linkage to merge the Clinical Practice Research Datalink – one of the largest databases of longitudinal medical records from primary care in the world – to a variety of other existing data sources that hold data on cardiovascular and cancer events, hospitalisation, and mortality.[16] Publications using this data infrastructure cover a vast range of topics, including studies showing the absence of an association between measles, mumps, and rubella (MMR) vaccine and autism,[17] cardiovascular risk after acute infection,[18] and the association between body mass index and cancer.[19]

Located in several low- and middle-income countries, health and demographic surveillance systems (HDSS) are effective and comprehensive data collection systems that primarily measure the fertility, mortality, and other self-reported health information of an entire population. However, such self-reports usually lack detail and accuracy about the clinical events and services received, and their retrospective nature means they quickly become dated. Linking an HDSS database to data from a health facility that serves the HDSS population produces a research infrastructure for generating directly observed data on access to and utilization of health facility services.[14]

Many HDSS sites, contrary to record linkage studies conducted in high-income countries, are in areas that lack unique national identifiers or suffer from data quality issues, such as incomplete records, spelling errors, and name and residence changes, all of which complicate both deterministic and probabilistic approaches when applied retrospectively. In these settings, a semi-automatic record linkage process that incorporates manual inspection of potential matches, such as interactive record linkage,[40, 41] is preferred. In our implementation of interactive record linkage, which we call point-of-contact interactive record linkage (PIRL), we carry out the manual inspection of potential matches identified by our linkage algorithm in the presence of the individual whose records are being linked. This prospective approach to record linkage has the advantage that any uncertainty surrounding their identity can be resolved during a brief interview, whereby extraneous information (e.g. household membership) can be referred to as an additional criterion to adjudicate between multiple potential matches. It also provides an opportunity to authenticate individuals who can legitimately be linked to more than one record in the HDSS because they have resided in more than one household. Finally, ethical and privacy concerns are properly addressed with PIRL as it offers an advantage to seek informed consent and individuals are made fully aware of how their data are being used.

There are numerous publicly and commercially available record linkage software packages. Herzog et al.[78] adapted a comprehensive checklist[105] for evaluating record linkage software, including questions regarding the amount of control the user has over the record linkage methodology, data management and standardisation, and post-linkage functions (see Appendix 10.8). Many of the available software packages are designed for batch linkages, such as those used in purely automated retrospective linkage.[106, 107] Given the novelty of the PIRL approach where searches are individually supervised, we opted to build our own software package to suit our specific needs. By designing our own software, we maintained full control over the specification of the linkage algorithm, including the match parameters, weights, agreement rules, string comparators, and how to handle missing data. We also required the ability to save session-specific notes that can be retrieved in future linkage sessions.

We introduced our PIRL software to prospectively link health records to HDSS records in a rural ward in northeast Tanzania. An analysis of the data created by our implementation of the software and how it compares to purely automated retrospective linkage has previously been published.[74] This paper describes our implementation of this software, and we attach a GitHub link[72] to the full source code for others to download and amend to their own research needs.

3.4 Methods

3.4.1 Data sources

The Kisesa observational HIV cohort study was established in 1994 and is located in a rural ward in the Magu district of Mwanza region in northwest Tanzania. It comprises demographic surveillance carried out through household interviews and population-based HIV surveillance based on individual serological tests and interviews. The HDSS databases include biannual rounds (31 to date) of household-based surveys that collect information on births, pregnancies, deaths, in- and out-migration, and spousal and parent-child relationships. One major weakness of the Kisesa HDSS is the lack of reconciling records of individuals who move households within the HDSS area. Therefore, while an HDSS ID is unique to a single individual, some individuals may have multiple HDSS IDs if they resided in more than one household in the HDSS area since the start of the HDSS in 1994. There have been eight rounds of HIV surveillance conducted every three years, with a detailed questionnaire on sexual behaviour and partnership factors, fertility outcomes, HIV-related knowledge, and use of health services. Individuals who participate in an HIV surveillance round are given a unique identifier, and their current unique identifier from the HDSS is also cross-referenced on their record.

A government-run health centre is situated in the Kisesa HDSS catchment area. Three clinics located in the Kisesa Health Centre were initially targeted as record linkage sites: the HIV care and treatment centre (CTC), the HIV testing and counselling clinic (HTC), and the antenatal clinic (ANC) which includes prevention of mother-to-child transmission services; all of which operate according to national guidelines and protocols. The CTC databases have been fully digitised, and data clerks regularly update and run data checks on these data. For the ANC and HTC clinics, we developed electronic data capture systems and digitised the paper-based logbooks.

3.4.2 Implementation

Our computer software utilises a probabilistic search algorithm to identify and rank potential matches in the HDSS database (n=110,000). The algorithm incorporates the following parameters or data fields: up to three names for the individual; sex; year, month, and day of birth; village and sub-village; up to three names of a household member; and up to three names for the ten-cell leader of the patient. A ten-cell leader is an individual who acts as a leader for a group of ten households and these positions have been relatively stable over time. The algorithm used for searching possible matches and ranking them is based on the Fellegi-Sunter record linkage model,[28, 29] with match probabilities (m_i) that have been adopted from a pilot study in the Agincourt HDSS.[38] The u_i probabilities, defined as chance agreement between two records which are true non-matches, were derived from the Kisesa HDSS data consistent with previous literature.[27] Let M be a set of true matches and U be a set of true non-matched record pairs. Two individual agreement probabilities are defined for each field i in record pair j as follows:

$$\text{match probability: } m_i = \text{P(field } i \text{ agrees} \mid j \in M) \tag{3.1}$$

$$\text{unmatch probability: } u_i = \text{P(field } i \text{ agrees} \mid j \in U) \tag{3.2}$$

For a given field with match probability m_i and unmatch probability u_i, the software calculates the matching weights as w_{ai} = $\log_2[m_i/u_i]$ for fields where both datasets agree, and w_{di} = $\log_2[(1\text{-}m_i)\ /(1\text{-}u_i)]$ where they disagree. Assuming independence of observations across the fields, the match score is computed by summing the weights across all fields.[29, 78]

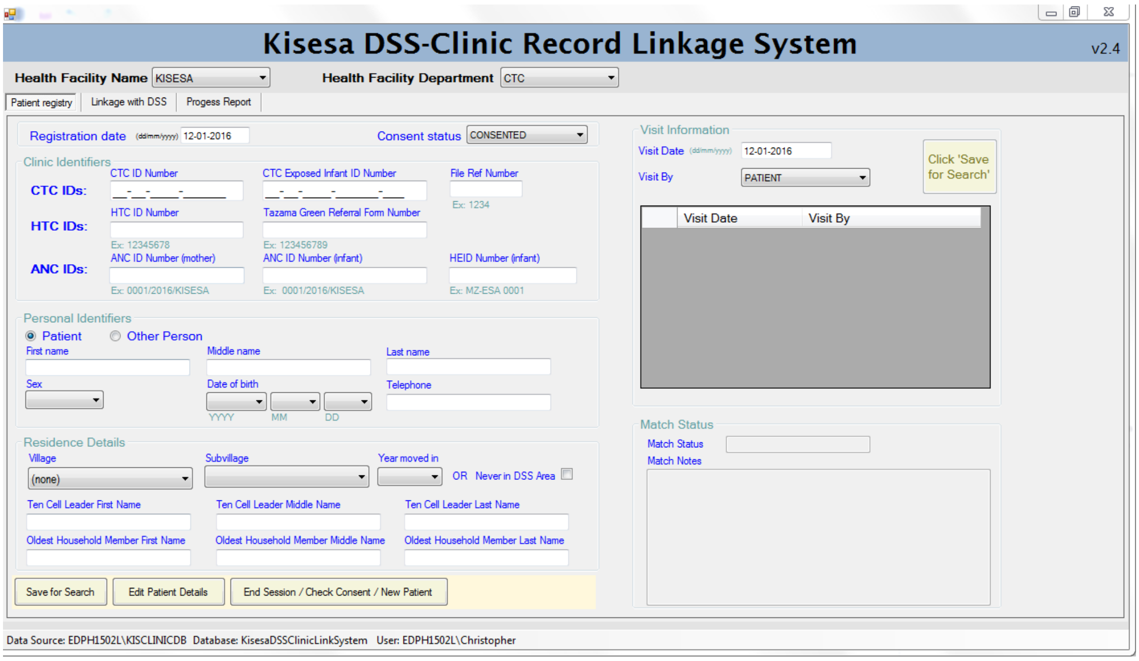
Agreement conditions vary for each of the parameters. Spelling errors, the use of more than one name (including nicknames), and interchangeable name order complicate locating an exact match between names in these databases; thus, the linkage algorithm allows for all pairwise comparisons between reported names and names found in the HDSS. In addition, the software uses a Jaro-Winkler string comparator approach to compare the name fields between the two data sources.[101] Previous research has shown the Jaro-Winkler method produces similar results to Double Metaphone and Soundex string comparators in a southern African context.[38] A Jaro-Winkler score ≥0.8 was considered a match for each collected name. Sex, village, and sub-village required an exact match, while the year of birth could differ by up to two years.

3.4.3 Operation

A full user guide including screen shots and step-by-step instructions on how we operationalise this software is attached (Supplementary File 1). Briefly, as individuals arrive to any of the target clinics, a fieldworker introduces him/herself and then invites the attendee to take part in the linkage study, which involved a brief interview. The primary goals of the brief interview are to explain the study, seek informed consent, and identify the HDSS records of all participants with a residency history in the HDSS.

Our team uses a dedicated desk located within the clinic, but out of the way of normal clinic operations, to conduct the brief interviews, and therefore did not interrupt or interfere with clinical practice. While we highly recommend ensuring privacy during each patient interaction, the interview only involves asking for demographic information, such as name, sex, birthdate, and residence details, and does not ask for any medical information. In addition, all collected data from a previous session is cleared from the system at the end of each patient interaction. Therefore, to enhance the accuracy of the data, we allow patients to watch their information be entered into the software and ask them to verify what has been collected.

The first step after obtaining written consent is to collect all clinic identifiers for the patient. The software uses these clinic identifiers to retrieve previously collected information and matches made on patients interviewed during a prior visit. After all clinic identifiers are collected, personal and residence details are entered into the system (Figure 3.1). Information from most of these fields contribute to the linkage algorithm described in the Implementation section above.



Kisesa DSS-Clinic Record Linkage System v2.4

Health Facility Name: KISESA
Health Facility Department: CTC

Registration date (dd/mm/yyyy): 12-01-2016
Consent status: CONSENTED

Patient ● Other Person ○

Village: (none)
OR Never in DSS Area ☐

Visit Date (dd/mm/yyyy): 12-01-2016
Visit By: PATIENT

Data Source: EDPH1502L\KISCLINICDB Database: KisesaDSSClinicLinkSystem User: EDPH1502L\Christopher

Figure 3.1: User interface of Point-of-contact Interactive Record Linkage (PIRL) software

Once all personal and residence details are entered, the user initiates an initial search through the HDSS data source. The software computes a match score for each record in the HDSS database, ranks them from highest to lowest based on match score, and outputs the top 20 records within 15 seconds. While manually searching through these potential matches, the user can view the full list of household members associated with each HDSS record. The user can then inquire with the patient to identify which HDSS record(s), if any, are a true match.

An important feature of this software is the ability to perform multiple search attempts for a single patient. If an initial search attempt does not result in a match, the user can further inquire into the possible use of nicknames, maiden names, or residency episodes at other addresses, and perform consecutive searches with this updated information. If one or more HDSS records are not found, the user can enter details of the missing records into a free-text field called “match notes.” These match notes are retrieved by clinic identifiers and can be used to guide interviews and searches during subsequent visits. When a clinic identifier is entered into the system that has already been collected, the software automatically displays the match status (e.g., matched, not matched) and saved matched notes to the user. The dates of all follow-up visits are automatically logged into the system.

Because we use this software in an area without reliable internet connectivity, we perform manual backups and syncs of the back-end data at the end of each working day as a way to mitigate any risk for loss of collected data. Full details on the import and export routines can be found in Annex 2 of the attached user guide (Supplementary File 1). Briefly, the data manager exports a backup file from each of the user’s machines using SQL Server Management Studio (SSMS). Then, the backup files are imported into SSMS on the data manager’s machine, and a SQL program automatically merges, updates, and collates the data collected from previous days. Finally, the data manager exports the combined backup file and imports it onto each of the user machines. Source code for these import and export routines can also be found on GitHub.

We employ data integrity checks within the software and on the back-end data. Due to the importance of clinical identifiers, all ID fields require double entry. Furthermore, HTC IDs are ensured through modulo-97 check digits, and ANC and CTC IDs have specific formats that the software confirms. The software also displays warning messages to the user if they attempt to match to a record that has an absolute difference in birth year of >10 years or the sum of the Jaro-Winkler name scores is ≤1.6.

To validate the matches in the back-end database, the lead author performs periodic and manual, back-end inspection of the data. These data integrity checks flag individuals who are matched to multiple HDSS records with large age differences (>10 years), of conflicting sex, within the same household, or with overlapping residency episodes in which one record’s start date occurred before another record’s end date. Over 18 months, only eight (0.2%) out of 3,456 matches were deemed unlikely and were deleted from the back-end database.

3.4.4 System requirements

The user interface (UI) portion of the software was coded using C# language in Microsoft Visual Studio 2013 Community edition. The database management system was coded in Microsoft SQL Server 2012 Express. The software has been developed for machines running a Windows 7 operating system.

Users who wish to edit source code to tailor the software to their specific needs will need both Visual Studio and SSMS. However, users who only need to run the software will need SSMS alone.

Full installation instructions can be found in Annex 1 of the attached user guide (Supplementary File 1).

3.5 USE CASES

3.5.1 Input dataset

Due to the nature of the software and its requirement for personally identifiable information, we are unable to provide real HDSS data used in our implementation of the software. However, we did create a dataset of 100 fake HDSS records that randomly sampled information found in the real data. Each field was sampled separately to break any links of information that could identify an individual. Spelling alterations, change of names, and other minor errors to birthdays or residence details were made to make the example cases described below more realistic to what we experience in the field. The data and a codebook for the fake input dataset are attached (Supplementary File 2). The script used to create the fake input dataset is also attached (Supplementary File 3).

3.5.2 Output datasets

The software creates four password-encrypted tables and stores them in SSMS. The first table, called the ‘Registry’, stores clinic identifiers, personal and residence details reported by the patient and entered by the fieldworker into the main view of the software (Figure 3.1). A new record is created for each search attempt. The second table, called ‘Matches’, stores all matches made to HDSS records, including the HDSS identifier, match score, and the rank of the match. The third table, called ‘Notes’, holds the collection of match notes made during an interview. The fourth table, called ‘Visits’, is a file containing all visit dates for each patient.

Three auto-generated identifiers are used to link records that pertain to a specific individual between the four back-end data tables: the local machine name, a session ID, and a record number. For each local machine, a session ID consisting of numerical values for year, month, day, hour, minute, and second gets automatically created at the beginning of a new session (e.g., ‘20170601093000’ for a session initiated at exactly 9:30:00am local time on 1 June 2017). Within each session, a six-digit record number is created and iterates for each search attempt within a session. Whenever a match is made (table 2), match notes are stored (table 3), or a visit date is recorded (table 4), the values for the machine name, session ID, and record number are stamped on those records.

An example output database from the cases below and its codebook are attached (Supplementary File 4).

3.5.3 Case 1

The patient enters the CTC and agrees to take part in this study. The fieldworker collects his CTC ID and enters it into the system along with the personal and residence details he reports (Table 3.1). The software displays the top 20 potential matches to the fieldworker. The fieldworker selects the top ranked record to view the entire household membership and confirms the reported co-resident is listed. There are minor spelling errors in the names, but the year of birth, years of residency, and residence details match exactly. Thus, the fieldworker assigns the match to this record and ends the search as all reported residency episodes were found. The fieldworker saves a match note that says, “All reported residency episodes found.” The fieldworker then stores the visit date and thanks the patient for his time.

Table 3.1. Personal identifiers used for three case patients with varying numbers of residency episodes

	Case 1	Case 2	Case 3		
Residency episode	1	1	1	2	3
Clinic ID(s)	CTC: 77-10-4545-253004	ANC: 1234/2017/KISESA HTC: 44447050	HTC: 44618061		
First name	PETER	PASTORY	SUZANNE	SUZANNE	SUZANNE
Second name	JAKKU	SWAKALA	LENARD	JONAS	JONAS
Third name		TIMOS	WILLIAMS	ZABRON	ZABRON
Sex	M	F	F	F	F
Year of birth	2004	1984	1980	1980	1980
Month of birth	8	9			
Day of birth	15				
Village	KANYAMA	KANYAMA	KISESA	Outside HDSS area	IHAYABUYAGA
Sub-village	CHANGABE	NYAN'HELELA	KISESA KATI		ILENDEJA
Residence start year	2012	2010	1995	2003	2006
Residence end year	2014	2014	2003	2006	2014
TCL first name[a]	HELENA	MICHAEL	MIZIMALLI		MABINA
TCL second name[a]	MSHIMO	MALIGANYA	NDALAHAWA		PALO
TCL third name[a]					
HH member first name	LUZALIE	JOSEPHI	KOYA		DOTTO
HH member second name	MATHIAS	BONIFASI	SAHANNI		SALU
HH member third name					
True HDSS ID[b]	22341597005	77537712004	10012368001	-	10025490004
True ID in fake input dataset	30	98	1	-	54

Abbreviations: ID - identifier; TCL - ten-cell leader; HH - household; HDSS - health and demographic surveillance system

[a] Ten-cell leader: a ten-cell leader is an individual who acts as a leader for a group of ten households and these positions have been relatively stable over time

[b] True HDSS ID of patient (found in fake input dataset), which is unknown in reality

3.5.4 Case 2

The patient enters the ANC and agrees to take part in the study. The fieldworker collects her ANC ID, but also notices she carries an HTC card, so they collect that information as well (these cross-clinic links are common in our fieldwork and allow us to link patient records across multiple services). The fieldworker also enters the personal and residence details she reports (Table 3.1). The software displays the top 20 potential matches to the fieldworker. The fieldworker selects the top ranked record to view the entire household membership and confirms the reported co-resident is listed. The years of residence are only off by one year, and the birth year and residence details match exactly. There are minor spelling mistakes in the names reported, but the reported names are switched in order on the HDSS record, which is not uncommon for the data in this setting. The fieldworker assigns the match to this record and ends the search as all reported residency episodes were found. The fieldworker saves a match note that says, “All reported residency episodes found.” The fieldworker then stores the visit date and thanks the patient for her time.

3.5.5 Case 3

The patient enters the HTC and agrees to take part in the study. The fieldworker collects her HTC ID and enters it into the system along with the personal identifiers she reports (Table 3.1). During the interview, she reports she had two residency episodes in different villages, one from 1995 to 2003 and the other from 2006 to 2014. The patient reports to have lived outside of the HDSS area between 2003 and 2006. The fieldworker enters the information for the most recent residency episode and initiates the search. The software displays the top 20 potential matches from the HDSS to the fieldworker. The fieldworker selects the top ranked record to view and confirm that the other household members are correct. There are minor spelling errors in the names and the year of birth is off by one year, but the residence details are the same, so the fieldworker assigns this record as a match.

The fieldworker continues moving down the list of potential matches and tries to find the record associated with the older residency episode. However, the fieldworker finishes going through the list without detecting the record. The fieldworker informs the patient that her record for the older residency episode was not found and asks if there was any reason why her personal details would have been different. She informs the fieldworker she was married in 2003 and provides her maiden name and the name of another household member for that episode. The fieldworker amends the personal details and attempts a second search. The fieldworker now finds the top ranked record to have a few spelling differences, but the years of residence, village, and birth year are all the same. Additionally, the household member is listed on the record. The fieldworker assigns the match to this record and ends the search as all reported residency episodes were found. The fieldworker saves a match note that says, “All reported residency episodes found.” The fieldworker then stores the visit date and thanks the patient for his time.

3.5.6 Return visits

When any of the case patients return to a linkage clinic, their clinic IDs when entered will retrieve the match status (in this case, “Matched’; if no matches were made, “Not matched”) and the saved match notes. In these cases, the fieldworker can quickly see no other searches are needed and can simply store the new visit date before thanking the patient again for their time. In the event a match note stated, “Missing a record for 2002–2007 in Kisesa Kati,” the fieldworker can focus the interview to obtain the personal details that were associated with that record.

3.6 CONCLUSIONS

The PIRL software – which combines a probabilistic search algorithm for identifying potential matches with a relatively simple human intervention – has shown promise for linking multiple data sources without a unique identifier in rural Tanzania. A key advantage of this software over other software that employ purely automated record linkage is the ability to perform multiple searches for the same individual. This is of importance for individuals whose records are more likely to contain out-of-date or inaccurate names or addresses, particularly for individuals with older residency episodes and women whose names change after marriage. Each search attempt on the HDSS database takes less than 15 seconds to complete. Excluding time spent obtaining written consent, the median duration of time we spend with each patient is six minutes.

A limitation of the search database in the current implementation of the software is that it can only be as current as the most recently completed HDSS round. In Kisesa, HDSS rounds are conducted for a few months roughly once per year, and extensive data cleaning delays the data availability by another few months. Therefore, recent residents, such as children and adults who first move into the HDSS area or infants born after the last HDSS round, will not have an HDSS record. The software allows the user to input the date of first residence in the HDSS area, so that these individuals can be flagged in subsequent analyses. During the first 18 months of operations in Kisesa, we flagged 1,576 (24.7%) patients as recent residents out of 6,376 clinic attendees who consented to the linkage study.

In this setting, a purely automated retrospective approach to record linkage would have only correctly identified about half of the true matches and resulted in high linkage errors, therefore highlighting immediate benefit of this prospective approach.[74] Linking health records to an HDSS database generates a rich data source of directly observed data on access to and utilization of health facility services at a subnational level.

3.7 Data and software availability

Software source code: https://github.com/LSHTM-ALPHAnetwork/PIRL_RecordLinkageSoftware

Archived source code as at time of publication: https://doi.org/10.5281/zenodo.998867

License: MIT

Due to ethical clearances, we are unable to share identifiable HDSS data or clinic identifiers used in our implementation of the software with anyone outside the study team. However, demographic data only for the HDSS are available via the INDEPTH Network's Sharing and Accessing Repository (iSHARE). Applications to access the anonymised data for collaborative analysis are encouraged and can be made by contacting the project coordinator for the Kisesa HDSS, Mark Urassa (urassamark@yahoo.co.uk), or by contacting the ALPHA Network team (alpha@lshtm.ac.uk).

3.8 Supplementary material

All supplementary material can be found online by clicking on the following links. They were too large to present in this document.

1. Supplementary File 1. Kisesa-HDSS record linkage user guide
    a. https://gatesopenresearch.s3.amazonaws.com/supplementary/12751/479f8ddd-3893-4b6e-8146-e056e438fc63.docx
2. Supplementary File 2. Fake input dataset with codebook
    a. https://gatesopenresearch.s3.amazonaws.com/supplementary/12751/f81a3410-eabf-4794-a2dd-8a5a27cc4d03.xlsx
3. Supplementary File 3. Script to create fake input dataset
    a. https://gatesopenresearch.s3.amazonaws.com/supplementary/12751/9e99fa11-65a4-47ae-9939-83db87f45191.txt
4. Supplementary File 4. Output datasets for case patients with codebook
    a. https://gatesopenresearch.s3.amazonaws.com/supplementary/12751/b5cd9ea1-0a64-4830-93b1-878644756409.xlsx

4 Paper B. Point-of-contact interactive record linkage (PIRL) between demographic surveillance and health facility data in rural Tanzania

Christopher T. Rentsch[1], Georges Reniers[1,2], Chodziwadziwa Kabudula[2], Richard Machemba[3], Baltazar Mtenga[3], Katie Harron[4], Paul Mee[5], Denna Michael[3], Redempta Natalis[6], Mark Urassa[3], Jim Todd[1,3], Basia Żaba[1]

[1]Department of Population Health, London School of Hygiene & Tropical Medicine, London, UK
[2]School of Public Health, University of the Witwatersrand, Johannesburg, South Africa
[3]The TAZAMA Project, National Institute for Medical Research, Mwanza, Tanzania
[4]Department of Health Services Research and Policy, London School of Hygiene & Tropical Medicine, London, UK
[5]MeSH Consortium, Faculty of Public Health and Policy, London School of Hygiene & Tropical Medicine, London, UK
[6]District Medical Officer, Magu District, Tanzania

London School of Hygiene & Tropical Medicine
Keppel Street, London WC1E 7HT
www.lshtm.ac.uk

Registry
T: +44(0)20 7299 4646
F: +44(0)20 7299 4656
E: registry@lshtm.ac.uk

LONDON
SCHOOL *of*
HYGIENE
&TROPICAL
MEDICINE

RESEARCH PAPER COVER SHEET

PLEASE NOTE THAT A COVER SHEET MUST BE COMPLETED FOR EACH RESEARCH PAPER INCLUDED IN A THESIS.

SECTION A – Student Details

Student	Christopher T. Rentsch
Principal Supervisor	Professor Basia Żaba and Dr. Georges Reniers
Thesis Title	Point-of-contact interactive record linkage between demographic surveillance and health facilities to measure patterns of HIV service utilisation in Tanzania

If the Research Paper has previously been published please complete Section B, if not please move to Section C

SECTION B – Paper already published

Where was the work published?	*International Journal for Population Data Science*		
When was the work published?	Dec 2017		
If the work was published prior to registration for your research degree, give a brief rationale for its inclusion			
Have you retained the copyright for the work?*	Yes	Was the work subject to academic peer review?	Yes

**If yes, please attach evidence of retention. If no, or if the work is being included in its published format, please attach evidence of permission from the copyright holder (publisher or other author) to include this work.*

SECTION C – Prepared for publication, but not yet published

Where is the work intended to be published?	
Please list the paper's authors in the intended authorship order:	
Stage of publication	

SECTION D – Multi-authored work

For multi-authored work, give full details of your role in the research included in the paper and in the preparation of the paper. (Attach a further sheet if necessary)	I contributed to the conception of the investigation, designed the study, created analysis plans, extracted and cleaned data, conducted the analysis, drafted the manuscript, collated feedback from co-authors, submitted manuscript, and liaised with journal editors

Student Signature: [redacted] **Date:** 8/8/2018

Supervisor Signature: [redacted] **Date:** 8/8/2018

4.1 OVERVIEW

In the previous chapters, I detailed the implementation of PIRL, including the field methodology and software, used in Kisesa to collect data for this PhD research. In this chapter, I report record linkage statistics, including match percentages and attributes associated with (un)successful linkage, from the first 18 months of primary data collection. This paper also includes a head-to-head comparison between PIRL and a fully automated linkage approach using the same linkage algorithm.

Objective 2. To identify individual characteristics associated with successful linkage using PIRL and compare PIRL with traditional, automated probabilistic record linkage.

4.2 Abstract

Introduction. Health and demographic surveillance systems (HDSS) have been an invaluable resource for monitoring the health status of populations, but often contain self-reported health service utilisation, which are subject to reporting bias.

Objectives. To implement point-of-contact interactive record linkage (PIRL) between demographic and health facility systems data, characterise attributes associated with (un)successful record linkage, and compare findings with a fully automated retrospective linkage approach.

Methods. Individuals visiting the Kisesa Health Centre were matched to their HDSS records during a short uptake interview in the waiting area of the health facility. The search algorithm was used to rank potential matches, from which the true match(es) were selected after consultation with the patient. Multivariable logistic regression models were used to identify characteristics associated with being matched to an HDSS record. Records matched based on respondent's clarifications were subsequently used as the gold-standard to evaluate fully automated retrospective record linkage by calculating sensitivity and positive predictive value (PPV).

Results. Among 2,624 individuals who reportedly lived in the HDSS coverage area, we matched 2,206 (84.1%) to their HDSS records. Characteristics associated with a higher odds of being matched were increased age (OR 1.07, 95% CI 1.02, 1.12; per 5-year increment), a later consent into the study (OR 2.07, 95% CI 1.37, 3.12; in the most recent six-month period), and fieldworker level of experience. The main drivers of the linkage algorithm were name, sex, year of birth, village, sub-village, and household member name. At the lowest match score threshold, automated retrospective linkage would have only correctly identified and linked 55% (1440/2612) of the records with a PPV of 55% (1440/2612).

Conclusion. Where resources are available, PIRL is a viable approach to link HDSS and other administrative data sources that outperforms purely retrospective approaches.

4.3 Introduction

Most analyses of health service use are limited to databases of patients enrolled in clinical care. These analyses lack a population perspective on service utilization, clinical outcomes, survival status, and patients who are lost to follow-up. In contrast, health and demographic surveillance systems (HDSS) comprehensively measure vital events but rely on self-reports of health services use. Such reports usually lack detail and accuracy about the clinical events and services received, and their retrospective nature means that they quickly become dated. Linking an HDSS database to data from a health facility that serves the HDSS population produces a nascent research infrastructure for generating directly observed data on access to and utilization of health facility services at the subnational level.[14] The linked clinical data could also be used to validate or substitute the self-reported health status and health service use data collected in the HDSS surveys.

Two popular methods of record linkage have been established, deterministic[24] and probabilistic,[25-27] to combine data sources holding different information on the same individual. Deterministic record linkage is a rule-based approach that usually requires exact matching between one or more identifiers existing in all data sources. However, when common unique identifiers are not available, probabilistic methods can be employed to assign weights based on the (dis)similarity of components (e.g., name, sex, and date of birth) between records. Few studies exist linking demographic surveillance and health facility data on the African continent, which is likely due to the lack of electronically-available clinic data and the limited number of shared variables collected in both data sources. Nevertheless, there are studies that suggest record linkage is feasible in some African settings. In Namibia, three databases – clinical, pharmaceutical, and laboratory – were retrospectively linked using patient name, sex, date of birth, and facility name; however, substantial missing data limited the success of the linkage to between 58% and 76% of records being matched.[37] In South Africa, a mix of deterministic (South Africa has a national identification number system) and probabilistic methods was employed to retrospectively link local health facility data to HDSS data with 88% of records being matched, which suggests linkage between these two data sources is achievable.[38]

Many HDSS sites, however, are in areas that lack unique national identifiers or suffer from data quality issues, such as incomplete records, spelling errors, and name and residence changes, all of which complicate both deterministic and probabilistic approaches when applied retrospectively using fully automated software. In these settings, ‘point-of-contact interactive record linkage’ (PIRL) can be used to improve matching rates and quality. This prospective approach to record linkage is conducted in the presence of the individual whose records are being matched, which contrasts with the more conventional approach where record linkage is done retrospectively. PIRL has the advantage that uncertainty surrounding their identity can be resolved during a brief interaction whereby extraneous information (e.g. household membership) can be referred to as an additional criterion to adjudicate between multiple possible matches. It also provides an opportunity to authenticate individuals who can legitimately be linked to more than one record in the HDSS because they have been resident in more than one household.

We introduced a PIRL system to link HDSS records with a local health facility that serves the HDSS population with the goal of producing a data source that could be used to monitor the utilisation of health services and the outcomes of patients after they have made contact with the health system. In this manuscript, we report on initial record linkage statistics, characterise patient and fieldwork attributes associated with (un)successful record linkage, and compare our findings with a fully automated linkage approach.

4.4 METHODS

4.4.1 Data sources

The Kisesa observational HIV cohort study was established in 1994 and is located in a rural ward in the Magu district of Mwanza region in northwest Tanzania. It comprises demographic surveillance carried out through household interviews that allow proxy reporting, and population-based HIV surveillance based on individual serological tests and interviews. The HDSS databases include biannual rounds (31 to date) of household-based surveys that collect information on births, pregnancies, deaths, in- and out-migration, and spousal and parent-child relationships. One major weakness of the Kisesa HDSS data is the lack of reconciling records of individuals who move households within the HDSS area. Therefore, some individuals may have multiple HDSS records if they resided in more than one household in the HDSS area since the start of the HDSS in 1994. There have been eight rounds of HIV surveillance conducted every three years, with a detailed questionnaire on sexual behaviour and partnership factors, fertility outcomes, HIV-related knowledge, and use of health services. Individuals who participate in an HIV surveillance round are given a unique identifier, and their current household-based identification from the HDSS is also cross referenced on their record.

A government-run health centre is located within the Kisesa HDSS catchment area. Three clinics located in the Kisesa Health Centre were initially selected as record linkage sites: the HIV care and treatment centre (CTC), the HIV testing and counselling clinic (HTC), and the antenatal clinic (ANC) which includes prevention of mother-to-child transmission (PMTCT) services; all of which operate according to national guidelines and protocols. The CTC databases have been fully digitised, and data clerks regularly update and run data checks on these data. For the ANC and HTC clinics, we developed electronic databases and digitised the paper-based logbooks using a double-entry system where two different fieldworkers independently capture each book, and any discrepancy between fields are reconciled in a cleaning stage.

4.4.2 Field team

Fieldwork started in Kisesa Health Centre on 1 June 2015 and results presented in this paper include all data collected through 31 December 2016. At the beginning of the study, the study team was comprised of four fieldworkers, one of whom had previous experience with management of health facility and HDSS data (fieldworker 1) and three others who had experience with management of health facility data only (fieldworkers 2, 3, and 4). Before the initial rollout of the software in June 2015, all fieldworkers and the field manager were provided formative training by the first author. The training session included instructions on how to obtain informed consent and conduct brief interviews and several demonstrations of the software. Fieldworkers who were hired after the initial rollout of the software were trained by the field manager and existing fieldworkers through shadowing and close oversight for at least one month before working on their own.

During the first four months, fieldworkers 1, 2, and 3 were assigned to a single clinic. Beginning in October 2015, the fieldworkers rotated between clinics. At any time over the study period, fieldworker 4 would substitute for any of the three primary fieldworkers in case of any absences. In July 2016, fieldworker 3 was replaced by a new hire (fieldworker 5) who had limited experience with health facility data and HDSS data.

4.4.3 Interview process

The population of interest in this research included all individuals who attended any of these three clinics. There were no restrictions based on age; if a patient was less than 18 years of age, s/he was required to have a parent or legal guardian present. Informed written consent was obtained from all individuals who participated in this project. As individuals arrived at the clinics, a fieldworker introduced him/herself and then described the study. The fieldworker then invited the attendee to a desk located within the clinic but out of the way of normal clinic operations to conduct the brief interactive record linkage interview. The primary goals of the interview were to identify the true HDSS record(s) and to confirm residence histories of all participants using computer software developed for this project (available open source: https://doi.org/10.5281/zenodo.998867).[72]

Our computer software utilises a probabilistic search algorithm to identify and rank potential matches in the HDSS database. The algorithm incorporated the following parameters or data fields: up to three names for the individual; sex; year, month, and day of birth; village and sub-village; up to three names of a household member; and up to three names for the ten-cell leader of the patient. A ten-cell leader is an individual who acts as a leader for a group of ten households and these positions have been relatively stable over time. The algorithm used for searching possible matches and ranking them is based on a the Fellegi-Sunter record linkage model,[28, 29] with match probabilities (m_i) that have been adopted from a similar study in the Agincourt HDSS.[38]

Let M be a set of true matches and U be a set of true non-matched record pairs. Two individual agreement probabilities were defined for each field i in record pair j as follows:

$$\text{match probability: } m_i = \text{P(field } i \text{ agrees} \mid j \in M) \qquad (4.1)$$

$$\text{unmatch probability: } u_i = \text{P(field } i \text{ agrees} \mid j \in U) \qquad (4.2)$$

The higher the ratio m_i/u_i, the more useful a field was for matching purposes. For a given field with match probability m_i and unmatch probability u_i, we calculated the matching weights as w_{ai} = $\log_2[m_i/u_i]$ for fields where both datasets agree, and w_{di} = $\log_2[(1\text{-}m_i)/(1\text{-}u_i)]$ where they disagree. Assuming independence of observations across the fields, we computed the match score by summing the weights across all fields with collected information.[29, 78] Incomplete fields did not add or subtract from the match score.

Agreement conditions varied for each of the parameters and match probabilities were calculated using an expectation-maximisation algorithm (Supplemental Table 1 in Appendix 10.9.1). Spelling errors and the use of more than one name (including nicknames) complicated locating an exact match between any two names in these databases. We used the Jaro-Winkler string comparator approach to compare the name fields between two records.[101] Previous research has shown the Jaro-Winkler method produces similar results to Double Metaphone and Soundex string comparators in a southern African context.[38]

The software computed a match score for each record in the HDSS database, ranked them from highest to lowest match score, and output the top 20 records. Our decision to display 20 records was guided by the pilot phase of the software in November 2014. During the pilot phase, no matches were found beyond the first 20 record-pairs with the highest match scores.

While searching through these potential matches, the fieldworker could view the full list of household members associated with each HDSS record. The fieldworker then inquired with the patient to identify which HDSS record(s), if any, were a true match. The software displays warning messages to the fieldworkers if they attempt to match to a record that has an absolute difference in birth year of >10 years or the sum of the Jaro-Winkler name scores was ≤1.6. If the first search attempt did not result in a match or the individual reported multiple residency episodes, the fieldworker performed another search using updated identifying information obtained during the brief interview. The software does not have a limit on the number of searches a fieldworker can make and each search takes less than 15 seconds to output potential matches.

4.4.4 Review of matches

Matches selected during the interviews were assumed to be true matches. If no HDSS record was found, the fieldworker saved relevant information in a free-text field, "match notes," regarding likely reasons why the search did not result in a match. During the pilot phase of the software in November 2014, we learned the most likely reasons for not finding a match were having no residence history in the HDSS coverage area and migrating into the area or born after the last HDSS round. The software was adapted to flag these individuals and they were excluded from the analysis.

The lead author performed periodic and manual, back-end inspection of the data to verify the matches made in the field. These data integrity checks flagged individuals who were matched to multiple HDSS records with large age differences (>10 years), of conflicting sex, within the same household, or with overlapping residency episodes in which one record's start date occurred before another record's end date. Over the study period, eight matches were deemed unlikely and were deleted for this analysis.

4.4.5 Privacy

All interactions with the software are logged and labelled with a unique username for each fieldworker. The data collected with the linkage software includes personal identifiers used by the linkage algorithm, clinic identifiers, and visit dates. No medical information is captured or stored in the record linkage software. Data are stored on password-protected laptops and in an encrypted form. Once a fieldworker ends a session with a patient, they cannot access the unencrypted data. At the end of each working day, a data manager collates the data collected on each laptop and performs a backup of the database.

4.4.6 Statistical analyses

We calculated the overall match percentage as the proportion of patients who were matched to at least one HDSS record (numerator) out of the number of patients who claimed residence history in the HDSS area (denominator). We excluded patients who reported no residence history in the HDSS area – either the patient reported never to have lived in the HDSS catchment area, or they recently moved into the area or were born after the last HDSS round, or both. The match percentages were then stratified by clinic and patient characteristics. Patient characteristics included sex, age, whether their sub-village was on a tarmac road, type of residence (e.g., rural, peri-urban, or urban), date of first visit, and which fieldworker performed the initial interview and search. For patients seen in the HIV testing and counselling clinic, we also stratified the match percentage by their HIV status as determined by the result of the HIV test they had on the day they consented to PIRL. Chi-square (χ^2) tests were used to assess if the match percentage differed by the patient characteristics or between the three clinics.

Multivariable logistic regression models were used to identify patient and fieldwork attributes that were associated with a successful match to an HDSS record. Variables were included in the model if their bivariate association with the outcome was significant at the $p<0.2$ level. A two-way interaction term between date of first visit and fieldworker was explored but not significant ($p=0.4$). Guided by the Akaike information criterion (AIC), the best fitting model included a transformed variable for age (per 5-year increase). The regression models were stratified by clinic.

The utility of the matching parameters in the linkage algorithm was explored by calculating two metrics among search attempts that resulted in a match. First, we calculated the proportion of all searches that included a non-missing value for each parameter (% collected). Second, we calculated the proportion of times where the collected information agreed with the information in the matched record (% agreement). For example, year of birth was collected for 99% of searches and agreed with the year of birth (±2 years) on the matched record 87% of the time.

4.4.7 Automated linkage

We performed a fully automated probabilistic record linkage approach using the same algorithm used in the PIRL software to understand how the algorithm would have performed in a non-interactive setting. There are many detailed sources of how to perform retrospective record linkage.[27, 76-79] Briefly, a patient registry database of all matched participants in this study was created containing the collected information for the matching parameters (including records with incomplete information) and a variable for the participants' true HDSS ID. If multiple search attempts were made on an individual, the information collected for the first search attempt was used. If an individual was matched to more than one HDSS record, the HDSS record associated with the most recent residency dates was flagged as the sole true match. A match score was calculated for all pairwise comparisons between the patient registry (n=2,612) and the full HDSS database (n=90,996). The HDSS record with the highest match score was selected for each record in the patient registry.

When performing retrospective linkage, a match score threshold is selected to determine what constitutes a link versus a non-link. The placement of the threshold can be a matter of trial and error.[34] Additionally, a match score is not a standardised metric and can be greatly influenced by the number of parameters used. For this analysis, various thresholds of percentiles were selected based on the distribution of match scores among true matches (Supplementary Figure 1 in Appendix 10.9.1). There are four possible outcomes from retrospective record linkage: true links (true positives), true non-links (true negatives), false matches (false positives), and missed matches (false negatives) (Figure 4.1). Using an epidemiologic perspective, sensitivity of a linkage algorithm was defined as the proportion of true matches that were linked, positive predictive value (PPV) was the proportion of links that were true matches, and the false match rate was the proportion of true non-matches that were linked (the inverse of PPV).[27, 77] Initially, the same 'full' algorithm used in the PIRL software was used for automated retrospective linkage. A sensitivity analysis was carried out to determine the effects of limiting the algorithm to only commonly collected and high-performing parameters identified in this manuscript.

		True match status		
		Match	**Non-match**	
Link status	**Link**	True links (TP)	False matches (FP)	Total links
	Non-link	Missed matches (FN)	True non-links (TN)	Total non-links
		Total matches	Total non-matches	Total record pairs

Figure 4.1: Classification diagram of record linkage outcomes against true match status. Abbreviations: TP = true positives; FP = false positives; FN = false negatives; TN = true negatives. Common calculations: sensitivity = TP/(TP+FN); positive predictive value = TP/(TP+FP); false match rate = FP/(FP+TN)

Statistical analyses were performed using SAS version 9.4 (SAS Institute Inc., Cary, NC, USA). Ethical approval was obtained from the Lake Zone Institutional Review Board (MR/53/100/450), Tanzanian National Research Ethics Review Committee, and the London School of Hygiene & Tropical Medi-cine (LSHTM #8852).

4.5 Results

4.5.1 Sample population

Between 1 June 2015 and 31 December 2016, we consented and conducted brief interviews with 6,376 clinic attendees, which was a median 14 new patients per day (interquartile range (IQR): 9-20). Excluding time spent obtaining written consent, the median duration of time spent using the software to search for potential matches was 6 minutes (IQR: 2-21 minutes). Among the 6,376 patients, 2,206 (34.6%) reported they had never lived in the HDSS coverage area, and 1,576 (24.7%) were recent residents (either born or moved into the area after the last HDSS round) (Table 4.1). Thus, 2,624 patients reported residence history in the HDSS area and were considered likely to have a record in the community database.

Table 4.1: Exclusion criteria among point-of-contact interactive record linkage (PIRL) participants in rural Tanzania by clinic, n=6,376

Exclusion criteria	**Overall** (n=6,376)	**CTC** (n=1,318)	**ANC** (n=2,583)	**HTC** (n=2,480)	***P*[a]**
Total excluded	3,752 (58.9)	762 (57.8)	1,298 (50.3)	1,692 (68.4)	<0.0001
Never lived in HDSS area	2,206 (34.6)	642 (48.7)	393 (15.2)	1,171 (47.3)	<0.0001
Recently born or moved into HDSS area	1,576 (24.7)	126 (9.6)	915 (35.4)	535 (21.6)	<0.0001

Abbreviations: CTC - HIV care and treatment centre; ANC - antenatal clinic; HTC - HIV testing and counselling clinic; HDSS - health and demographic surveillance system

Note: all statistics are given in n(%)

[a]Clinic differences tested for statistical significance with chi-square (χ^2) tests

4.5.2 Match statistics

Of the 2,624 patients who reported residence history in the HDSS area, 2,206 (84.1%) were matched to one or more HDSS records (Table 4.2). By clinic, the match percentage was 86.0% in the CTC, 83.8% in the ANC, and 83.1% in the HTC (p=0.36). Overall, the match percentage did not differ by sex (84.2% among females vs. 83.6% among males; p=0.72) (Table 2). Patients who were older had higher match percentages than their younger counterparts (89.2% among 50+ years vs. 83.4% among 15-49 years and 86.2% among <15 years, respectively; p=0.04). Additionally, patients who resided in a sub-village that had no road or was rural, were first seen after August 2015, or were interviewed by fieldworkers 1, 2, or 3 (three of the original fieldworkers) had higher match percentages than those who resided in a sub-village that had a road or was urban, were first seen in the first three months of the study, or were interviewed by fieldworkers 4 or 5 (less experienced fieldworkers) (all p<0.005). Many of these associations were upheld in the stratified analyses by clinic. However, in the CTC and HTC, there was no significant association between a patient's date of first visit and being matched. In the ANC, match percentages did not differ by age (88.8% among <15 years, 83.5% among 15-49 years, 66.7% among 50+ years; p=0.19) but did differ significantly by sex (84.2% among females vs. 70.0% among males; p=0.04). Of note, only 30 (2.3%) of individuals seen in the ANC were male, the high majority (n=28; 93.3%) of whom were children aged 6 years or younger, and only three women reported an age of 50+ years. Lastly, in the HTC, there was no statistical difference between the match percentages by HIV test result received on the day of consent to record linkage (83.5% among positives, 83.1% among negatives, and 84.2% among inconclusive/unknowns; p=0.99).

Table 4.2: Match percentages among eligible point-of-contact interactive record linkage (PIRL) participants in rural Tanzania, by patient characteristic and clinic, n=2,624

	Overall			**CTC**			**ANC**			**HTC**		
Characteristic	**Matched** (n=2,206)	**Not matched** (n=418)	***P*[a]**	**Matched** (n=478)	**Not matched** (n=78)	***P*[a]**	**Matched** (n=1,077)	**Not matched** (n=208)	***P*[a]**	**Matched** (n=651)	**Not matched** (n=132)	***P*[a]**
Sex												
Female	1,769 (84.2)	331 (15.8)	0.7181	307 (85.0)	54 (15.0)	0.4030	1,053 (84.2)	197 (15.8)	0.0446	409 (83.6)	80 (16.4)	0.6310
Male	433 (83.6)	85 (16.4)		170 (87.6)	24 (12.4)		21 (70.0)	9 (30.0)		242 (82.3)	52 (17.7)	
Age, years												
<15	131 (86.2)	21 (13.8)	0.0431	26 (81.3)	6 (18.8)	0.0369	87 (88.8)	11 (11.2)	0.1887	18 (81.8)	4 (18.2)	0.5896
15-49	1,836 (83.4)	365 (16.6)		329 (84.3)	61 (15.6)		985 (83.5)	195 (16.5)		522 (82.7)	109 (17.3)	
50+	231 (89.2)	28 (10.8)		122 (92.4)	10 (7.6)		2 (66.7)	1 (33.3)		107 (86.3)	17 (13.7)	
Sub-village of residence, has road												
Yes	1,318 (81.4)	302 (18.6)	<0.0001	227 (82.0)	50 (18.0)	0.0034	746 (82.0)	164 (18.0)	0.0027	345 (79.7)	88 (20.3)	0.0029
No	886 (88.9)	111 (11.1)		249 (90.6)	26 (9.5)		331 (88.7)	42 (11.3)		306 (87.7)	43 (12.3)	
Sub-village of residence, type												
Rural	703 (89.0)	87 (11.0)	<0.0001	212 (88.3)	28 (11.7)	0.3595	237 (89.1)	29 (10.9)	0.0084	254 (89.4)	30 (10.6)	0.0005
Peri-urban	696 (84.6)	127 (15.4)		140 (85.9)	23 (14.1)		380 (84.8)	68 (15.2)		176 (83.0)	36 (17.0)	
Urban	805 (80.2)	199 (19.8)		124 (83.2)	25 (16.8)		460 (80.8)	109 (19.2)		221 (77.3)	65 (22.7)	
Date of first visit												
Jun-Aug 2015	845 (81.5)	192 (18.5)	0.0050	303 (86.3)	48 (13.7)	0.4326	350 (78.8)	94 (21.2)	0.0014	192 (79.3)	50 (20.7)	0.1513
Sep-Dec 2015	503 (88.3)	67 (11.8)		118 (88.1)	16 (12.0)		228 (89.8)	26 (10.2)		157 (86.3)	25 (13.7)	
Jan-Jun 2016	503 (84.0)	96 (16.0)		33 (80.5)	8 (19.5)		299 (85.4)	51 (14.6)		171 (82.2)	37 (17.8)	

Table 4.2: Match percentages among eligible point-of-contact interactive record linkage (PIRL) participants in rural Tanzania, by patient characteristic and clinic, n=2,624

	Overall			CTC			ANC			HTC		
Characteristic	**Matched** (n=2,206)	**Not matched** (n=418)	***P*[a]**	**Matched** (n=478)	**Not matched** (n=78)	***P*[a]**	**Matched** (n=1,077)	**Not matched** (n=208)	***P*[a]**	**Matched** (n=651)	**Not matched** (n=132)	***P*[a]**
Jul-Dec 2016	355 (84.9)	63 (15.1)		24 (80.0)	6 (20.0)		200 (84.4)	37 (15.6)		131 (86.8)	20 (13.3)	
Fieldworker												
1 - originally trained	731 (86.7)	112 (13.3)	0.0001	412 (87.1)	61 (12.9)	<0.0001	196 (86.0)	32 (14.0)	0.3075	118 (86.1)	19 (13.9)	0.0237
2 - originally trained	951 (84.9)	169 (15.1)		46 (93.9)	3 (6.1)		747 (84.1)	141 (15.9)		156 (85.7)	26 (14.3)	
3 - originally trained	387 (82.2)	84 (17.8)		10 (66.7)	5 (33.3)		49 (76.6)	15 (23.4)		324 (83.5)	64 (16.5)	
4 - substitute	59 (69.4)	26 (30.6)		11 (52.6)	9 (47.4)		9 (90.9)	1 (9.1)		40 (71.4)	16 (28.6)	
5 - recently trained	89 (78.1)	25 (21.9)			[b]		75 (79.8)	19 (20.2)		13 (65.0)	7 (35.0)	
HIV test result at first visit												
Positive										106 (83.5)	21 (16.5)	0.9855
Negative		-			-			-		529 (83.1)	108 (17.0)	
Inconclusive/unknown										16 (84.2)	3 (15.8)	

Abbreviations: CTC - HIV care and treatment centre; ANC - antenatal clinic; HTC - HIV testing and counselling clinic; HDSS - health and demographic surveillance system; IQR - interquartile range

Note: all statistics are given in n (%), unless otherwise noted

[a] Statistical differences tested for significance with chi-square (χ^2), Fisher's Exact, or Wilcoxon Rank-Sum tests

[b] Recently hired fieldworker who had not yet worked in CTC

4.5.3 Logistic regression

The results from the multivariable logistic regression models largely agreed with the bivariate analyses. A multivariable model including all patients suggested that a five-year increase in age was associated with a 7% increase in the odds of being matched (odds ratio (OR) 1.07, 95% confidence interval (CI) 1.02, 1.12) (Table 4.3). In addition, patients who resided in a sub-village that had no road were 44% more likely to be matched than those who resided in a sub-village that had a road (95% CI 1.02, 2.03). Compared to the initial three months of linkage operations, patients who were first seen later in the study period were twice as likely to be matched (OR 2.07, 95% CI 1.37, 3.12 for first visits between July and December 2016). Lastly, patients who were consented by the substitute or recently trained fieldworker were significantly less likely to be matched than those who were consented by one of the originally trained fieldworkers (OR 0.30, 95% CI 0.18, 0.52 for fieldworker 4, and OR 0.36, 95% CI 0.20, 0.66 for fieldworker 5). There were no significant associations with being matched by sex or type of sub-village in the overall model.

Table 4.3: Results from multivariable logistic regression models estimating the associations between being matched to an HDSS record with various patient characteristics in rural Tanzania, overall and by clinic

	Overall	CTC	ANC	HTC
Characteristic	**OR (95% CI)**	**OR (95% CI)**	**OR (95% CI)**	**OR (95% CI)**
Sample size *(number missing)*	2,624 *(22)*	556 *(6)*	1,285 *(10)*	783 *(6)*
Sex				
Female	1	1	1	1
Male	0.89 (0.67, 1.17)	1.34 (0.77, 2.33)	**0.32 (0.13, 0.81)**	0.92 (0.61, 1.37)
Age, per 5-year increase	**1.07 (1.02, 1.12)**	**1.17 (1.06, 1.28)**	0.95 (0.87, 1.05)	1.07 (0.99, 1.16)
Sub-village of residence has road				
Yes	1	1	1	1
No	**1.44 (1.02, 2.03)**	**2.69 (1.22, 5.95)**	1.39 (0.86, 2.25)	0.95 (0.48, 1.85)
Sub-village of residence, type				
Rural	1.44 (0.97, 2.14)	0.62 (0.25, 1.52)	1.54 (0.87, 2.74)	**2.41 (1.10, 5.31)**
Peri-urban	1.13 (0.89, 1.53)	0.92 (0.47, 1.79)	1.21 (0.83, 1.76)	1.34 (0.78, 2.31)
Urban	1	1	1	1
Date of first visit				
Jun-Aug 2015	1	1	1	1
Sep-Dec 2015	**1.95 (1.43, 2.66)**	1.54 (0.75, 3.13)	**2.98 (1.79, 4.95)**	**2.26 (1.17, 4.36)**
Jan-Jun 2016	**1.44 (1.09, 1.91)**	1.20 (0.39, 3.65)	**2.03 (1.30, 3.17)**	**2.42 (1.17, 5.01)**
Jul-Dec 2016	**2.07 (1.37, 3.12)**	0.89 (0.23, 3.43)	**2.43 (1.23, 4.82)**	**5.15 (2.06, 12.89)**
Fieldworker who performed first search				
1 - originally trained	0.93 (0.70, 1.23)	0.44 (0.12, 1.70)	0.69 (0.41, 1.17)	1.03 (0.53, 2.00)
2 - originally trained	1	1	1	1
3 - originally trained	0.77 (0.56, 1.05)	**0.12 (0.02, 0.72)**	**0.47 (0.23, 0.95)**	1.84 (0.90, 3.79)
4 - substitute	**0.30 (0.18, 0.52)**	**0.12 (0.03, 0.61)**	1.09 (0.13, 9.46)	**0.45 (0.21, 0.96)**
5 - recently trained	**0.36 (0.20, 0.66)**	[a]	**0.43 (0.19, 0.97)**	**0.17 (0.05, 0.53)**
HIV test result at first visit				
Positive	-	-	-	0.94 (0.55, 1.62)
Negative				1
Inconclusive/unknown				0.82 (0.22, 2.99)

Abbreviations: HDSS - health and demographic sentinel surveillance; CTC - HIV care and treatment centre; ANC - antenatal clinic; HTC - HIV testing and counselling clinic; OR = adjusted odds ratio; CI = confidence interval; ref = referent category

Note: bolded OR (95% CI) are significant at a p<0.05 level

[a] Recently hired fieldworker who has not yet worked in CTC

In the multivariable analyses stratified by clinic, males were 68% less likely to be matched than females in the ANC (OR 0.32, 95% CI 0.13, 0.81); however, sex was not associated with being matched in the CTC or HTC. The association between increased age and being matched found in the overall model was stronger in the CTC model (OR 1.17, 95% CI 1.06, 1.28) and similar in the HTC (OR 1.07, 95% CI 0.99, 1.16); however, the association was not found in the ANC. Conversely, the increased odds of being matched later in the study period compared earlier in the study period was not found in the CTC, but still found in the ANC and HTC. Interestingly, a positive or inconclusive/unknown HIV test result was not associated with being matched (OR 0.94, 95% CI 0.55, 1.62 for positive result; OR 0.82, 95% CI 0.22, 2.99 for inconclusive/unknown result).

4.5.4 Linkage algorithm

PIRL performed well in this setting. In addition to the 2,206 matched individuals who reported they had a residency history in the HDSS area, HDSS records were also found for 406 (10.8%) of the patients who did not initially report a residence history in the HDSS area (the name “Kisesa” refers to a ward, a village within the ward, and a sub-village within that in which the health facility is located, which makes it conceivable that patients may report not living in Kisesa because they interpreted the question to mean village or sub-village rather than ward). Additionally, some of the individuals reported having multiple residency episodes within the HDSS area, thus qualifying them to have more than one HDSS ID record. In total, we matched 3,434 HDSS records to 2,612 individuals. We selected the HDSS record associated with the most recent residency dates for the remaining calculations. Of the 2,612 matches, 1,871 (71.6%) were ranked with the highest score by the search algorithm, and 306 (11.7%) were ranked with the second highest score. The remaining 435 (16.7%) matched records were ranked between third and twentieth by the computer algorithm. The mean match score was higher for matched records ranked first (mean match score 25.6, standard deviation (SD) 10.2) than matched records ranked second (mean match score 19.4, SD 9.5) or third and below (mean match score 12.2, SD 8.6). Interestingly, the median number of parameters used to search was only slightly higher for matched records ranked first (11, IQR: 9-11) than for matched records ranked second (10, IQR: 9-11) or third and below (10, IQR: 9-11), however this difference was statistically significant (p<0.01).

The matching parameters with the highest completeness during the first search attempt were first name, second name, third name, sex, year of birth, village, sub-village, and first and second name of a household member (all >83%) (Figure 4.2). These parameters also had the highest levels of agreement between the information collected and the matched HDSS record (all >64%), apart from third name, which had only 5.7% agreement. Fieldworkers took advantage of the linkage software's ability to perform multiple searches by updating the identifiers given during the brief interviews. A table that compares the completeness and agreement of all parameters between the first and matched search attempt can be found in the supplemental material (Supplemental Table 1 in Appendix 10.9.1). Briefly, the previously defined parameters with the highest levels of completeness and agreement for the first search had similar levels of completeness but increased levels of agreement for the search that resulted in a match.

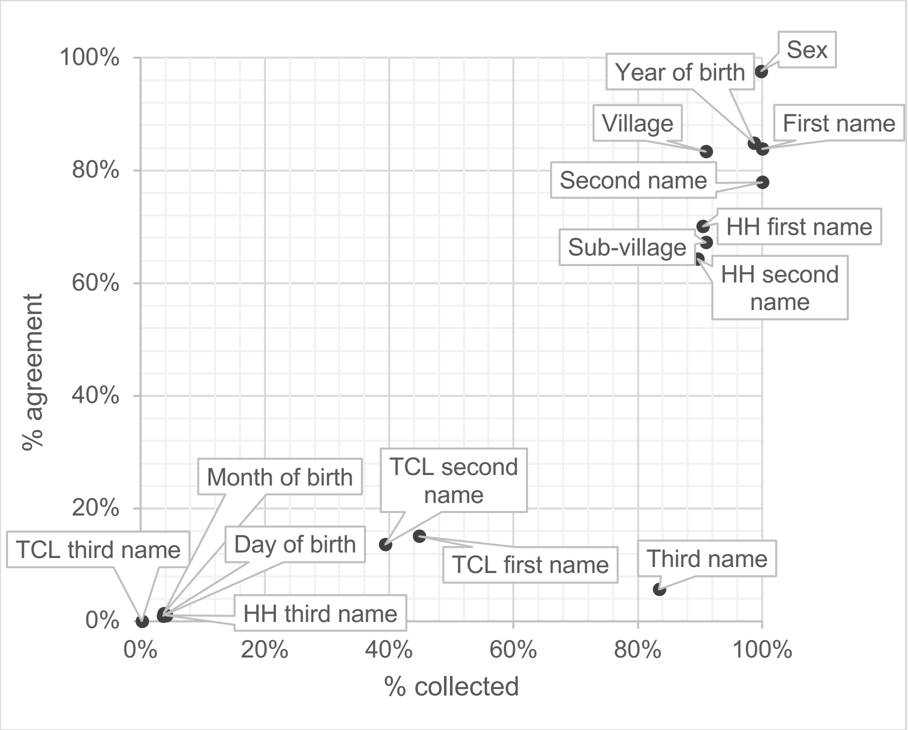


Figure 4.2: Quality measures of a probabilistic record linkage algorithm used to link health facility and HDSS databases in rural Tanzania, first search attempt.

Notes: HH = household member; TCL = ten-cell leader, an individual for a group of ten households; % collected = proportion of matched records with completed information; % agreement = proportion of matched records with agreeing information

4.5.5 Comparisons with automated linkage

Utilising the linked database resulting from PIRL as the gold standard, we applied a fully automated retrospective record linkage approach to compare the performance of the linkage algorithm. The full range of match scores among true matches was nearly completely enveloped by the range of match scores among true non-matches (Supplementary Figure 1 in Appendix 10.9.1). We calculated the sensitivity and PPV of the full algorithm at 10th-, 30th-, 50th-, 70th-, and 90th-percentile match score thresholds.

As the match score threshold was increased, sensitivity (the proportion of the 2,612 gold standard matches that were correctly identified and linked) decreased from 55% (1440/2612) to 10% (247/2612), and PPV (the proportion of linked records that were true matches) increased from 55% (1440/2612) to 85% (247/292) (Figure 4.3).

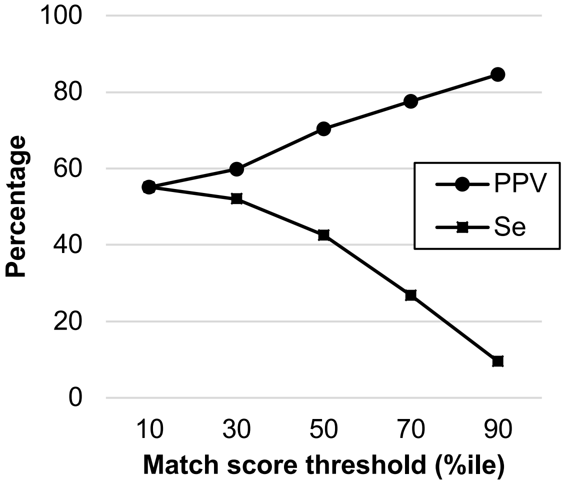


Figure 4.3: Sensitivity (Se) and positive predictive value (PPV) of automated retrospective record linkage at various match score percentile thresholds, full algorithm

Individual characteristics differed between the PIRL dataset and automated linked dataset at each match score threshold. Chiefly, the automated linkage resulted in a dataset that over-represented children aged five years or younger and under-represented adults aged between 18-34 years (all p<0.0001) (Table 4.4). Additionally, females were under-represented and males were over-represented in datasets created at higher match score thresholds (both p<0.02). Remarkably, the sensitivity analysis using an algorithm limited to only first name, second name, sex, year of birth, village, sub-village, and first and second name of a household member suggested the limited algorithm performed similarly to the full algorithm in terms of the algorithm's sensitivity and PPV, and the comparison between the automated linked datasets (Supplemental Figures 2 and 3, Supplemental Table 2 in Appendix 10.9.1).

Table 4.4: Distribution of individual characteristics in the dataset matched using point-of-contact interactive record linkage (PIRL) compared with those matched using a purely automated probabilistic approach using the full algorithm, by match score threshold

		Automated: full algorithm					
	PIRL match	**Threshold=10%ile**		**Threshold=50%ile**		**Threshold=90%ile**	
Characteristic	**n (%)**	**n (%)**	***p*-value***	**n (%)**	***p*-value***	**n (%)**	***p*-value***
Total matched (PPV)	2,612	2,612 (55.1)		1,579 (70.3)		292 (84.6)	
Sex							
Female	2,061 (78.9)	2,036 (78.0)	0.4004	1,185 (75.1)	0.0038	213 (73.0)	0.0191
Male	551 (21.1)	576 (22.1)		394 (25.0)		79 (27.1)	
Age, in years							
<5	125 (4.8)	198 (7.6)	<0.0001	132 (8.4)	<0.0001	46 (15.8)	<0.0001
5-17	393 (15.1)	464 (17.8)		239 (15.2)		35 (12.0)	
18-34	1,384 (53.0)	1,301 (49.9)		770 (48.8)		125 (42.8)	
35-49	522 (20.0)	433 (16.6)		301 (19.1)		68 (23.3)	
50-64	160 (6.1)	162 (6.2)		105 (6.7)		15 (5.1)	
65+	28 (1.1)	52 (2.0)		30 (1.9)		3 (1.0)	
Village of residence							
Kisesa	999 (38.3)	982 (37.6)	0.9340	586 (37.1)	0.8100	111 (38.0)	0.3320
Kanyama	521 (20.0)	529 (20.3)		302 (19.1)		46 (15.8)	
Kitumba	424 (16.2)	444 (17.0)		262 (16.6)		48 (16.4)	
Isangijo	257 (9.8)	258 (9.9)		176 (11.2)		39 (13.4)	
Ihayabuyaga	152 (5.8)	138 (5.3)		89 (5.6)		21 (7.2)	
Igekemaja	141 (5.4)	150 (5.7)		94 (6.0)		13 (4.5)	
Welamasonga	118 (4.5)	111 (4.3)		70 (4.4)		14 (4.8)	
Marital status[a]							
Never married	362 (24.0)	272 (24.1)	0.9997	179 (22.5)	0.4266	33 (22.3)	0.6089
Married once	724 (48.0)	540 (47.8)		403 (50.6)		72 (48.7)	
Remarried	175 (11.6)	132 (11.7)		99 (12.4)		22 (14.9)	
Separated/Widowed	249 (16.5)	187 (16.5)		116 (14.6)		21 (14.2)	
Pregnant at last HDSS round[b]							
No	1,057 (95.7)	758 (95.5)	0.8425	529 (95.0)	0.5292	101 (98.1)	0.3094
Yes	48 (4.3)	36 (4.5)		28 (5.0)		2 (1.9)	

Table 4.4: Distribution of individual characteristics in the dataset matched using point-of-contact interactive record linkage (PIRL) compared with those matched using a purely automated probabilistic approach using the full algorithm, by match score threshold

		Automated: full algorithm					
	PIRL match	**Threshold=10%ile**		**Threshold=50%ile**		**Threshold=90%ile**	
Characteristic	**n (%)**	**n (%)**	***p*-value***	**n (%)**	***p*-value***	**n (%)**	***p*-value***
Enrolled in school at last HDSS round[c]							
No	378 (72.0)	282 (67.6)	0.1454	185 (68.3)	0.2725	25 (52.1)	0.0038
Yes	147 (28.0)	135 (32.4)		86 (31.7)		23 (47.9)	

Abbreviations: HDSS - health and demographic sentinel surveillance

*Statistical differences tested for significance with chi-square (χ^2) or Fisher's Exact tests

[a]This question was only given to individuals aged 15 years or older

[b]This question was only given to females between 15 and 49 years of age

[c]This question was only given to individuals between 5 and 25 years of age

4.6 Discussion

PIRL – which combines a probabilistic search algorithm for identifying potential matches with a relatively simple human intervention – shows promise for linking multiple data sources in rural Tanzania. We matched 84% of individuals who reported any residence history in the HDSS area to at least one HDSS record. Session-specific notes stored in the software and discussions with fieldworkers suggested likely reasons (usually in combination with each other) why an HDSS record was not found for individuals who reported a residence history. First, the chances an HDSS enumerator contacted any respondent in a household was reduced as the household size decreased, particularly in households with one or two members. Second, HDSS rounds were usually conducted during the work day and may fail to capture individuals whose employment requires them away from home for extended periods of time. Lastly, given the sensitive nature of attending a clinic for HIV testing or care or antenatal services, fieldworkers were trained to use caution when a patient seemed unwilling to divulge the other personal information, such as names they may use at home (and be listed on their HDSS record), when a record could not be found. In these instances, we stopped searching for HDSS records in the hopes that the patient would be more amenable to sharing more information during any repeat visit.

During the study period, we had no refusals to provide informed written consent from clinic attendees who agreed to sit down with a fieldworker. We believe a more likely approach individuals who did not wish to participate may have taken was to passively refuse participation by not agreeing to meet with a fieldworker. During high-volume clinic days, the number of clinic attendees far exceeded the number of individuals we could enrol in record linkage, and patients who were willing to participate self-selected to queue for the fieldworkers.

Matching statistics improved as fieldwork progressed. Individuals who consented into the study with one of the more experienced fieldworkers or later in the study period were more likely to be matched than those who consented into the study with a recently hired fieldworker or at the beginning of the study period. These characteristics are indicators of an increasing maturity of the PIRL system and the increasing knowledge of the fieldworkers. Two of the three clinics (ANC and HTC) improved their match percentage compared to the first three months of fieldwork, which was likely due to the fieldworkers gaining understanding of the computer software. The lack of association with time and being matched in the CTC was likely due to the comparatively greater experience of fieldworker 1 who was the sole worker in the CTC during the first three months of the study period.

Increased age was another important characteristic associated with matching success, which has been shown elsewhere to be negatively associated with being matched using retrospective record linkage.[38] In theory, older individuals are likely to have spent a longer time in the HDSS area and thus have a more visible footprint in the database compared to younger individuals who are often more mobile. However, records for older individuals may contain out-of-date or inaccurate information, such as names, addresses, and dates of birth. A benefit of PIRL is the ability to perform multiple searches through the HDSS database while interviewing the individual whereas these issues would not get resolved using purely retrospective methods. There was also some evidence in the CTC and HTC that individuals from more rural areas of the HDSS area without a nearby road were independently more likely to be matched than those who lived near a main road. One explanation of this phenomenon could be due to the higher rate of migration within and into the urban and peri-urban areas, which have a higher density of households than in rural areas. A patient's sex was associated with being matched among ANC clinic attendees, where the small number of males were infants and were not likely to have an established record in the HDSS. Lastly, there was no evidence of an association between an HIV test result in the HTC and being matched to an HDSS record. Our belief was that HIV-positive individuals may be less likely to divulge identifying information required for record linkage; however, it is important to note the HTC clients in this study may not have been aware of their HIV status at the time of consenting to the study since record linkage interviews were conducted prior to HIV testing and counselling.

The results of the automated retrospective linkage substantiated the benefit of PIRL. At the 10th-percentile match score threshold, the algorithm had only 55% sensitivity and 55% PPV. In record linkage literature, the inverse of PPV is called the 'false-match rate' and is interpreted as the proportion of incorrectly linked records in a dataset.[77] Increasing the match score threshold resulted in lower sensitivity but with gains in PPV and thus a decreasing false match rate. At the 90th-percentile threshold, the algorithm had 10% sensitivity and the false-match rate was 15%. The choice of an acceptable level of false matches in a dataset depends on how the linked data are to be used. In our case, an appropriate amount of linkage error may be theorised as the maximum level at which secondary data analyses using the linked data would be unbiased. However, our results suggested that individual characteristics including age and sex were not properly represented in the automated linked datasets at any threshold. Therefore, analyses using data from automated linkage in this setting would potentially be biased. Further research is planned to measure the impact of varying linkage error rates on secondary data analyses (Chapter 5).

There were two other past attempts to link clinic and HDSS data in Kisesa. One study linked individuals' ANC records with their HDSS records using those whose ANC IDs were captured in an HDSS survey as the gold standard; out of 16,601 records, 75% were matched to an HDSS record with 70% sensitivity and 98% PPV.[39] Another study in Kisesa linked HTC clinic records to the HDSS using those whose HTC IDs were captured in an HIV surveillance round as the gold standard; out of 10,994 records, 37% were matched to an HDSS record with 18% sensitivity and a PPV of 69%.[11] The main limitations in each of these retrospective linkages was the poor data quality of the clinic ID variables captured in the HDSS and HIV surveillance data, respectively. PIRL is an approach that does not rely on previously collected identifiers that may suffer from poor data quality issues, such as high levels of missingness.

A key advantage of PIRL over a purely automated approach is the ability to perform multiple searches for the same individual. The match score that is calculated for each search attempt is not standardised and can be heavily influenced by both the quantity and quality of parameters used to search. The highest performing parameters during the first search attempt (first and second name, sex, year of birth, village, sub-village, and first and second name of a household member) all experienced 2-11% increased levels of agreement (a quality measure) between the first and matched search attempts. Concurrently, the change in the level of completeness (a quantity measure) in these parameters only changed between 0-3%. Therefore, these results suggest the amendments made to identifying information gathered during brief interviews was a key driver to locating a match – a feature of our PIRL system that is not common in purely automated linkage approaches.

We introduced a PIRL system to link HDSS records with a local health facility that serves the HDSS population with the goal of producing a data source that could be used to monitor the utilisation of health services and the outcomes of patients after they have made contact with the health system. The linked clinical data could also be used to validate or substitute the self-reported health status and health service use data collected in the HDSS surveys. Depending on available support, we conclude PIRL should be continued and expanded in Kisesa to other clinics in the HDSS area. We believe PIRL may be a effective solution for smaller-scale research projects where data quality is a principal concern.

4.7 CONCLUSIONS

Where resources are available, PIRL is a promising tool for linking multiple sources of data in a setting that lacks unique identifiers. We developed PIRL software that incorporated a probabilistic algorithm and allowed for multiple search attempts for an individual. A high majority (84%) of the individuals who reported residence history in the area were matched to one or more of their HDSS records. In this setting, an automated retrospective approach to record linkage at the lowest thresholds would have only correctly identified about half of the true matches and resulted in high linkage errors, therefore highlighting immediate benefit of this prospective approach. The data infrastructure produced by PIRL has the potential to become an invaluable resource for monitoring access to and utilization of health facility services at subnational levels.

4.8 SUPPLEMENTARY MATERIAL

All supplementary material for this publication can be found in Appendix 10.9.1.

- Supplemental Table 1. Agreement conditions, match (m) probabilities, proportion collected, and proportion of records with agreement for each field (i) in the probabilistic algorithm, by first and matched search attempts
- Supplemental Figure 1. Log frequency of match scores calculated for all pairwise comparisons using full algorithm, by true match status
- Supplemental Figure 2. Log frequency of match scores calculated for all pairwise comparisons using limited algorithm, by true match status
- Supplemental Figure 3. Sensitivity (Se) and positive predictive value (PPV) of automated retrospective record linkage at various match score percentile thresholds, full (F) vs. limited (L) algorithm
- Supplemental Table 2. Distribution of individual characteristics in the dataset matched using point-of-contact interactive record linkage (PIRL) compared with those matched using a purely automated probabilistic approach using a full and limited algorithm, by match score threshold

5 Paper C. Impact of linkage quality on inferences drawn from analyses using data with high rates of linkage errors in rural Tanzania

Christopher T. Rentsch[1], Katie Harron[2], Mark Urassa[3], Jim Todd[1,3], Georges Reniers[1,4], Basia Żaba[1]

[1]Department of Population Health, Faculty of Epidemiology and Population Health, London School of Hygiene & Tropical Medicine, London, UK
[2]UCL GOS Institute of Child Health, London, UK
[3]The TAZAMA Project, National Institute for Medical Research, Mwanza, Tanzania
[4]MRC/Wits Rural Public Health and Health Transitions Research Unit (Agincourt), School of Public Health, Faculty of Health Sciences, University of the Witwatersrand, Johannesburg, South Africa

London School of Hygiene & Tropical Medicine
Keppel Street, London WC1E 7HT
www.lshtm.ac.uk

Registry
T: +44(0)20 7299 4646
F: +44(0)20 7299 4656
E: registry@lshtm.ac.uk

LONDON
SCHOOL *of*
HYGIENE
&TROPICAL
MEDICINE

RESEARCH PAPER COVER SHEET

PLEASE NOTE THAT A COVER SHEET MUST BE COMPLETED FOR EACH RESEARCH PAPER INCLUDED IN A THESIS.

SECTION A – Student Details

Student	Christopher T. Rentsch
Principal Supervisor	Professor Basia Żaba and Dr. Georges Reniers
Thesis Title	Point-of-contact interactive record linkage between demographic surveillance and health facilities to measure patterns of HIV service utilisation in Tanzania

If the Research Paper has previously been published please complete Section B, if not please move to Section C

SECTION B – Paper already published

Where was the work published?			
When was the work published?			
If the work was published prior to registration for your research degree, give a brief rationale for its inclusion			
Have you retained the copyright for the work?*		Was the work subject to academic peer review?	

**If yes, please attach evidence of retention. If no, or if the work is being included in its published format, please attach evidence of permission from the copyright holder (publisher or other author) to include this work.*

SECTION C – Prepared for publication, but not yet published

Where is the work intended to be published?	*BMC Medical Research Methodology*
Please list the paper's authors in the intended authorship order:	Christopher T. Rentsch, Katie Harron, Mark Urassa, Jim Todd, Georges Reniers, Basia Żaba
Stage of publication	Submitted (under review)

SECTION D – Multi-authored work

For multi-authored work, give full details of your role in the research included in the paper and in the preparation of the paper. (Attach a further sheet if necessary)	I contributed to the conception of the investigation, designed the study, created analysis plans, extracted and cleaned data, conducted the analysis, drafted the manuscript, collated feedback from co-authors, submitted manuscript, and liaised with journal editors

Student Signature: [redacted] **Date:** 8/8/2018

Supervisor Signature: [redacted] **Date:** 8/8/2018

5.1 Overview

In the previous chapter, I compared PIRL with a fully automated linkage approach using the same linkage algorithm and found that the latter resulted in substantial levels of linkage errors. In this final methodological chapter, I measure the impact of linkage quality on inferences drawn from secondary analyses using data with high rates of linkage errors.

Objective 2. To identify individual characteristics associated with successful linkage using PIRL and compare PIRL with traditional, automated probabilistic record linkage.

5.2 Abstract

Background. Studies based on high-quality linked data in developed countries show that even minor linkage errors can impact bias and precision of subsequent analyses. We evaluated the impact of linkage quality on inferences drawn from analyses using data with substantial linkage errors in rural Tanzania.

Methods. Gold standard links were available for community-based HIV surveillance data and digitised medical records at three clinics serving the surveillance population based on point-of-contact interactive record linkage. Automated probabilistic record linkage was subsequently used to create linked datasets at minimum, low, medium, and high match score thresholds. Cox proportional hazards regression models were used to compare HIV care registration rates by testing modality (sero-survey vs. clinic) in each analytic dataset created by the automated linkage. We assessed linkage quality using three approaches: quantifying linkage errors, comparing characteristics between linked and unlinked data, and evaluating bias and precision of the primary exposure (testing modality) regression estimate.

Results. Between 2014-2017, 405 individuals with gold standard links were newly diagnosed with HIV in sero-surveys (n=263) and clinics (n=142). Automated probabilistic linkage correctly identified 233 individuals (positive predictive value [PPV]=65%) at the low threshold and 95 individuals (PPV=90%) at the high threshold. Significant differences were found between linked and unlinked records in primary exposure and outcome variables and for adjusting covariates at every threshold. As expected, differences attenuated with increasing threshold. Testing modality was significantly associated with time to CTC registration in the gold standard data (adjusted hazard ratio [HR] 4.98 for clinic-based testing, 95% confidence interval [CI] 3.34, 7.42). Increasing false matches weakened the association (HR 2.76 at minimum match score threshold, 95% CI 1.73, 4.41). Increasing missed matches (i.e., increasing match score threshold and positive predictive value of the linkage algorithm) was strongly correlated with a reduction in the precision of coefficient estimate (R^2=0.97; p=0.03).

Conclusions. Similar to studies with more negligible levels of linkage errors, false matches in this setting reduced the magnitude of the association; missed matches reduced precision. Adjusting for these biases could provide more robust results using data with considerable linkage errors.

5.3 BACKGROUND

A growing number of demographic and epidemiological research studies are conducted using linked datasets from multiple sources.[15] In the absence of unique identifiers, record linkage – the matching of an individual's records between two or more data sources[28, 29] – often relies on a set of personal identifiers (e.g., names, address, date of birth) that are reported with error or are dynamic (e.g., name or residence changes). Errors arising during the linkage process because of imperfect identifiers can result in two types of linkage errors: false matches (records of two different individuals are erroneously linked) and missed matches (records belonging to the same individual are not linked). These linkage errors have been shown to impact the bias and precision of subsequent analyses.[45, 46] False matches typically weaken associations between variables captured in different datasets and bias coefficients toward a null association[46] while missed matches result in a decreased analytic sample size and thus statistical power, and potentially underestimate exposures and outcomes of interest. [108, 109] Globally, there is a lack of guidance on how to measure the impact of linkage errors on analyses of linked data.[110, 111] However, the few studies that exist are predominantly conducted in settings with very low linkage errors, such as the United Kingdom, United States, and Australia.[47, 112] Whether and how analyses are affected by more substantial linkage errors remains unknown.

A recent Wellcome Trust report detailed how record linkage adds to the value of medical research in low- and middle-income countries.[15] A unique challenge exists in these settings, particularly in sub-Saharan Africa, where there is an overall lack of electronic data available for linkage and relatively poor quality of variables that could be used by a linkage algorithm. Because of this, very few record linkage projects have been undertaken throughout the region,[11, 37-39] and the absence of gold standard linked data complicate those that have used automated linkage. In a rural ward of ~35,000 residents in northwest Tanzania with a history of community-based HIV surveillance, we developed and implemented a novel approach to record linkage, which we term point-of-contact interactive record linkage (PIRL).[72-74] PIRL, described later in more detail, is a semi-automatic record linkage process that incorporates human inspection of potential matches identified by a probabilistic linkage algorithm whilst in the presence of the individual whose records are being linked, which contrasts with a more conventional approach where record linkage is done automatically with no human involvement. PIRL has the advantage that uncertainty surrounding identities can be resolved during a brief interaction whereby extraneous information (e.g., household membership) can be referred to as an additional criterion to adjudicate between multiple potential matches. Largely due to the interaction with those who are the target of the linkage and the ability to perform repeated searches through the database, PIRL has been shown to outperform automated linkage for identifying matches, which have been affected by the substantial data quality issues in similar settings.[74] The gold standard linked database created by PIRL allows for the first known attempt to evaluate the impact of linkage errors on subsequent analyses in a setting with substantial linkage errors.

The linked data infrastructure created by PIRL includes gold standard links between HIV serological survey data and manually digitised medical records from three clinics serving the surveillance population, two of which offer HIV testing services while the third enrols HIV-positive individuals into care. As an illustrative example to evaluate linkage errors, we tested whether individuals who receive their first HIV diagnosis during a village-based HIV serological survey enrol for HIV care services quicker than those who receive their first HIV diagnosis in a clinic that also offers HIV testing. For this analysis, we first assessed the relationship between diagnosis location and time to enrolment into HIV care in the gold standard linked data. We then conducted automated record linkage, a process that included no human interaction or involvement like PIRL, to quantify linkage errors in this setting. Finally, we determined whether and how linkage errors impacted the analysis of the primary research question by comparing the characteristics of linked and unlinked records and the bias and precision of regression coefficients.

5.4 Methods

5.4.1 Data sources

The Kisesa observational HIV cohort study was established in 1994 and is located in a rural ward in the Magu district of Mwanza region in northwest Tanzania.[68] The study includes multiple rounds of health and demographic surveillance system (HDSS) surveys that cover the entire population of ~35,000 residents, and multiple rounds of population-based HIV sero-surveys, in which adults aged 15 years or older living in the Kisesa HDSS study area are invited to attend temporary village-based clinics for a personal interview and HIV test. A government-run health centre serving the HDSS population includes an HIV testing and counselling clinic (HTC), an antenatal clinic (ANC) offering HIV testing, and an HIV care and treatment centre (CTC). For the HTC and ANC, we developed electronic databases and digitised the paper-based logbooks using a double-entry system where two different fieldworkers independently capture each book, and any discrepancy between fields were reconciled in a third cleaning stage. The CTC databases have been fully digitised, and data clerks regularly update and run data checks on these data. Ethical approval was obtained from the National Institute for

Medical Research, Tanzania (reference no. NIMR/HQ/R.8c/Vol.II/436 and MR/53/100/450), and the London School of Hygiene and Tropical Medicine (Project ID #8852). Informed written consent (including consent to link data sources in the PIRL study) was obtained from all participants.

5.4.2 Linkage

Participants' records from all sero-survey rounds were cross-referenced with their HDSS identifiers as part of the identification process during the survey interview. Records from the three clinics were linked to the HDSS database using PIRL, which has been described elsewhere.[73, 74] Briefly, as individuals arrived to any of the three clinics and consented to be in the study, fieldworkers entered their personal and residence details into specialised computer software,[72] which used a probabilistic linkage algorithm to search the HDSS database. The algorithm used to search for possible matches was based on the Fellegi-Sunter record linkage model,[28, 29] and incorporated the following data fields: up to three names for the individual; sex; year, month, and day of birth; village and sub-village; up to three names of a household member; and up to three names for the ten-cell leader of the patient. A ten-cell leader is an individual who acted as a leader for a group of ten households and these positions have been relatively stable over time. While searching through potential matches, the fieldworker could view the full list of household members associated with each HDSS record as an additional step to adjudicate true matches. The fieldworker then interacted with the patient to identify which HDSS record(s), if any, were a true match.

Multiple data checks were performed within the software and on the back-end database to ensure the links made with PIRL were true matches. First, the software displayed warning messages to the fieldworkers if they attempted to match to a record that had an absolute difference in birth year of >10 years, or the entered names did not agree with the names listed on the selected HDSS record as measured by a Jaro-Winkler string comparator.[101] The linkage algorithm allowed for all pairwise comparisons between listed names on clinic and HDSS records because the order of names is relaxed in this setting and HDSS records only hold up to two names while other data sources often store more than two names. Further, the lead author performed periodic and manual, back-end inspection of the data to verify the matches made in the field. These data integrity checks flagged individuals who were matched to multiple HDSS records with large age differences (>10 years), of conflicting sex, within the same household, or with overlapping household residency episodes in which one record's start date occurred before another record's end date. Over the study period, eight PIRL matches were deemed unlikely and deleted.

Using links made during the sero-survey and PIRL as the gold standard, we performed automated probabilistic record linkage using the same algorithm used in the PIRL software but limited to identifiers collected in the sero-survey and clinic databases. Automated record linkage has been well described.[27, 76-79] Briefly, a match score (i.e., the weighted likelihood a record-pair is a link or non-link) was calculated for all pairwise comparisons between the patient registry and the HDSS database. The HDSS record with the highest match score was selected for each record in the patient registry. When performing automated linkage, a match score threshold is selected to determine what constitutes a link versus a non-link. The placement of the threshold can be a matter of trial and error.[34] Additionally, a match score is not a standardised metric and can be greatly influenced by the number of identifiers used in the linkage algorithm. To show how the impact of linkage errors on subsequent analyses were affected by the placement of the match score threshold, we created separate analytic datasets at various thresholds based on percentiles of the distribution of match scores among true matches: (a) all matches above the minimum match score, (b) low or 25th percentile, (c) medium or 50th percentile, and (d) high or 75th percentile. Higher thresholds represent more conservative definitions on what constituted a true match. The PIRL links made between the CTC and HDSS databases were then used for the entire sample to identify those who registered for HIV care.

5.4.3 Analytic sample

We included all individuals with a gold standard link who received their first positive HIV diagnosis in the sero-survey, HTC, or ANC between December 2014 and October 2017. Individuals were excluded if they were younger than 15 years (to be consistent with the 15-year age limit in the sero-survey), had evidence of a previous positive HIV diagnostic test or registered for HIV care prior to their HIV test (repeat testers), or reported residence outside the HDSS area or were not seen in the 2016/17 HDSS round (non-residents). Repeat testers and non-residents were excluded because these groups are likely to achieve the outcome (registered for HIV care) at different rates than individuals newly diagnosed with HIV and residents. We extracted demographic and spatial characteristics including sex, age, rurality of sub-village (rural, peri-urban, or urban), whether the sub-village of residence had a paved road, and geodesic distance between an individual's household and the CTC.

5.4.4 Statistical analyses

Chi-square and Fisher's exact tests were used to assess differences between individuals who were diagnosed with HIV by testing modality, i.e. in the community-based sero-survey versus walk-in clinic (either HTC or ANC) during the study period in the gold standard data. At each match score threshold, we classified links made by the automated linkage as true, false, or missed matches and compared characteristics between these groups using standardised differences.[113] Standardised differences of 0.2, 0.5, and 0.8 represented small, moderate, and large standardised differences, respectively, comparing true matches with false and missed matches.[114] Cox proportional hazards regression models were used to compare HIV care registration rates by testing modality (sero-survey vs. clinic) in each dataset created by the automated linkage. Individuals were censored at first CTC visit, death, or 90 days after positive HIV diagnosis. Models were adjusted for age, sex, rurality of sub-village, whether the sub-village had a paved road, and distance to the CTC. We evaluated for bias in precision by comparing regression coefficients and standard errors of the primary exposure variable (testing modality) in the gold standard data with those obtained at each selected match score threshold. Statistical analyses were performed using SAS version 9.4 (SAS Institute Inc., Cary, NC, USA).

5.5 RESULTS

5.5.1 Gold standard links

During the study period, 263 and 142 individuals with gold standard links received their first positive HIV diagnosis in the sero-survey and clinics, respectively (total n=405). Among clinic patients, 126 (89%) HIV diagnoses occurred in the HTC and the remaining 16 (11%) diagnoses were made in the ANC. Participants diagnosed in the sero-survey were more likely to be older, from more rural areas, and reside further from the CTC than those who were diagnosed in a clinic (all p<0.02) (Table 5.1). Over half (n=75 [53%]) of individuals diagnosed in a clinic subsequently registered for HIV care by the study cut-off date, compared to 42 (16%) of those diagnosed in the sero-survey (p<0.0001).

Table 5.1: Characteristics of patients in the analytic sample

Characteristic	**Sero-survey participants** (n=263)	**Clinic patients** (n=142)	**p-value**
Clinic			
ANC	-	16 (11.3)	-
HTC	-	126 (88.7)	
Sex			
Female	173 (65.8)	98 (69.0)	0.5092
Male	90 (34.2)	44 (31.0)	
Age, years			
15-29	62 (23.6)	51 (35.9)	0.0222
30-39	96 (36.5)	53 (37.3)	
40-49	59 (22.4)	22 (15.5)	
50+	46 (17.5)	16 (11.3)	
Village			
Igekemaja	27 (10.3)	14 (9.9)	0.0167
Ihayabuyaga	30 (11.4)	6 (4.2)	
Isangijo	27 (10.3)	14 (9.9)	
Kanyama	38 (14.5)	23 (16.2)	
Kisesa	73 (27.8)	51 (35.9)	
Kitumba	32 (12.2)	26 (18.3)	
Welamasonga	36 (13.7)	8 (5.6)	
Rurality of sub-village			
Rural	140 (53.2)	55 (38.7)	0.0204
Peri-urban	54 (20.5)	39 (27.5)	
Urban	69 (26.2)	48 (33.8)	
Sub-village had paved road			
Yes	109 (41.4)	70 (49.3)	0.1290
No	154 (58.6)	72 (50.7)	
Distance from household to CTC, km			
<1	53 (20.2)	37 (26.1)	0.0162
1-1.9	58 (22.1)	45 (31.7)	
2-4.9	60 (22.8)	29 (20.4)	
5-11	92 (35.0)	31 (21.8)	
Registered at CTC	42 (16.0)	75 (52.8)	<0.0001

Abbreviations: CTC - HIV care and treatment centre; ANC - antenatal clinic; HTC - HIV testing and counselling clinic

Note: all statistics are given in n (%); differences tested using chi-square

5.5.2 Automated linkage

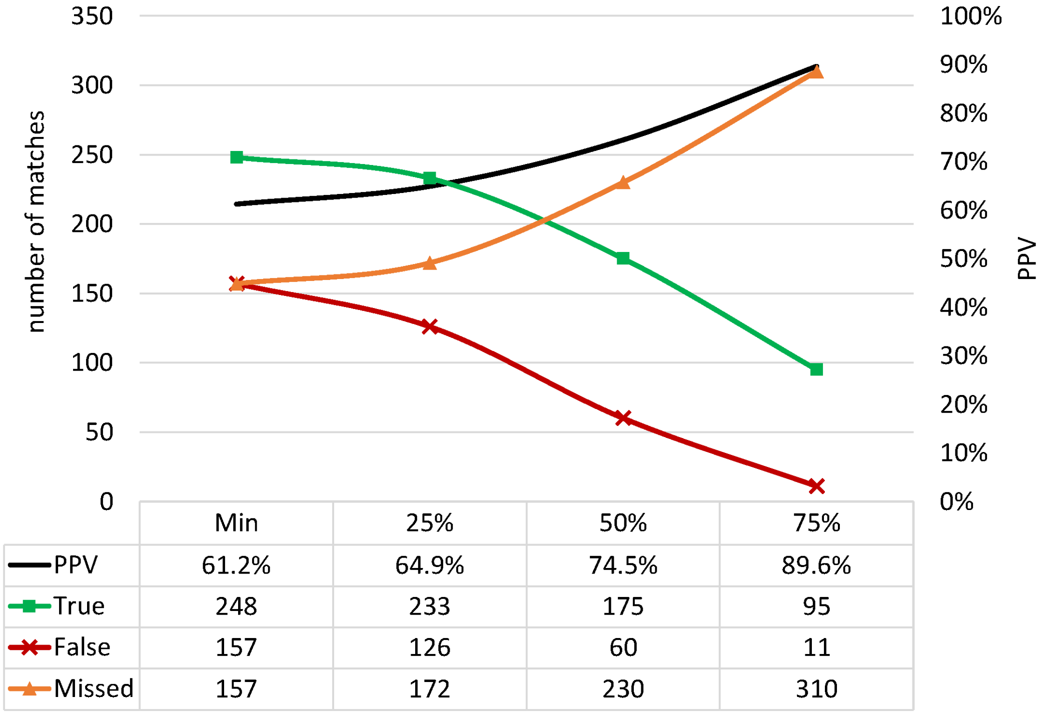
Most identifiers used by the linkage algorithm were complete or nearly complete in the sero-survey and clinic databases, including two names, year of birth, sex, village, and sub-village information (all ≥99.3% complete) (Table 5.2). A majority (72%) of sero-survey records also included two names of another household member, 48% included two names of the household's ten-cell leader, and 13% had a third name for the individual. Most (89%) clinic records held information on a third name for the individual, >75% up to two names of the household's ten-cell leader, and 12% included two names for another household member. The HDSS database had high levels of completeness (all >99%) on all identifiers used by the linkage algorithm except for a third name, which is not collected in the HDSS system.

Table 5.2: Completeness of matching identifiers in clinic data and demographic surveillance data

	% records with complete information		
Matching identifier	**Sero-surveys** (n=263)	**Clinic data** (n=142)	**HDSS data** (n=99,866)
First name	100.0%	100.0%	100.0%
Second name	100.0%	100.0%	100.0%
Third name	13.3%	88.7%	-
Year of birth	100.0%	100.0%	99.4%
Sex	100.0%	100.0%	100.0%
Village	100.0%	99.3%	100.0%
Sub-village	100.0%	99.3%	100.0%
TCL first name	48.3%	91.5%	99.4%
TCL second name	48.3%	74.6%	99.4%
Household member first name	71.5%	11.3%	99.9%
Household member second name	71.5%	11.3%	99.9%

Abbreviations: HDSS - health and demographic surveillance system; TCL - ten-cell leader

Of the 405 gold standard links, automated probabilistic linkage correctly identified 248 individuals, falsely matched 157 individuals, and missed 157 individuals at the minimum match score threshold, which resulted in a positive predictive value (PPV) of 61% (Figure 5.1). Increasing the match score threshold to a more conservative definition of a match resulted in a decrease in the number of true (n=95) and false (n=11) matches and an increase in the number missed matches (n=310) and PPV (90%) at the 75th-percentile match score threshold.



	Min	25%	50%	75%
PPV	61.2%	64.9%	74.5%	89.6%
True	248	233	175	95
False	157	126	60	11
Missed	157	172	230	310

Figure 5.1: Positive predictive value (PPV) and number of true, false, and missed matches, by match score threshold

5.5.3 Linked sample characteristics

The frequency of the primary exposure variable, the location in which an individual received their first positive HIV diagnostic test, differed between true, false, and missed matches at all match score thresholds (Table 5.3). Compared to linked true matches, false and missed matches were more likely to receive their HIV-positive test in a clinic than the sero-survey. Increasing the threshold minimised but did not eliminate the differences between true matches and false matches.

The frequency of the outcome variable, registering at the CTC, also differed significantly between true matches and false matches, particularly at lower match score thresholds. Compared to linked true matches, false matches were less likely to have registered at the CTC at every match score threshold except for the high threshold.

There were also differences between true, false, and missed matches with respect to variables used as adjusting factors. False matches were more likely to be younger, from more rural areas, and reside at greater distances from the CTC. There were minimal differences between true and false matches by sex in analytic samples created using lower match score thresholds; however, false matches were more likely than true matches to be male at the medium and high match score thresholds.

Table 5.3: Characteristics of records linked by automated linkage according to linkage status, by match score threshold

Threshold:	Minimum					Low (25th percentile)					Medium (50th percentile)					High (75th percentile)				
	True matches (n=248)	**False matches** (n=157)	**St. diff.**	**Missed matches** (n=157)	**St. diff.**	**True matches** (n=233)	**False matches** (n=126)	**St. diff.**	**Missed matches** (n=172)	**St. diff.**	**True matches** (n=175)	**False matches** (n=60)	**St. diff.**	**Missed matches** (n=230)	**St. diff.**	**True matches** (n=95)	**False matches** (n=11)	**St. diff.**	**Missed matches** (n=310)	**St. diff.**
Exposure variable																				
Test location																				
Sero-survey	185 (74.6)	78 (49.7)	0.53	78 (49.7)	0.53	185 (79.4)	77 (61.1)	0.41	78 (45.3)	0.75	163 (93.1)	52 (86.7)	0.22	100 (43.5)	1.26	93 (97.9)	11 (100.0)	0.21	170 (54.8)	1.18
Clinic	63 (25.4)	79 (50.3)		79 (50.3)		48 (20.6)	49 (38.9)		94 (54.7)		12 (6.9)	8 (13.3)		130 (56.5)		2 (2.1)	0 (0.0)		140 (45.2)	
Outcome variable																				
Registered at CTC	70 (28.2)	7 (4.5)	0.68	47 (29.9)	0.04	60 (25.8)	7 (5.6)	0.58	57 (33.1)	0.16	34 (19.4)	6 (10.0)	0.27	83 (36.1)	0.38	16 (16.8)	2 (18.2)	0.04	101 (32.6)	0.37
Adjusting factors																				
Sex																				
Female	165 (66.5)	106 (67.5)	0.02	106 (67.5)	0.02	154 (66.1)	85 (67.5)	0.03	117 (68.0)	0.04	113 (64.6)	32 (53.3)	0.23	158 (68.7)	0.09	57 (60.0)	4 (36.4)	0.49	214 (69.0)	0.19
Male	83 (33.5)	51 (32.5)		51 (32.5)		79 (33.9)	41 (32.5)		55 (32.0)		62 (35.4)	28 (46.7)		72 (31.3)		38 (40.0)	7 (63.6)		96 (31.0)	
Age, years																				
15-29	60 (24.2)	72 (45.9)	0.48	53 (33.8)	0.32	52 (22.3)	59 (46.8)	0.53	61 (35.5)	0.38	33 (18.9)	22 (36.7)	0.41	80 (34.8)	0.39	20 (21.1)	3 (27.3)	0.41	93 (30.0)	0.21
30-39	89 (35.9)	44 (28.0)		60 (38.2)		86 (36.9)	33 (26.2)		63 (36.6)		67 (38.3)	17 (28.3)		82 (35.7)		37 (38.9)	3 (27.3)		112 (36.1)	
40-49	52 (21.0)	19 (12.1)		29 (18.5)		49 (21.0)	17 (13.5)		32 (18.6)		41 (23.4)	12 (20.0)		40 (17.4)		22 (23.2)	4 (36.4)		59 (19.0)	
50+	47 (19.0)	22 (14.0)		15 (9.6)		46 (19.7)	17 (13.5)		16 (9.3)		34 (19.4)	9 (15.0)		28 (12.2)		16 (16.8)	1 (9.1)		46 (14.8)	
Village																				
Igekemaja	32 (12.9)	10 (6.4)	0.38	9 (5.7)	0.40	29 (12.4)	9 (7.1)	0.41	12 (7.0)	0.32	20 (11.4)	6 (10.0)	0.45	21 (9.1)	0.27	13 (13.7)	2 (18.2)	0.85	28 (9.0)	0.28
Ihayabuyaga	20 (8.1)	17 (10.8)		16 (10.2)		18 (7.7)	14 (11.1)		18 (10.5)		14 (8.0)	9 (15.0)		22 (9.6)		7 (7.4)	0 (0.0)		29 (9.4)	
Isangijo	23 (9.3)	13 (8.3)		18 (11.5)		23 (9.9)	10 (7.9)		18 (10.5)		20 (11.4)	5 (8.3)		21 (9.1)		9 (9.5)	0 (0.0)		32 (10.3)	
Kanyama	39 (15.7)	23 (14.6)		22 (14.0)		37 (15.9)	16 (12.7)		24 (14.0)		26 (14.9)	11 (18.3)		35 (15.2)		12 (12.6)	1 (9.1)		49 (15.8)	
Kisesa	81 (32.7)	52 (33.1)		43 (27.4)		73 (31.3)	42 (33.3)		51 (29.7)		56 (32.0)	11 (18.3)		68 (29.6)		28 (29.5)	2 (18.2)		96 (31.0)	
Kitumba	25 (10.1)	31 (19.7)		33 (21.0)		25 (10.7)	27 (21.4)		33 (19.2)		17 (9.7)	11 (18.3)		41 (17.8)		11 (11.6)	2 (18.2)		47 (15.2)	
Welamasonga	28 (11.3)	11 (7.0)		16 (10.2)		28 (12.0)	8 (6.3)		16 (9.3)		22 (12.6)	7 (11.7)		22 (9.6)		15 (15.8)	4 (36.4)		29 (9.4)	
Rurality of sub-village																				
Rural	118 (47.6)	69 (43.9)	0.08	77 (49.0)	0.08	113 (48.5)	60 (47.6)	0.10	82 (47.7)	0.04	83 (47.4)	36 (60.0)	0.30	112 (48.7)	0.06	51 (53.7)	8 (72.7)	0.42	144 (46.5)	0.15
Peri-urban	55 (22.2)	36 (22.9)		38 (24.2)		52 (22.3)	24 (19.0)		41 (23.8)		39 (22.3)	13 (21.7)		54 (23.5)		19 (20.0)	1 (9.1)		74 (23.9)	
Urban	75 (30.2)	52 (33.1)		42 (26.8)		68 (29.2)	42 (33.3)		49 (28.5)		53 (30.3)	11 (18.3)		64 (27.8)		25 (26.3)	2 (18.2)		92 (29.7)	
Sub-village had paved road																				
Yes	115 (46.4)	75 (47.8)	0.03	64 (40.8)	0.11	105 (45.1)	59 (46.8)	0.04	74 (43.0)	0.04	81 (46.3)	22 (36.7)	0.20	98 (42.6)	0.07	40 (42.1)	3 (27.3)	0.32	139 (44.8)	0.06
No	133 (53.6)	82 (52.2)		93 (59.2)		128 (54.9)	67 (53.2)		98 (57.0)		94 (53.7)	38 (63.3)		132 (57.4)		55 (57.9)	8 (72.7)		171 (55.2)	
Distance from household to CTC, km																				
<1	52 (21.0)	38 (24.2)	0.10	38 (24.2)	0.16	49 (21.0)	31 (24.6)	0.10	41 (23.8)	0.08	39 (22.3)	12 (20.0)	0.35	51 (22.2)	0.09	19 (20.0)	2 (18.2)	0.85	71 (22.9)	0.20
1-1.9	69 (27.8)	45 (28.7)		34 (21.7)		62 (26.6)	33 (26.2)		41 (23.8)		48 (27.4)	9 (15.0)		55 (23.9)		24 (25.3)	0 (0.0)		79 (25.5)	
2-4.9	51 (20.6)	32 (20.4)		38 (24.2)		51 (21.9)	28 (22.2)		38 (22.1)		36 (20.6)	18 (30.0)		53 (23.0)		17 (17.9)	3 (27.3)		72 (23.2)	
5-11	76 (30.6)	42 (26.8)		47 (29.9)		71 (30.5)	34 (27.0)		52 (30.2)		52 (29.7)	21 (35.0)		71 (30.9)		35 (36.8)	6 (54.5)		88 (28.4)	

0.2, 0.5, and 0.8 can be considered as small, medium, and large effect sizes respectively

Abbreviations: St. diff. - standardised differences; CTC - HIV care and treatment centre; ANC - antenatal clinic; HTC - HIV testing and counselling clinic

Note: Comparisons made to true matches within each match score threshold stratum; all statistics are given in n (%)

5.5.4 Modelled estimates

There was a significant association between testing modality and time to registration at the CTC in the linked gold standard data in favour of those receiving their diagnosis at a walk-in clinic (adjusted hazard ratio [HR] 4.98, 95% confidence interval [CI] 3.34, 7.42) (Table 5.4). Bias was present at each match score threshold in the automated linked datasets. The significant positive association was still found, though much attenuated, at the minimum threshold (HR 2.76, 95% CI 1.73, 4.41) and low threshold (HR 3.32, 95% CI 2.00, 5.51) (Figure 5.2). The association was not found at the medium threshold (HR 2.37, 95% CI 0.96, 5.87) nor the high threshold (HR 1.70, 95% CI 0.17, 16.87). An increase in the number of missed matches from the analytic dataset (i.e. increasing the match score threshold and positive predictive value of the linkage algorithm) was strongly correlated with a reduction in the precision of the primary exposure coefficient (R^2=0.97; p=0.03).

Table 5.4: Comparison of regression model diagnostics by match score threshold

Sample	n	β	SE	χ^2	p	HR (95% CI)	PPV
Gold standard	405	1.61	0.2033	62.4	<.0001	4.98 (3.34, 7.42)	-
Probabilistic linkage threshold, by match score threshold							
minimum	405	1.02	0.2383	18.2	<.0001	2.76 (1.73, 4.41)	0.612
low	359	1.20	0.2579	21.7	<.0001	3.32 (2.00, 5.51)	0.649
medium	235	0.86	0.4621	3.5	0.0615	2.37 (0.96, 5.87)	0.745
high	106	0.53	1.1707	0.2	0.6501	1.70 (0.17, 16.87)	0.896

Abbreviations: n - sample size; β - primary exposure coefficient; SE - standard error; χ^2 - chi-square; p - p-value; HR - hazard ratio; CI - confidence interval; PPV - automated linkage algorithm's positive predictive value

*Adjusted for age, sex, sub-village, and distance from household to CTC

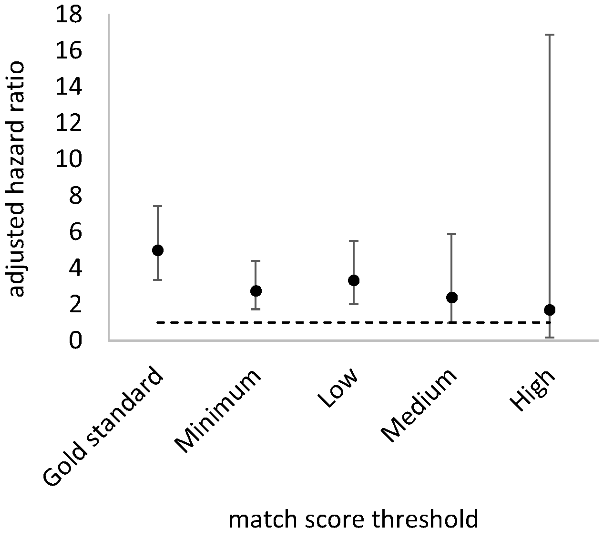


Figure 5.2: Association between the primary exposure and outcome variables by match score threshold

5.6 Discussion

This paper provides original evidence that bias and precision in analyses using linked data are impacted by substantial linkage errors similarly to how they are impacted by more negligible linkage errors. With the recent availability of gold standard linked data in this East African setting, we asked a timely research question and assessed how our conclusions would have changed if instead of using gold standard linked data, we used automated record linkage, a less resource-intensive but less accurate form of record linkage. We evaluated the quality of automated linkage and identified potential sources of bias by quantifying false and missed matches, comparing characteristics between linked and unlinked data, and comparing regression coefficients at various match score thresholds in sensitivity analyses. High levels of linkage errors in this setting introduced bias at all match score thresholds. False matches reduced the magnitude of the association between the tested exposure and outcome while increasing numbers of missed matches reduced the precision of these estimates, which is comparable to analyses in settings with higher quality data.[46, 47, 108, 109, 115]

We used standardised differences to identify variables that were more affected by linkage error and potential sources of bias as was done in previous studies.[116] We found strong evidence of selection bias based on who was included in the analytic datasets since frequencies of the primary exposure, outcome, and some adjusting variables differed significantly between true, false, and missed matches at all match score thresholds. Increasing the match score threshold attenuated differences between true and false matches but also exacerbated differences between true and missed matches. The trade-off between false and missed matches when comparing characteristics between linked and unlinked data has also been found in other settings with low levels of linkage errors.[117] Importantly, the proportion of individuals who registered at the CTC (outcome) was 29% in the gold standard data and ranged between 17%-19% in the linked datasets. Therefore, if our research question was to obtain the proportion or rate of individuals who registered at the CTC our conclusions would also have been meaningfully different at every match score threshold.

We found measurable bias in the regression coefficient of the primary exposure at every match score threshold. Selection bias is likely to have impacted the analyses given that selection into the linked datasets was related to both the exposure and outcome.[118, 119] Therefore, conditioning or limiting the analyses to records that were linked could therefore induce a protective relationship between the exposure and outcome, as we found in this analysis. One method to potentially correct for this bias is to use multiple imputation to handle missing values due to unlinked records,[115] which could employ the match weights from the linkage procedure to inform priors during the imputation process.[120, 121] We also found that the number of missed matches increased at higher thresholds which resulted in a decreased analytic sample size and thus statistical power as evidenced by larger standard errors and wider confidence intervals compared to lower thresholds. Our identification of bias towards a null association with gains in precision at these lower thresholds is substantiated by previous research that showed similar trends in settings with minimal linkage errors.[46, 115]

A strength of this analysis was the access to individual-level data collected in the PIRL software, clinics, and sero-surveys. This information is often only available to individuals performing the linkage and not to researchers conducting analyses[42, 44, 122, 123] and allowed us to have full control of the automated linkage process including data pre-processing to improve the quality of the variables used in the algorithm. Most of the identifiers used by the automated linkage algorithm had no or very little missing data, including names, year of birth, sex, village and sub-village. While the algorithm embedded in the PIRL software utilised a larger set of personal identifiers, this restricted set of variables has been shown to drive the success of the linkage algorithm in our PIRL software.[74]

There were some limitations. First, the magnitude of the tested association between the selected exposure and outcome was large in the gold standard data, which was probably why the conclusions of the primary regression analysis were similar in the automated linked datasets at the lower match score thresholds even after measurable attenuation in the estimate. It is likely that a more modest association found in the gold standard data would have resulted in a null association and therefore different conclusions as has been found in other studies.[117] Second, the relatively small sample size in the gold standard data did not allow us to assess linkage bias at match score thresholds higher than the 75th percentile.

5.7 CONCLUSIONS

Recently, there has been increased attention on how errors arising during the linkage process impacts inferences drawn from analyses using imperfectly matched data, but predominately in high-income countries with negligible linkage errors. We provided original evidence that the impact of linkage quality is similar in a low-income country setting with substantial linkage errors. Until future analyses investigate methods to adjust for these biases and provide more robust results using data with considerable linkage errors, our results suggest that other researchers in similar settings desiring to perform probabilistic record linkage should allocate resources toward PIRL or similar system.

6 Paper D. Non-disclosure of HIV testing history in population-based surveys: implications for the estimation of the UNAIDS 90-90-90 target

Christopher T. Rentsch[1], Georges Reniers[1,2], Richard Machemba[3], Emma Slaymaker[1], Milly Marston[1], Alison Wringe[1], Jeffrey W. Eaton[4], Annabelle Gourlay[1], Brian Rice[5], Chodziwadziwa Whiteson Kabudula[2], Mark Urassa[3], Jim Todd[1,3], Basia Żaba[1]

[1]Department of Population Health, London School of Hygiene & Tropical Medicine, London, UK
[2]MRC/Wits Rural Public Health and Health Transitions Research Unit (Agincourt), School of Public Health, Faculty of Health Sciences, University of the Witwatersrand, Johannesburg, South Africa
[3]The TAZAMA Project, National Institute for Medical Research, Mwanza, Tanzania
[4]Department of Infectious Disease Epidemiology, Imperial College London, London, UK
[5]Department of Social and Environmental Health Research, London School of Hygiene & Tropical Medicine, London, UK

London School of Hygiene & Tropical Medicine
Keppel Street, London WC1E 7HT
www.lshtm.ac.uk

Registry
T: +44(0)20 7299 4646
F: +44(0)20 7299 4656
E: registry@lshtm.ac.uk

LONDON
SCHOOL of
HYGIENE
&TROPICAL
MEDICINE

RESEARCH PAPER COVER SHEET

PLEASE NOTE THAT A COVER SHEET MUST BE COMPLETED FOR EACH RESEARCH PAPER INCLUDED IN A THESIS.

SECTION A – Student Details

Student	Christopher T. Rentsch
Principal Supervisor	Professor Basia Żaba and Dr. Georges Reniers
Thesis Title	Point-of-contact interactive record linkage between demographic surveillance and health facilities to measure patterns of HIV service utilisation in Tanzania

If the Research Paper has previously been published please complete Section B, if not please move to Section C

SECTION B – Paper already published

Where was the work published?			
When was the work published?			
If the work was published prior to registration for your research degree, give a brief rationale for its inclusion			
Have you retained the copyright for the work?*		Was the work subject to academic peer review?	

**If yes, please attach evidence of retention. If no, or if the work is being included in its published format, please attach evidence of permission from the copyright holder (publisher or other author) to include this work.*

SECTION C – Prepared for publication, but not yet published

Where is the work intended to be published?	*AIDS*
Please list the paper's authors in the intended authorship order:	Christopher T. Rentsch, Georges Reniers, Richard Machemba, Emma Slaymaker, Milly Marston, Alison Wringe, Jeffrey W. Eaton, Annabelle Gourlay, Brian Rice, Chodziwadziwa Whiteson Kabudula, Mark Urassa, Jim Todd, Basia Żaba
Stage of publication	Submitted (under review)

SECTION D – Multi-authored work

For multi-authored work, give full details of your role in the research included in the paper and in the preparation of the paper. (Attach a further sheet if necessary)	I contributed to the conception of the investigation, designed the study, created analysis plans, extracted and cleaned data, conducted the analysis, drafted the manuscript, collated feedback from co-authors, submitted manuscript, and liaised with journal editors

Student Signature: [redacted] **Date:** 8/8/2018

Supervisor Signature: [redacted] **Date:** 8/8/2018

6.1 OVERVIEW

The previous chapters have described the methodological aspects of PIRL. In this first of two chapters highlighting the utility of the PIRL data infrastructure, I estimate the accuracy of the ‘first 90’ (the proportion of PLHIV who are diagnosed) by augmenting the standard estimation method with linked, directly-observed HIV testing records. This paper serves as a critical first step in guiding UNAIDS and other stakeholders to formulate updated algorithms to better estimate widely used targets that assist programmes and organisations to track progress and prioritise further programme implementation.

Objective 3. To measure patterns of HIV service utilisation using the linked data infrastructure created by PIRL.

6.2 ABSTRACT

Background. The proportion of people living with HIV (PLHIV) who know their HIV status (the 'first 90' of the UNAIDS 90-90-90 targets) is often estimated using self-reported HIV testing history in population-based HIV serological surveys. Any bias in this estimate caused by non-disclosure of testing history has the potential to mislead organisations on where gaps exist in HIV care and treatment programmes.

Methods. We used three rounds of population-based HIV serological surveillance conducted between 2010 and 2016 from the Kisesa observational HIV cohort study in Tanzania and linked HIV testing history and medical records from a local HIV care and treatment clinic to identify participants who had received a previous diagnostic HIV test. We fitted generalised estimating equations logistic regression models to detect associations with non-disclosure of HIV testing history adjusting for demographic, behavioural, and clinical characteristics. We compared estimates of the 'first 90' using self-reported survey data only and augmented estimates using information from linked HIV testing history and clinical data to quantify absolute and relative bias in diagnosis based on self-report.

Results. Numbers of participants in each of the survey rounds ranged from 7,171 to 7,981 with an average HIV prevalence of 6.9%. Up to 33% of those who tested HIV-positive and 34% of those who tested HIV-negative did not correctly disclose their HIV testing history. The variable most associated with non-disclosure was not knowing or refusing to indicate whether a condom was used at last sex (OR 2.93, 95% CI 1.84, 4.67). The proportion of PLHIV who reported knowing their status increased from 34% in 2010 to 65% in 2016. Updating these estimates to include information from the linked data resulted in a relative bias between 9.4% and 12.4%.

Conclusions. In this population, relying on self-reported testing history in population-based HIV serological surveys under-estimated the percentage diagnosed by a relative factor of at least 9.4%. Given we did not capture HIV testing or care that occurred outside of the study area, bias may still be under-estimated. Research should be employed in other surveillance systems that benefit from linked data to investigate how bias may vary between settings.

6.3 Background

The effectiveness of HIV testing and counselling (HTC) services is principally measured by the number of people living with HIV (PLHIV) who know their HIV status.[124] These services are the gateway to receiving further HIV prevention, care, and treatment services. In 2014, the Joint United Nations Programme on HIV/AIDS (UNAIDS) launched a series of targets known as '90-90-90', which stated by 2020, 90% of all PLHIV will be diagnosed (the 'first 90'), 90% of people diagnosed with HIV will receive antiretroviral treatment (ART) (the 'second 90'), and 90% of people receiving ART will achieve viral suppression (the 'third 90').[1] Routinely updated estimates of the 90-90-90 targets assist programmes and organisations to track progress and prioritise further programme implementation.

In eastern and southern Africa, where more than half of global PLHIV[8, 9] are resident, estimates of the 'first 90' are derived from national population-based surveys that include HIV serological testing (henceforth termed, 'sero-surveys').[67] When the survey includes a question directly asking respondents to report their last HIV test result, the estimate of the 'first 90' in the year of the survey is simply the proportion of people who reported they were diagnosed with HIV at their last HIV test out of the total number who tested HIV-positive in the survey. If the survey does not include a direct question about the knowledge of HIV status, such as Demographic and Health Surveys (DHS),[125] the estimate of the 'first 90' in the year of the survey is the average of two indicators: (i) the percentage of people diagnosed with HIV in the survey who report ever having been tested for HIV and receiving their last test result (upper bound), and (ii) the percentage of PLHIV on ART as reported with national ART programme data (lower bound).[67] However, the accuracy of the upper bound estimate may be affected by respondents who are hesitant to report their HIV testing history, or by the training and ability of the interviewers to ask sensitive questions.[124, 126]

Inaccurate disclosure of HIV testing history in surveys may lead to bias in modelled estimates of the 'first 90.' In this report, we used data from a health and demographic surveillance site (HDSS) in northwest Tanzania with population-based HIV sero-surveillance and linked medical records to measure the extent of undisclosed HIV testing history among sero-survey participants who had attended a previous sero-survey or had previously registered for HIV care and treatment. We modelled associations with non-disclosure of HIV testing history and stratified results by HIV test result. Finally, we measured the discrepancy between an estimate of the 'first 90' using survey data and an augmented estimate using linked HIV testing history and medical records.

6.4 METHODS

6.4.1 Data sources

The Kisesa observational HIV cohort study was established in 1994 and is located in a rural ward in the Magu district of Mwanza region in northwest Tanzania.[68] The study conducts annual or bi-annual rounds of household-based demographic surveillance (31 to date) and has completed eight rounds of population-based HIV sero-surveys, in which adults aged 15 years or older living in the Kisesa HDSS coverage area are invited to attend a temporary village-based clinic for a personal interview and to give a finger-prick blood specimen to be anonymously tested for HIV. Beginning in 2007, voluntary HTC services were offered at the sero-survey clinics to those who wanted to know their HIV status, based on a second, separate blood specimen for the HIV test. Sero-survey records indicate whether the participant received their test result. Participants' records from all sero-survey rounds are linked with a unique permanent identifier, and temporary household-based identifiers from the HDSS are also cross-referenced on each sero-survey record. This analysis included all sero-survey participants in each of the three most recent rounds: Sero 6 (2010), Sero 7 (2013) and Sero 8 (2016).

A government-run health centre is located within the Kisesa HDSS catchment area, including an HIV care and treatment centre (CTC). The CTC databases have been fully digitised, and data clerks regularly update and run data checks on these data. Medical records from the CTC have been linked to the HDSS database using point-of-contact interactive record linkage (PIRL), described elsewhere.[73, 74] Ethical approval for each of the sero-survey rounds and the PIRL study was obtained from the Tanzanian National Institute for Medical Research Lake Zone Institutional Review Board and the London School of Hygiene & Tropical Medicine. Informed written consent was obtained from all participants in both studies.

6.4.2 Outcome

The Kisesa sero-surveys, akin to DHS, include indirect questions about knowledge of HIV status by asking about HIV testing history. To determine risk factors associated with non-disclosure of HIV testing history, the regression analyses only included participants with evidence of a previous diagnostic HIV test (during a sero-survey or as noted in their medical records). UNAIDS estimates the upper bound of the 'first 90' using the following two questions: (i) "Have you ever had HIV testing and counselling?"; and (ii) "Did you find out your test results after your last test?" Those who responded affirmatively to both questions were classified as disclosing their HIV testing history. All others were classified as having an undisclosed HIV testing history.

6.4.3 Covariates

We extracted demographic, behavioural, and clinical characteristics from each sero-survey round. Demographic variables included sex, age, education level (no primary, some primary, or primary or higher), sub-village of residence (rural, peri-urban, or urban), whether the sub-village of residence has a road, and marital status (never or ever married/cohabitated). Behavioural variables included the reported number of sex partners in the last 12 months and reported condom use at last sex. Clinical variables included whether participants visited a health provider (e.g., hospital, health centre, dispensary, antenatal clinic, vaccination clinic, visit from home-based care worker, private pharmacy, or traditional healer) in the last 12 months, and for those who tested HIV-positive, whether the participant had initiated ART prior to the sero-survey as noted in CTC records, and the time since HIV diagnosis using the first HIV-positive test date in a sero-survey. For individuals who did not have a recorded positive HIV test date in a sero-survey, we used the first HIV-positive test date as listed in their medical records.

6.4.4 Statistical analyses

Chi-square and Fisher's exact tests were used to assess differences between participants with and without a previous diagnostic HIV test, as well as between participants who did and did not disclose their HIV testing history. Some participants attended more than one sero-survey round. Thus, generalised estimating equations (GEE) logistic regression models were used to account for the correlated data. We fit crude and adjusted GEE logistic regression models for all participants to detect differences in disclosure of HIV testing history by HIV test result. We also fit analogous models limited to participants who tested HIV-positive during the sero-survey. Statistical analyses were performed using SAS version 9.4 (SAS Institute Inc., Cary, NC, USA).

6.4.5 Estimating bias in the 'first 90'

Following UNAIDS guidelines,[67] we estimated the 'first 90' for each sero-survey round by averaging the proportion of participants who tested HIV-positive who reported ever having been tested for HIV and receiving their last test result (upper bound), and the percentage of adult PLHIV on ART as reported with national ART programme data (lower bound). Estimates of the percentage of adult PLHIV on ART for each year coinciding with a sero-survey round were obtained from UNAIDS AIDSinfo[9]; however, these were national estimates as longitudinal, sub-national ART coverage estimates were not available. We then updated the estimate of the 'first 90' by adding information from the linked HIV testing history and medical records. For each sero-survey round, we calculated two measures of bias:

$$\text{absolute bias: } Bias_{abs} = \hat{u} - \hat{o} \quad (6.1)$$

$$\text{relative bias: } Bias_{rel} = (\hat{u} - \hat{o})/\hat{u} \quad (6.2)$$

where $\hat{o}$ is the original estimate of the 'first 90' using self-reported survey responses, and $\hat{u}$ is the updated estimate of the 'first 90' by augmenting $\hat{o}$ with linked data. In a sensitivity analysis to account for possible misspecification of ART coverage, we determined the robustness of the bias estimates by increasing and decreasing the national ART coverage estimates by half, and by setting them to their theoretical maximum equal to the upper bound.

6.5 RESULTS

There were 7,981 participants (61% female) in Sero 6, 7,607 participants (62% female) in Sero 7, and 7,161 participants (62% female) in Sero 8, of whom 860 (10.8%), 1,232 (16.2%), and 1,786 (24.9%) respectively had received a diagnostic HIV test in a previous sero-survey round or registered in the CTC (Figure 6.1).

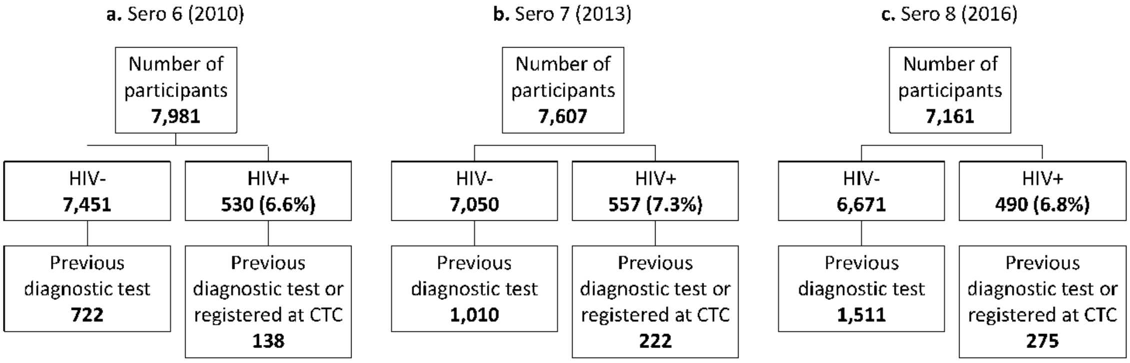

a. Sero 6 (2010)
Number of participants 7,981
HIV- 7,451
HIV+ 530 (6.6%)
Previous diagnostic test 722
Previous diagnostic test or registered at CTC 138
b. Sero 7 (2013)
Number of participants 7,607
HIV- 7,050
HIV+ 557 (7.3%)
Previous diagnostic test 1,010
Previous diagnostic test or registered at CTC 222
c. Sero 8 (2016)
Number of participants 7,161
HIV- 6,671
HIV+ 490 (6.8%)
Previous diagnostic test 1,511
Previous diagnostic test or registered at CTC 275

Figure 6.1: Number of participants with previous diagnostic testing or HIV care in three rounds of HIV serological surveys in Kisesa, Tanzania, by HIV status, 2010-2016

Among participants who tested HIV-negative during a sero-survey, those who had previously received a diagnostic HIV test (i.e., repeat HIV-negative testers) were older, had more education, were from more urbanised areas, reported more sex partners in the last 12 months, reported less condom use at last sex, and reported more health service use than those who had not previously received a diagnostics HIV test (all p<0.05) (Supplemental Table 1 in Appendix 10.9.2). The differences between participants with and without a previous HIV diagnostic test among people who tested HIV-positive during a survey were narrower (Supplemental Table 2 in Appendix 10.9.2). In this group, the only statistically significant difference found in multiple survey rounds was that participants who had previously received a diagnostic HIV test were from more urbanised areas than those who had not previously received a diagnostics HIV test (p<0.04).

6.5.1 Non-disclosure of HIV testing history

Among participants with a previous diagnostic HIV test and who tested HIV-positive in the sero-survey, 39/138 (28%) in Sero 6, 73/222 (33%) in Sero 7, and 64/275 (23%) in Sero 8 did not disclose their HIV testing history (Figure 6.2). Among participants with a previous diagnostic HIV test and who tested HIV-negative in the sero-survey, 142/722 (20%) in Sero 6, 340/1,010 (34%) in Sero 7, and 352/1,511 (23%) in Sero 8 did not disclose their HIV testing history. In bivariate analyses, there was a statistically significant difference between the level of non-disclosure of HIV testing history by HIV test result in Sero 6 (p=0.02), but not in Sero 7 (p=0.82) or Sero 8 (p=0.99). In addition, participants with less education and who reported not knowing or refused to report number of sex partners in the last 12 months were more likely to not disclose their HIV testing history in all three sero-survey rounds (all p<0.03) (Table 6.1).

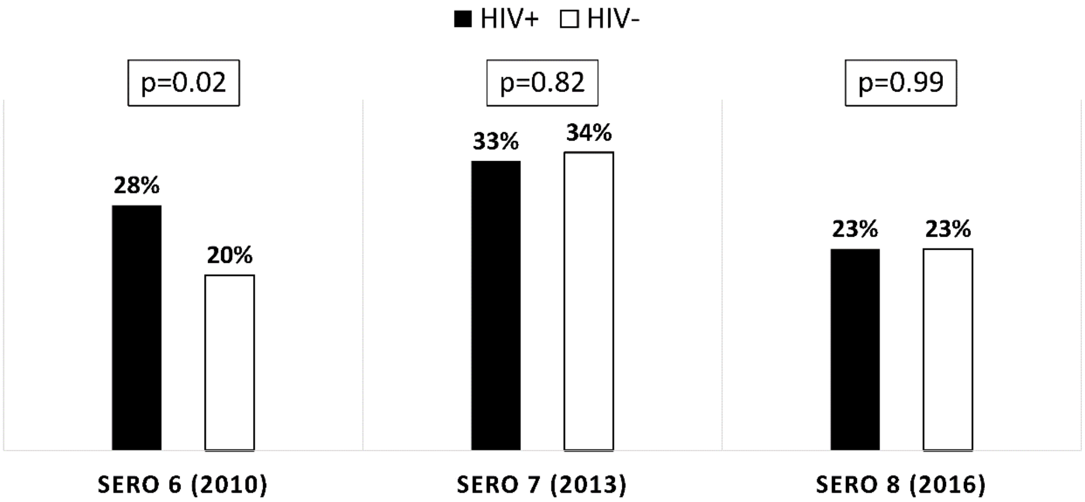


Figure 6.2: Proportion of population-based HIV serological survey participants in Kisesa, Tanzania who did not disclose their HIV testing history, by HIV test result

Table 6.1: Characteristics of survey participants with evidence of previous HIV testing, by disclosure of testing history

	Sero 6 (2010)			Sero 7 (2013)			Sero 8 (2016)		
	Disclosed	**Undisclosed**		**Disclosed**	**Undisclosed**		**Disclosed**	**Undisclosed**	
	n=679	**n=181**	**p-value**	**n=819**	**n=413**	**p-value**	**n=1,370**	**n=416**	**p-value**
HIV test result in sero									
HIV+	99 (72)	39 (28)	0.0233	149 (67)	73 (33)	0.8235	211 (77)	64 (23)	0.9933
HIV-	580 (80)	142 (20)		670 (66)	340 (34)		1,159 (77)	352 (23)	
Demographic characteristic									
Sex									
Female	407 (80)	103 (20)	0.4817	470 (62)	292 (38)	<0.0001	854 (74)	297 (26)	0.0003
Male	270 (78)	77 (22)		349 (75)	117 (25)		516 (82)	115 (18)	
Age, years									
15-29	181 (79)	49 (21)	0.6757	170 (64)	94 (36)	0.7189	301 (72)	119 (28)	0.0020
30-49	377 (80)	95 (20)		444 (67)	219 (33)		683 (80)	169 (20)	
50+	121 (77)	37 (23)		205 (67)	100 (33)		386 (75)	128 (25)	
Education level									
No primary	136 (70)	59 (30)	0.0014	158 (51)	152 (49)	<0.0001	315 (67)	158 (33)	<0.0001
Some primary	88 (80)	22 (20)		97 (66)	51 (24)		156 (78)	43 (22)	
Primary or higher	455 (82)	100 (18)		564 (73)	210 (27)		899 (81)	215 (19)	
Sub-village of residence, type									
Rural	355 (75)	119 (25)	0.0048	352 (59)	242 (41)	<0.0001	680 (76)	218 (24)	0.3432
Peri-urban	184 (85)	33 (15)		234 (68)	110 (32)		358 (79)	94 (21)	
Urban	140 (83)	29 (17)		233 (79)	61 (21)		332 (76)	104 (24)	
Sub-village of residence, has road									
No	399 (75)	133 (25)	0.0003	413 (60)	280 (40)	<0.0001	762 (76)	243 (24)	0.3145
Yes	280 (85)	48 (15)		406 (75)	133 (25)		608 (78)	173 (22)	
Current marital status									
Never married/cohabitated	66 (71)	27 (29)	0.0454	88 (63)	52 (37)	0.3352	145 (68)	69 (32)	0.0010
Ever married/cohabitated	613 (80)	154 (20)		731 (67)	361 (33)		1,225 (78)	347 (22)	

Table 6.1: Characteristics of survey participants with evidence of previous HIV testing, by disclosure of testing history

	Sero 6 (2010)			Sero 7 (2013)			Sero 8 (2016)		
	Disclosed n=679	Undisclosed n=181	p-value	Disclosed n=819	Undisclosed n=413	p-value	Disclosed n=1,370	Undisclosed n=416	p-value
Behavioural characteristic									
Number of sex partners in last 12 months									
Don't know/refused	8 (44)	10 (56)	0.0034	17 (46)	20 (54)	0.0003	46 (64)	26 (36)	0.0279
0	90 (81)	21 (19)		102 (58)	74 (42)		202 (75)	66 (25)	
1	462 (80)	116 (20)		599 (68)	288 (32)		1,005 (77)	298 (23)	
2 or more	119 (78)	34 (22)		101 (77)	31 (23)		117 (82)	26 (18)	
Condom use at last sex									
Don't know/refused	96 (76)	31 (24)	0.4236	619 (64)	347 (36)	0.0001	155 (69)	69 (31)	0.0009
No	536 (79)	141 (21)		155 (72)	61 (28)		1,133 (77)	336 (23)	
Yes	47 (84)	9 (16)		45 (90)	5 (10)		82 (88)	11 (12)	
Clinical characteristic									
Visited health provider in last 12 months									
No	84 (79)	22 (21)	0.9373	105 (55)	86 (45)	0.0002	343 (71)	142 (29)	0.0003
Yes	595 (79)	159 (21)		714 (69)	327 (31)		1,027 (79)	274 (21)	
HIV+ only									
Time since HIV diagnosis, years									
First positive test during sero	23 (61)	15 (39)	0.2204	34 (69)	15 (31)	0.0358	30 (64)	17 (36)	0.1089
<5	48 (73)	18 (27)		62 (62)	38 (38)		79 (77)	24 (23)	
5-9	12 (86)	2 (14)		34 (64)	19 (36)		74 (81)	17 (19)	
10+	16 (80)	4 (20)		19 (95)	1 (5)		28 (82)	6 (18)	
Initiated antiretroviral therapy									
Yes	50 (79)	13 (21)	0.0682	55 (66)	28 (34)	0.8346	53 (75)	18 (25)	0.6308
No	49 (65)	26 (35)		94 (68)	45 (32)		158 (77)	46 (23)	

Abbreviations: HIV - human immunodeficiency virus; sero - HIV serological survey

Note: all statistics are given in n(row %); differences tested for significance with chi-square (χ^2) and Fisher's exact tests

6.5.2 Regression models

After accounting for the correlated data, there was no evidence that participants who tested HIV-positive differentially disclosed their HIV testing history compared to those who tested HIV-negative during a sero-survey (crude odds ratio [cOR] 1.10, 95% confidence interval [CI] 0.90, 1.34) (Table 6.2). After adjusting for other covariates, non-disclosure of HIV testing history was associated with sex (females vs. males: adjusted odds ratio [aOR] 1.54, 95% CI 1.27, 1.88), education (no primary vs. primary or higher: aOR 2.01, 95% CI 1.67, 2.41), marital status (never vs. ever married/cohabitated: aOR 1.88, 95% CI 1.43, 2.47), and reported condom use at last sex (do not know/refused to answer vs. yes: aOR 2.93, 95% CI 1.84, 4.67; no vs. yes: aOR 2.16, 95% CI 1.39, 3.36). There were no significant associations between non-disclosure of HIV testing history and age, sub-village of residence, and reported number of sex partners in the last 12 months, after adjusting for other covariates.

Among participants who tested HIV-positive during the sero-survey, participants' whose positive test during the survey was their first were three times more likely (aOR 3.03, 95% CI 1.39, 6.62) and those with a first recorded diagnosis of less than five years were 2.4 times more likely (aOR 2.36, 95% CI 1.18, 4.69) to not disclose their HIV testing history than their counterparts who had first been diagnosed over 10 years prior to the survey. There was no evidence of an association between non-disclosure of HIV testing history and whether the participant was on ART (aOR 1.08, 95% CI 0.72, 1.62).

Table 6.2: Associations with non-disclosure of HIV testing history among participants of population-based HIV serological surveys

	All participants, n=2,747				**HIV+ only, n=454**		
	cOR (95% CI)		**aOR (95% CI)**		**cOR (95% CI)**		**aOR (95% CI)**
HIV test result in any sero							
HIV+ vs. HIV-	1.10 (0.90, 1.34)		1.08 (0.87, 1.34)				
Sero round							
Sero 8	1.11 (0.92, 1.34)		0.86 (0.66, 1.13)		0.76 (0.48, 1.20)		0.75 (0.45, 1.24)
Sero 7	1.87 (1.53, 2.27)	***	1.30 (0.95, 1.77)		1.22 (0.77, 1.94)		0.99 (0.55, 1.78)
Sero 6	1		1		1		1
HIV test result by sero round							
HIV+ vs. HIV-, Sero 8	1.00 (0.74, 1.35)		0.94 (0.68, 1.30)				
HIV+ vs. HIV-, Sero 7	0.95 (0.69, 1.29)		0.88 (0.64, 1.21)				
HIV+ vs. HIV-, Sero 6	1.59 (1.06, 2.40)	*	1.53 (0.99, 2.36)				
Demographic characteristic							
Sex							
Female	1.46 (1.25, 1.71)	***	1.54 (1.27, 1.88)	***	1.38 (0.93, 2.04)		1.39 (0.87, 2.23)
Male	1		1		1		1
Age, years							
15-29	1.25 (1.05, 1.48)	*	1.18 (0.97, 1.44)		1.87 (1.22, 2.87)	**	1.46 (0.92, 2.32)
30-49	1		1		1		1
50+	1.15 (0.96, 1.38)		1.08 (0.88, 1.32)		1.36 (0.87, 2.12)		1.43 (0.89, 2.29)
Education level							
No primary	2.19 (1.85, 2.58)	***	2.01 (1.67, 2.41)	***	1.65 (1.11, 2.44)	*	1.36 (0.88, 2.10)
Some primary	1.23 (0.98, 1.56)		1.26 (0.99, 1.61)		1.24 (0.70, 2.19)		1.25 (0.68, 2.28)
Primary or higher	1		1		1		1
Sub-village of residence, type							
Rural	1.50 (1.24, 1.81)	***	1.19 (0.89, 1.60)		1.47 (0.95, 2.27)		1.41 (0.66, 3.01)
Peri-urban	1.10 (0.88, 1.37)		0.94 (0.73, 1.22)		1.21 (0.74, 1.97)		1.01 (0.53, 1.92)
Urban	1		1		1		1

Table 6.2: Associations with non-disclosure of HIV testing history among participants of population-based HIV serological surveys

	All participants, n=2,747				**HIV+ only, n=454**			
	cOR (95% CI)		**aOR (95% CI)**		**cOR (95% CI)**		**aOR (95% CI)**	
Sub-village of residence, has road								
No	1.51 (1.30, 1.76)	***	1.35 (1.06, 1.72)	*	1.27 (0.88, 1.82)		1.04 (0.55, 1.95)	
Yes	1		1		1		1	
Current marital status								
Never married/cohabitated	1.48 (1.20, 1.83)	***	1.88 (1.43, 2.47)	***	1.77 (1.00, 3.15)	*	1.46 (0.74, 2.89)	
Ever married/cohabitated	1		1		1		1	
Behavioural characteristic								
Number of sex partners in last 12 months								
Don't know/refused	2.87 (1.88, 4.36)	***	1.32 (0.83, 2.08)		1.89 (0.67, 5.29)		0.99 (0.29, 3.39)	
0	1.48 (1.10, 1.99)	*	0.82 (0.56, 1.19)		1.10 (0.55, 2.21)		0.69 (0.27, 1.77)	
1	1.25 (0.97, 1.59)		1.00 (0.75, 1.33)		1.08 (0.57, 2.05)		0.87 (0.39, 1.90)	
2 or more	1		1		1		1	
Condom use at last sex								
Don't know/refused	3.48 (2.29, 5.29)	***	2.93 (1.84, 4.67)	***	4.94 (1.50,16.31)	**	4.46 (1.24,16.04)	*
No	1.99 (1.32, 3.02)	**	2.16 (1.39, 3.36)	***	3.71 (1.14,12.09)	*	3.19 (0.91,11.14)	
Yes	1		1		1		1	
Clinical characteristic								
Visited health provider in last 12 months								
No	1.44 (1.21, 1.70)	***	1.50 (1.24, 1.80)	***	1.53 (0.96, 2.45)		1.70 (1.01, 2.85)	*
Yes	1		1		1		1	
HIV+ only								
Times since HIV diagnosis, years								
First positive test during sero					3.33 (1.61, 6.87)	**	3.03 (1.39, 6.62)	**
<5					2.65 (1.37, 5.13)	*	2.36 (1.18, 4.69)	*
5-9					1.86 (0.93, 3.75)		1.75 (0.84, 3.64)	
10+					1		1	

Table 6.2: Associations with non-disclosure of HIV testing history among participants of population-based HIV serological surveys

	All participants, n=2,747		**HIV+ only, n=454**	
	cOR (95% CI)	**aOR (95% CI)**	**cOR (95% CI)**	**aOR (95% CI)**
Initiated antiretroviral therapy				
Yes			1.01 (0.70, 1.47)	1.09 (0.72, 1.64)
No			1	1

Abbreviations: cOR - crude unadjusted odds ratio; aOR - adjusted odds ratio; CI - confidence interval; HIV - human immunodeficiency virus; sero - HIV serological survey

*p<0.05; **p<0.01; ***p<0.001

6.5.3 Implications for the estimate of the ‘first 90’

The number of participants who tested HIV-positive was 530 (6.6%) in Sero 6, 557 (7.3%) in Sero 7, and 490 (6.8%) in Sero 8. Of these, the number of participants who reported having ever been tested for HIV and received their last HIV test result (the upper bound of the ‘first 90’ estimate) was 268 (50.6%) in Sero 6, 275 (49.4%) in Sero 7, and 328 (66.9%) in Sero 8. Estimates of ART coverage (the lower bound of the ‘first 90’ estimate) in Tanzania for each year coinciding with each sero-survey round were 18% in 2010, 39% in 2013, and 62% in 2016. Thus, cross-sectional estimates of the ‘first 90’ using self-reported survey data combined with the national ART estimates were 34.3% in 2010, 44.2% in 2013, and 64.5% in 2016 (Figure 6.3). However, there was evidence from the linked records that some patients who tested positive during the sero-survey did not correctly disclose their HIV testing history (39 participants in Sero 6, 70 participants in Sero 7, and 65 participants in Sero 8). After augmenting the estimates of the upper bound with this information, the estimate of the ‘first 90’ increased to 38.0% in 2010, 50.5% in 2013, and 71.1% in 2016. These increases corresponded to an absolute bias in the ‘first 90’ of 3.7 percentage points in Sero 6, 6.2 percentage points in Sero 7, and 6.7 percentage points in Sero 8, and a relative bias of 9.6% in Sero 6, 12.4% in Sero 7, and 9.4% in Sero 8.

In sensitivity analyses, decreasing estimates of ART coverage by half resulted in increased estimates of relative bias to 10.9% in Sero 6, 15.4% in Sero 7, and 12.0% in Sero 8; increasing estimates of ART coverage by 50% resulted in decreased estimates relative bias to 8.6% in Sero 6, 10.4% in Sero 7, and 7.7% in Sero 8; and setting the lower bound equal to the upper bound resulted in the highest estimates of relative bias to 12.6% in Sero 6, 20.2% in Sero 7, and 16.6% in Sero 8.

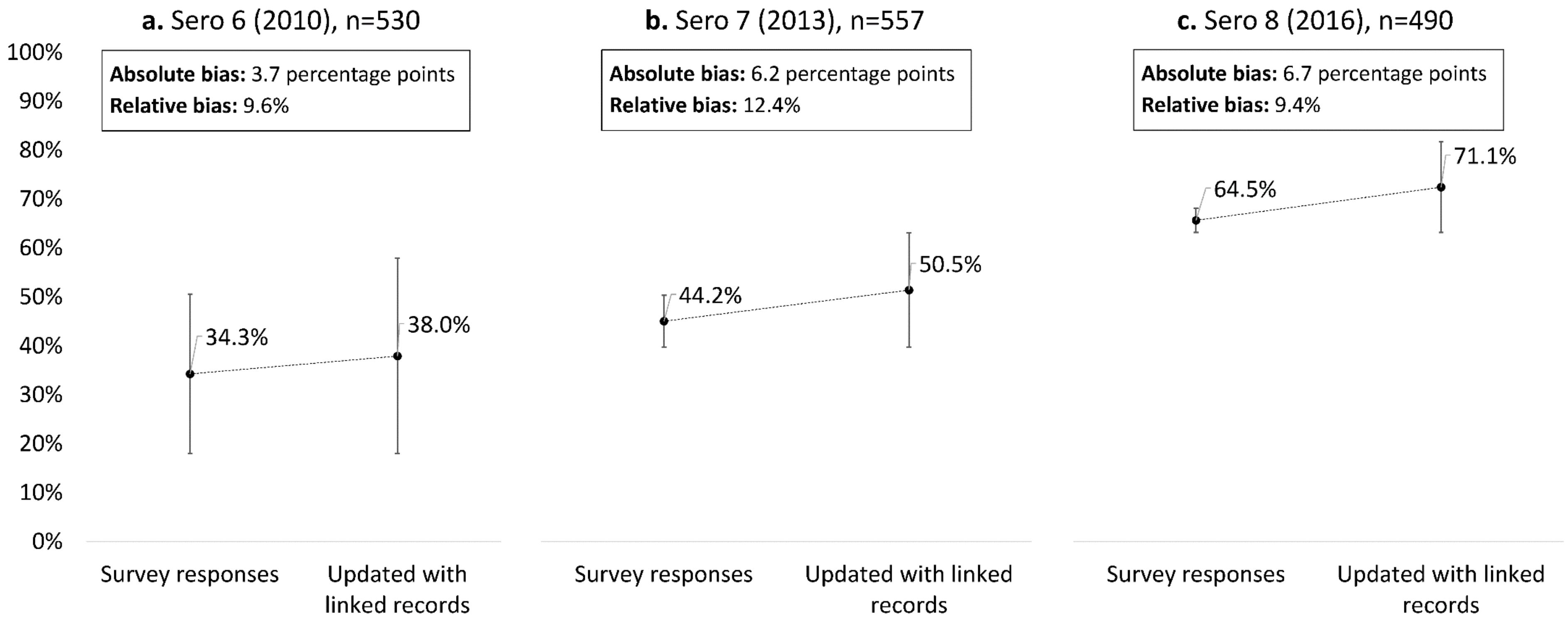


Figure 6.3: Estimates of the proportion of PLHIV who know their status (the 'first 90' of the UNAIDS 90-90-90 target) using population-based HIV serological survey data and linked clinic data in Kisesa, Tanzania, 2010-2016

6.6 Discussion

Using the data infrastructure available in the Kisesa HDSS, non-disclosure of previous HIV testing history resulted in a relative bias in the proportion of PLHIV who reported knowing their HIV sero-status (the 'first 90' of the widely used UNAIDS 90-90-90 targets) between 9.4% to 12.4% during the 2010-2016 population-based surveys. The size of the relative bias was directly related to the proportion of participants who did not disclose their HIV testing history – more non-disclosures resulted in higher bias. Importantly, our estimates of the number of individuals who received a previous HIV diagnostic test or HIV care are likely to be under-estimated since we did not capture HIV testing that occurs outside of the sero-surveys or HIV care outside of the Kisesa HDSS area. Therefore, our augmented estimates of the 'first 90' and the resulting bias estimates are likely under-estimates. In populations where non-disclosure of testing history is significant, the current UNAIDS estimation of the 'first 90' using self-reported survey responses are likely to be biased downward. This bias has the potential to lead large, international organisations, such as the US President's Emergency Plan for AIDS Relief (PEPFAR), to misdiagnose the gaps in HIV care and treatment programmes and misallocate resources.

The upper bound of the UNAIDS estimate of the 'first 90' (proportion reporting ever tested and receiving last test result) includes individuals who may truly not know their HIV-positive status because the last test result they received was negative (i.e., sero-converters), which is why UNAIDS use this as an upper bound rather than a direct estimate of knowledge of HIV status. The rationale for using ART coverage as the lower bound is that the proportion of PLHIV who are diagnosed ('first 90') cannot be lower than the proportion of PLHIV receiving ART. We found that changes in ART coverage drove the increase in the 'first 90' over the study period as it increased from 18% to 62% (a 224% relative change) compared to an increase from 58% to 80% (a 39% relative change) in the upper bound. While estimates of ART coverage were not available at the sub-national level, we accounted for potential measurement errors in sensitivity analyses and found that setting ART coverage equal to the upper bound increased the estimate of relative bias in the 'first 90' to 12.6%-20.2% over the study period. This finding is consistent with an analysis using national survey data from neighbouring Kenya in 2012 that showed a 16.5% relative increase in the proportion of diagnosed PLHIV after correcting the estimate for undisclosed HIV infection with available biomarker data.[127]

The proportion of PLHIV who reported knowing their HIV sero-status doubled from 34% in 2010 to 65% in 2016 in this rural Tanzanian community. There is a lack of longitudinal, national or sub-national estimates of the 'first 90' or its components available to which we can compare our findings. UNAIDS has estimated that 70% of PLHIV in Tanzania knew their sero-status in 2016.[8, 9] which is similar to our sub-national estimate in that year but cannot be directly compared. A national survey conducted by the National Bureau of Statistics in 2011 showed the 54.1% of those surveyed in the Mwanza region had ever been tested for HIV and received their last test result,[128] which is reasonably consistent with our finding of 50.6% of participants surveyed in the 2010 survey.

There were substantial levels of non-disclosure of HIV testing history in this sample. Between 1 in 3 and 1 in 5 participants did not accurately report their HIV testing history during a sero-survey. There was no evidence that participants who tested HIV-positive were more or less likely to disclose their HIV testing history than those who tested HIV-negative after adjusting for other factors. In unadjusted analyses, not knowing or refusing to indicate the number of sex partners in the last 12 months or condom use at last sex had the highest associations with non-disclosure. In the adjusted model, not knowing or refusing to indicate whether a condom was used at last sex remained independently associated with non-disclosure. Being prepared to answer sensitive questions about sexual behaviour may indicate a person's willingness to be open about other sensitive topics, such as HIV testing history. Those who reported no condom use at last sex were more likely to be female, which was also associated with non-disclosure of HIV testing history. Several studies have shown women perceive and are targets of HIV-related stigma more than men,[129-132] which may impede their willingness to discuss their testing history. In addition, the women in our sample were significantly younger and reported less education than men, both factors which were associated with non-disclosure of HIV testing history. Even though women had higher participation rates in the Kisesa sero-surveys than men, we hypothesize that men who self-select to participate may be more comfortable discussing HIV testing history, and more broadly that participation may not correlate with disclosure equally among men and women.

In a model limited to participants who tested HIV-positive during a sero-survey, those who were first diagnosed with HIV infection during the sero-survey and those with a recent diagnosis of less than five years were up to three times more likely to not disclose their HIV infection compared to their counterparts who had lived with HIV for at least 10 years prior to the survey. There is some evidence among PLHIV in Tanzania that time since HIV diagnosis is negatively correlated with internalised stigma,[133] and that HIV-related stigma is significantly associated with concealment of HIV status.[133, 134] Therefore, individuals who have been aware of their positive HIV status for longer durations of time may have had less stigma regarding their HIV status and were more comfortable to report their HIV testing history.

Our study had limitations. First, we had a relatively small sample size of PLHIV and therefore lacked sufficient power detect significant associations with non-disclosure of HIV testing history among PLHIV. Second, participation rates in the sero-surveys have declined over time to 43% of eligible adults during the most recent survey in 2016. Individuals who choose to participate may be differentially inclined to report HIV testing history than those who do not participate. Comparative research should be employed in other HDSS sites to investigate how the corrective factor in the measurement of the ‘first 90’ may vary between settings.

6.7 Conclusions

There was substantial non-disclosure of previous HIV testing history in population-based surveys from Tanzania, which resulted in an under-estimate of the first UNAIDS 90-90-90 target between 10% and 20%. There were likely previous HIV diagnostic tests not captured in this analysis, and therefore the bias estimates are likely to still be under-estimated. The factor most associated with non-disclosure of HIV testing history was refusing to answer other sensitive questions, a finding that could potentially be used to augment estimates of the ‘first 90’ derived from other population-based surveys that do not benefit from linked HIV testing and medical records. Comparative research should be employed in other HDSS sites that benefit from linked HIV testing and clinical data to investigate how the corrective factor may vary between settings.

6.8 Supplementary material

All supplementary material for this publication can be found in Appendix 10.9.2.

- Supplemental Table 1. Characteristics among participants of population-based HIV serological surveys in Magu, Tanzania who tested HIV-negative, by whether they had evidence of a previous diagnostic HIV test
- Supplemental Table 2. Characteristics among participants of population-based HIV serological surveys in Magu, Tanzania who tested HIV-positive, by whether they had evidence of a previous diagnostic HIV test

7 Paper E. Linkage to care and antiretroviral therapy initiation by testing modality among individuals newly diagnosed with HIV in Tanzania, 2014-2017

Christopher T. Rentsch[1], Alison Wringe[1], Richard Machemba[2], Denna Michael[2], Mark Urassa[2], Jim Todd[1,2], Georges Reniers[1,3], Basia Żaba[1]

[1]Department of Population Health, London School of Hygiene & Tropical Medicine, London, UK
[2]The TAZAMA Project, National Institute for Medical Research, Mwanza, Tanzania
[3]MRC/Wits Rural Public Health and Health Transitions Research Unit (Agincourt), School of Public Health, Faculty of Health Sciences, University of the Witwatersrand, Johannesburg, South Africa

London School of Hygiene & Tropical Medicine
Keppel Street, London WC1E 7HT
www.lshtm.ac.uk

Registry
T: +44(0)20 7299 4646
F: +44(0)20 7299 4656
E: registry@lshtm.ac.uk

LONDON
SCHOOL of
HYGIENE
&TROPICAL
MEDICINE

RESEARCH PAPER COVER SHEET

PLEASE NOTE THAT A COVER SHEET MUST BE COMPLETED FOR EACH RESEARCH PAPER INCLUDED IN A THESIS.

SECTION A – Student Details

Student	Christopher T. Rentsch
Principal Supervisor	Professor Basia Żaba and Dr. Georges Reniers
Thesis Title	Point-of-contact interactive record linkage between demographic surveillance and health facilities to measure patterns of HIV service utilisation in Tanzania

If the Research Paper has previously been published please complete Section B, if not please move to Section C

SECTION B – Paper already published

Where was the work published?			
When was the work published?			
If the work was published prior to registration for your research degree, give a brief rationale for its inclusion			
Have you retained the copyright for the work?*		Was the work subject to academic peer review?	

**If yes, please attach evidence of retention. If no, or if the work is being included in its published format, please attach evidence of permission from the copyright holder (publisher or other author) to include this work.*

SECTION C – Prepared for publication, but not yet published

Where is the work intended to be published?	*Tropical Medicine & International Health*
Please list the paper's authors in the intended authorship order:	Christopher T. Rentsch, Alison Wringe, Richard Machemba, Denna Michael, Mark Urassa, Jim Todd, Georges Reniers, Basia Żaba
Stage of publication	Submitted (under review)

SECTION D – Multi-authored work

For multi-authored work, give full details of your role in the research included in the paper and in the preparation of the paper. (Attach a further sheet if necessary)	I contributed to the conception of the investigation, designed the study, created analysis plans, extracted and cleaned data, conducted the analysis, drafted the manuscript, collated feedback from co-authors, submitted manuscript, and liaised with journal editors

Student Signature: [redacted] **Date:** 8/8/2018

Supervisor Signature [redacted] **Date:** 8/8/2018

7.1 Objective

The previous chapter demonstrated the utility of the linked data infrastructure created by PIRL by obtaining and measuring bias in longitudinal, population-level estimates of the 'first 90' (the proportion of PLHIV who were diagnosed). In this chapter, I further establish the value of PIRL by measuring linkage to care and ART initiation rates (the 'second 90') among individuals newly diagnosed between 2014 and 2017.

Objective 3. To measure patterns of HIV service utilisation using the linked data infrastructure created by PIRL.

7.2 ABSTRACT

Objective. To measure linkage to care and antiretroviral therapy (ART) initiation among individuals newly diagnosed with HIV in a rural Tanzanian community.

Methods. We included all new HIV diagnoses of adults made between 2014-2017 during community- or facility-based HIV testing and counselling (HTC) in a rural ward in northwest Tanzania. Community-based HTC included population-level HIV serological testing (sero-survey), and facility-based HTC included a stationary, voluntary HTC clinic (VCT) and an antenatal clinic offering provider-initiated HTC (ANC-PITC). Cox regression models were used to compare linkage to care rates by testing modality and identify associated factors. Among those in care, we compared initial CD4 cell counts and ART initiation rates by testing modality.

Results. A total of 411 adults were newly diagnosed, of whom 10% (27/265 sero-survey), 18% (3/14 facility-based ANC-PITC), and 53% (68/129 facility-based VCT) linked to care within 90 days. Individuals diagnosed using facility-based VCT were seven times (95% CI: 4.5-11.0) more likely to link to care than those diagnosed in the sero-survey. We found no difference in linkage rates between those diagnosed using facility-based ANC-PITC and sero-survey (p=0.26). Among individuals in care, 63% of those in the sero-survey had an initial CD4 count >350 cells/mm^3 compared to 29% of those using facility-based VCT (p=0.02). The proportion who initiated ART within one year of linkage to care was similar for both groups (94% sero-survey vs. 85% facility-based VCT; p=0.16).

Conclusions. Community-based sero-surveys are important for earlier diagnosis of HIV-positive individuals; however, interventions are essential to facilitate linkage to care.

7.3 BACKGROUND

HIV testing and counselling (HTC) is the first critical step for subsequent linkage to care and initiating ART. However, linkage to care following a positive HIV diagnosis remains low in sub-Saharan Africa. A 2015 meta-analysis found that linkage to care within 12 months of diagnosis was only 61% (95% confidence interval [CI] 48-72%) among individuals diagnosed using facility-based, voluntary HTC (VCT) and 55% (95% CI 39-71%) among those diagnosed using facility-based, provider-initiated HTC (PITC).[135] Further, linkage to care may vary by region throughout the region,[10] emphasising the need for locally appropriate interventions to improve linkage to care and subsequent access to ART.

In order to expand access to HIV testing and increase linkage with care and treatment services, the World Health Organisation (WHO) recommended community-based HTC with facilitated linkage to care services (for example, a lay-counsellor follow-up to encourage a clinic visit) in addition to traditional, facility-based HTC.[65] Community-based HTC includes services that are delivered using mobile and home-based approaches thus removing structural, logistical, and social barriers to HTC.[66] While community-based HTC can increase the number of individuals who know their HIV status,[12, 136-138] it may also increase the proportion of people living with HIV who know their status but do not link to HIV care services. A 2015 systematic review found that only 30% of individuals diagnosed with HIV using mobile and home-based HTC without facilitated linkage were linked to care within 12 months.[135] Major limitations of previous systematic reviews[135, 139] on linkage to care following community- and facility-based HTC are that linkage outcomes were reported without differentiation between newly and previously diagnosed HIV-positive individuals and the use of both directly observed and self-reported linkage to care. Individuals who previously tested HIV-positive and have not yet linked to care are likely to differ from newly identified patients with regard to barriers that may prevent service uptake.[140, 141]

Population-level HIV serological surveys (sero-surveys) are a unique form of community-based HTC that include repeated rounds of HIV testing, in which temporary clinics are constructed in numerous locations throughout a community to test all eligible individuals in the population and refer those who test HIV-positive to register for care at a stationary HIV care and treatment centre (CTC).[142] While diagnoses during a sero-survey are made relatively closer to participants' homes, the stationary CTC may be further away, and transportation costs could become a barrier to obtaining care.[143] Some sero-survey systems have offered transportation allowances and volunteer escorts to mitigate such barriers and facilitate linkage to care.[13] In addition, repeated rounds of sero-surveys, with a unique identifier linking all previous test results, allow for the identification of individuals who are newly diagnosed with HIV. Whether individuals newly diagnosed with HIV during a sero-survey link to care at different rates than those diagnosed in facility-based VCT or PITC remains unknown.

In 2015, we introduced a system to link individuals' sero-survey records with directly observed HIV testing and care records in a community in northwest Tanzania.[72-74] In this paper, we use the linked data to compare the time from a new HIV-positive diagnosis to successful linkage to care by testing modality (community-based HTC provided during sero-surveys versus facility-based VCT or PITC provided at stationary clinics). We further explored demographic and spatial characteristics associated with linkage to care. Finally, among those in care, we compared initial CD4 cell count and ART initiation rates by testing modality.

7.4 Methods

7.4.1 Data sources

The Kisesa observational HIV cohort study was established in 1994 and is located in a rural ward in the Magu district of Mwanza region in northwest Tanzania.[68] The study includes annual or bi-annual rounds of health and demographic surveillance surveys (HDSS) that cover the entire population of approximately 35,000 residents, and multiple rounds of sero-surveys, in which adults aged 15 years or older living in the Kisesa HDSS are invited to attend a temporary village-based clinic for a personal interview and HIV test (Figure 7.1). A government-run health centre is located within the Kisesa HDSS area, including a stationary VCT clinic, an antenatal clinic (ANC) offering PITC (ANC-PITC), and a CTC. For the stationary VCT and ANC facilities, we developed electronic databases and digitised the paper-based logbooks using a double-entry system where two different fieldworkers independently captured each book, and any discrepancy between fields were reconciled in a third cleaning stage. The CTC databases have been fully digitised, and data clerks regularly update and run data checks on these data.

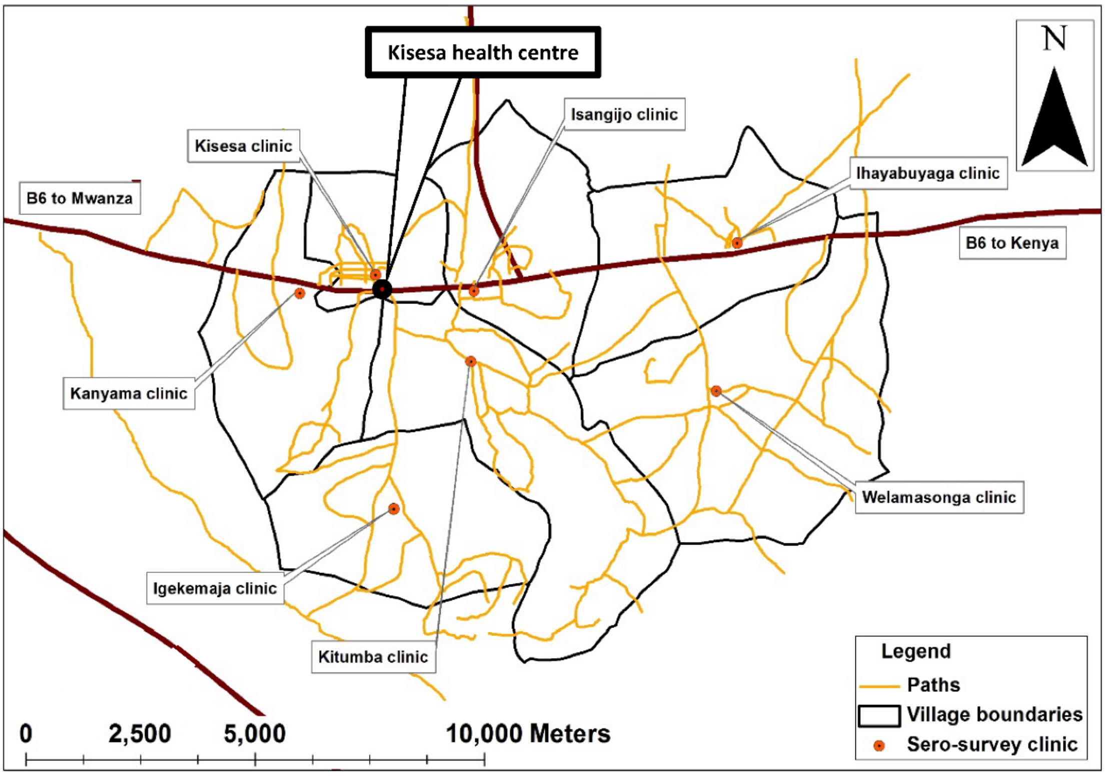


Figure 7.1: Location of sero-survey clinics and Kisesa health centre in Kisesa, Tanzania

7.4.2 Data linkage

Participants' records from all sero-survey rounds are cross-referenced with their HDSS identifiers as part of the identification process during the survey interview. Records from the three clinics were linked to the HDSS database using point-of-contact interactive record linkage (PIRL), which has previously been described in detail.[73, 74] Briefly, as individuals arrived to any of the three stationary clinics and consented to be in the study, fieldworkers entered their personal and residence details into specialised computer software,[72] which used a probabilistic linkage algorithm to search the HDSS database. While searching through potential matches, the fieldworker could view the full list of household members associated with each HDSS record as an additional step to adjudicate true matches. The fieldworker then interacted with the patient to identify which HDSS record(s), if any, were a true match.

7.4.3 Analytic sample

This analysis included all sero-survey participants, and HTC and ANC users who received a new HIV-positive diagnosis between December 2014 to October 2017. Date ranges varied by source (Table 7.1). Individuals younger than 15 years of age who received their HIV diagnosis in a clinic were excluded (to be consistent with the 15-year age limit in the sero-survey). Individuals were also excluded if their records were not linked with PIRL, had evidence of a positive HIV diagnostic test (repeat testers), or reported residence outside the HDSS area or were not seen in the 2016/17 HDSS survey (non-residents). We extracted demographic characteristics including sex, age, rurality of sub-village (rural, peri-urban, or urban), whether the sub-village of residence had a road, and geodesic distance between an individual's household and the CTC.

Table 7.1: HIV testing date range and exclusion criteria, by testing modality

	Facility-based		Community-based
	ANC-PITC	**VCT**	**Sero-survey**
Minimum HIV+ test date	30/12/2014	15/06/2015	09/09/2015
Maximum HIV+ test date	27/12/2016	03/10/2017	26/02/2016
Number of HIV+ diagnoses	24	159	476
Exclusion criteria			
Previous diagnostic HIV+ test	6 (25.0)	12 (7.6)	204 (42.8)
Non-resident	1 (4.2)	18 (11.3)	13 (2.7)*
Total in analytic sample	17	129	265

Abbreviations: HIV - human immunodeficiency virus; ANC - antenatal clinic; PITC - provider-initiated HIV testing and counselling; VCT - voluntary HIV testing and counselling; sero-survey - population-based HIV serological surveillance; HIV+ - HIV-positive

Notes: PITC offered through a stationary, antenatal clinic; VCT offered through a stationary, HIV testing and counselling clinic

*These individuals were residents during the 2015/16 sero-survey but subsequently moved out of the area

7.4.4 Outcomes

The CTC data included all registrations and visits up to November 2017 at Kisesa health centre. The primary outcome was successful linkage to care, defined as the first visit to the CTC including consultation with a clinician within 90 days of diagnosis. Limiting the time frame to 90 days provided a fairer comparison between those were diagnosed in a health facility and those who were diagnosed in the sero-survey, given the longer time period between the end of the sero-survey and available CTC registrations. Among those who linked to care, secondary outcomes were initial CD4 cell count (within one year after linking to care) and ART initiation within 90, 180, and 365 days of linkage to care.

There were other health centres in wards near the Kisesa HDSS surveillance area (all about 5-10 kilometres away from Kisesa health centre via a main road) that offered CTC services during the study period. Of note, this analysis only captured CTC registrations that occurred at Kisesa health centre.

7.4.5 Statistical analyses

We used chi-square tests to compare demographic and spatial characteristics between individuals who did and did not link to care. A Cox proportional hazards regression model was used to compare linkage to care rates by testing modality and identify associated factors. Individuals were censored at first CTC visit, death, or 90 days after positive HIV diagnosis. We considered all demographic and spatial characteristics and their interaction terms with testing modality for inclusion in an adjusted model. Interaction terms were eliminated from the model using likelihood ratio tests for significance. Remaining terms were assessed for multi-collinearity and dropped in a stepwise fashion until there was no further evidence of multi-collinearity in the model.

Among those who linked to care, we compared initial CD4 cell counts and the proportion of individuals who initiated ART within 90, 180, and 365 days by testing modality using chi-square or Fisher's Exact tests. Given the proximity between the CTC and the stationary VCT and ANC clinics (<25 metres), we performed sensitivity analysis by excluding individuals who linked to care on the same day as receiving their HIV diagnosis in either of the stationary clinics as all remaining individuals would be required to use transportation to visit the CTC on a subsequent day. We also performed sensitivity analysis on ART initiation rates by excluding individuals who had a CD4 cell count >500 cells/mm^3 at CTC registration to mirror treatment guidelines during part of the study period. Statistical analyses were performed using SAS version 9.4 (SAS Institute Inc., Cary, NC, USA).

7.4.6 Ethics

Ethical approvals were obtained from the Tanzanian National Institute for Medical Research and Lake Zone Institutional Review Board (reference no. NIMR/HQ/R.8c/Vol.II/436 and MR/53/100/450), and the London School of Hygiene and Tropical Medicine (Project ID #8852). Informed written consent was obtained from all participants.

7.5 RESULTS

7.5.1 Sample characteristics

Between 2014-2017, 659 adults received a positive HIV diagnostic test (476 community-based sero-survey, 159 facility-based VCT, 24 facility-based ANC-PITC). After excluding individuals who were previously diagnosed (222, 33.7%) and non-residents (32, 4.9%), 411 individuals remained in the analytic sample (265 community-based sero-survey, 129 facility-based VCT, 17 facility-based ANC-PITC) (Table 7.1 on page 136).

Among the 411 individuals who were newly diagnosed with HIV, 98 (23.8%) linked to care within 90 days of their diagnosis. Of note, only 8 individuals in this sample linked to care between 91-365 days of diagnosis. By testing modality, linkage to care was higher among those diagnosed using facility-based VCT (68, 52.7%) than those diagnosed using facility-based ANC-PITC (3, 17.7%) or community-based sero-survey (27, 10.2%) ($p<0.0001$) (Table 2). Individuals who resided in villages further from the CTC were less likely to link to care than those who resided in neighbouring villages (further: 8.9% Welamasonga, 12.8% Ihayabuyaga, 19.5% Isangijo; neighbouring: 16.1% Igekemaja, 22.4% Kisesa, 33.3% Kanyama, 35.9% Kitumba; $p=0.009$) (Figure 7.1 on page 135). There were no significant bivariate associations between linkage to care and sex, age, rurality of sub-village, whether the sub-village had a paved road, and distance between household and CTC (all $p>0.09$).

Table 7.2: Characteristics of individuals who received their first positive HIV diagnosis between 2015-2017 in Kisesa, Tanzania, by whether they subsequently linked to care

Characteristic	Linked to care (n=98)	Did not link to care (n=313)	p-value
Testing modality			
Facility-based VCT	68 (52.7)	61 (47.3)	<0.0001
Facility-based ANC-PITC	3 (17.7)	14 (82.3)	
Community-based sero-survey	27 (10.2)	238 (89.8)	
Sex			
Male	40 (28.8)	99 (71.2)	0.0934
Female	58 (21.3)	214 (78.7)	
Age, years			
15-29	31 (25.8)	89 (74.2)	0.9237
30-39	35 (23.8)	112 (76.2)	
40-49	18 (22.5)	62 (77.5)	
50+	14 (21.9)	50 (78.1)	
Village			
Igekemaja	12 (26.1)	34 (73.9)	0.0094
Ihayabuyaga	5 (12.8)	34 (87.2)	
Isangijo	8 (19.5)	33 (80.5)	
Kanyama	20 (33.3)	40 (66.7)	
Kisesa	26 (22.4)	90 (77.6)	
Kitumba	23 (35.9)	41 (64.1)	
Welamasonga	4 (8.9)	41 (91.1)	
Rurality of sub-village			
Urban	24 (22.0)	85 (78.0)	0.1496
Peri-urban	30 (31.3)	162 (78.6)	
Rural	44 (21.4)	66 (68.8)	
Sub-village has paved road			
Yes	44 (25.7)	127 (74.3)	0.4487
No	54 (22.5)	186 (77.5)	
Distance from household to CTC, km			
<1	19 (23.2)	63 (76.8)	0.1214
1-1.9	32 (29.9)	75 (70.1)	
2-4.9	24 (27.0)	65 (73.0)	
5-11	23 (17.3)	110 (82.7)	

Abbreviations: VCT - voluntary HIV testing and counselling; ANC - antenatal clinic; PITC - provider-initiated HIV testing and counselling; sero-survey - population-based HIV serological surveillance; CTC - HIV care and treatment centre

Note: all statistics are given in n (%); differences assessed using chi-square tests

7.5.2 Associations with linkage to care

Median time from diagnoses to linkage to care was 20 days (interquartile range [IQR] 4-47 days) among those diagnosed in the community-based sero-survey and 1 day (IQR 1-14 days) among those diagnosed using facility-based VCT. All three individuals diagnosed using facility-based ANC-PITC who linked to care did so on the same day as diagnosis. In an unadjusted model, individuals diagnosed using facility-based VCT were seven times more likely to linked to care than those diagnosed in the community-based sero-survey (hazard ratio [HR] 7.01, 95% CI 4.47-10.97) (Table 7.3). There was no statistical evidence that individuals diagnosed using facility-based ANC-PITC linked to care at higher rates than those diagnosed in the community-based sero-survey (HR 1.90, 95% CI 0.58-6.27).

Table 7.3: Associations with linkage to care among individuals receiving their first HIV+ diagnosis in a population-based HIV serological survey or health facility in Kisesa, Tanzania between 2014-2017, n=411

Covariate	**cHR (95% CI)**		**aHR (95% CI)**	
Testing modality				
Facility-based VCT	7.01 (4.47-10.97)	***	6.95 (4.39-11.00)	***
Facility-based ANC-PITC	1.90 (0.58-6.27)		2.00 (0.59-6.75)	
Community-based sero-survey	1		1	
Sex				
Male	1.40 (0.93-2.09)		1.44 (0.93-2.23)	*
Female	1		1	
Age, years				
15-29	1.19 (0.63-2.24)		0.97 (0.50-1.86)	
30-39	1.08 (0.58-2.01)		1.12 (0.59-2.11)	
40-49	1.01 (0.50-2.02)		1.10 (0.54-2.25)	
50+	1		1	
Village				
Igekemaja	3.15 (1.02-9.78)	*		
Ihayabuyaga	1.44 (0.39-5.35)			
Isangijo	2.21 (0.67-7.33)			
Kanyama	4.26 (1.46-12.47)	**	-	
Kisesa	2.59 (0.90-7.41)			
Kitumba	4.74 (1.64-13.70)	**		
Welamasonga	1			
Rurality of sub-village				
Urban	1.01 (0.61-1.66)		0.46 (0.16-1.33)	
Peri-urban	1.57 (0.99-2.50)		0.91 (0.42-1.95)	
Rural	1		1	
Sub-village has paved road				
Yes	1.12 (0.75-1.67)		1.10 (0.59-2.07)	
No	1		1	
Distance from household to CTC, km				
<1	1.36 (0.74-2.49)		2.22 (0.76-6.45)	
1-1.9	1.85 (1.08-3.16)	*	1.86 (0.80-4.34)	
2-4.9	1.62 (0.91-2.86)		1.40 (0.76-2.59)	
5-11	1		1	

Abbreviations: HIV - human immunodeficiency virus; HIV+ - HIV-positive; cHR - crude unadjusted hazard ratio; aHR - adjusted hazard ratio; CI - confidence interval; VCT - voluntary HIV testing and counselling; ANC - antenatal clinic; PITC - provider-initiated HIV testing and counselling; sero-survey - population-based HIV serological surveillance; CTC - HIV care and treatment centre; km - kilometres

*p<0.05; **p<0.01; ***p<0.0001

The final adjusted model included sex, age, rurality of sub-village (urban, peri-urban, rural), whether the sub-village had access to a paved road, and distance between household and CTC. The associations between linkage to care and testing modality remained after adjustment (HR 6.95, 95% CI 4.39-11.00 facility-based VCT; HR 2.00, 95% CI 0.59-6.75 facility-based ANC-PITC) (Figure 7.2). No other significant associations were found after adjustment.

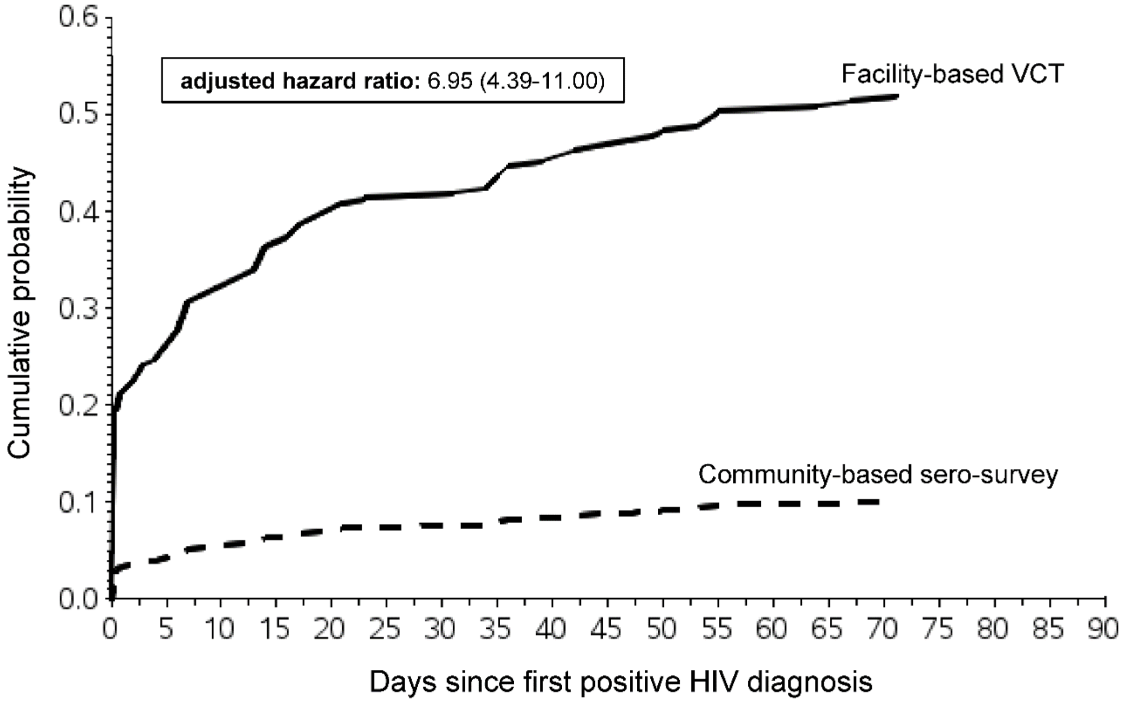


Figure 7.2: Adjusted cumulative probability of registration for HIV care after first positive HIV diagnosis by testing modality in Kisesa, Tanzania between 2014-2017, n=411

Notes: VCT – voluntary counselling and testing; facility-based provider-initiated counselling and testing had too few individuals for curve to be drawn; allowed for 90 days of follow-up, no event >71 days; adjusted for sex, age, rurality of sub-village (urban, peri-urban, rural), whether the sub-village had access to a paved road, and distance between household and Kisesa health centre

7.5.3 Initial CD4 count and ART initiation

Among individuals who linked to care, the proportion of individuals whose initial CD4 cell count was >500 cells/mm^3 was higher among those diagnosed in the community-based sero-survey (42%) than facility-based VCT (16%) (p=0.05). None of the three individuals diagnosed using facility-based ANC-PITC had a CD4 laboratory result on record.

Among the 68 individuals diagnosed using facility-based VCT and linked to care, 55 (80.9%) initiated ART within 90 days, 59 (86.8%) within 180 days, and 64 (94.1%) within 365 days. Among the 27 individuals diagnosed in the community-based sero-survey and linked to care, 17 (63.0%) initiated ART within 90 days, 21 (77.8%) within 180 days, and 23 (85.2%) within 365 days. All three individuals who were diagnosed using facility-based ANC-PITC and linked to care initiated ART on their first visit to the CTC. At each time window, there was no statistically significant difference by testing modality of the proportion of individuals who, having linked to care, initiated ART (all p>0.11). The proportion of individuals initiating ART increased further, yet conclusions remained the same, when restricting to those whose initial CD4 cell count was <500 cells/mm^3, which was the national guideline for when to initiate treatment for part of the study period.

7.5.4 Sensitivity analysis

Of the 71 individuals diagnosed using facility-based HTC (68 VCT and 3 ANC-PITC), 38 (54%) linked to care on the same day they received their HIV diagnosis (the ANC, VCT, and CTC are located within the same health centre). After excluding these individuals from the adjusted model (which included all three from facility-based ANC-PITC), the association between testing modality and linkage to care remained, although attenuated (HR 3.89, 95% CI 2.30-6.58 facility-based VCT vs. community-based sero-survey). In this restricted model, there was a clear stepped increase in the likelihood of linkage to care by proximity between household and the CTC. Compared to individuals whose households were ≥5 km away from the CTC, those who lived closer to the CTC were significantly more likely to link to care (HR 4.67, 95% CI 1.16-18.76 for <1km; HR 4.69, 95% CI 1.51-14.56 for 1-1.9km; HR 2.66, 95% CI 1.17-6.06 for 2-4.9km).

7.6 Discussion

Overall linkage to care was low (24%) among adults newly diagnosed with HIV between 2014 and 2017 in this rural Tanzanian population. However, individuals who received their first HIV diagnosis using facility-based VCT had seven-fold greater linkage to care than individuals diagnosed using community-based sero-surveys. Among individuals who linked to HIV care services, individuals diagnosed in the community-based sero-survey had proportionately higher initial CD4 cell counts >500 cells/mm^3 than those diagnosed using facility-based VCT. However, ART initiation rates were similar irrespective to HIV testing modality. These findings highlight the need for interventions to accompany community-based sero-surveys, which are important for expanding testing coverage and identifying more recent infections, to help link individuals who are diagnosed with HIV into care.

The low level of linkage to care in our sample is concerning, particularly among those diagnosed during the sero-survey, but is comparable to that documented during previous or ongoing trials in South Africa[144] and Zambia.[145] Our finding of higher uptake of HIV care services among individuals diagnosed using facility-based VCT than those diagnosed using community-based HTC is corroborated by a 2015 meta-analysis of studies reporting rates of linkage to care throughout sub-Saharan Africa. Pooling data from 31 studies, community-based HTC achieved approximately 30% linkage, facility-based PITC (overall, and not restricted to ANC-PITC) achieved 55% linkage, and facility-based VCT achieved 61% linkage,[135] compared to our findings of 10%, 18%, and 53%, respectively. The overall higher uptake of linkage to care in the meta-analysis could be due to a number of factors. First, the studies in the systematic review did not differentiate newly diagnosed individuals from those who had previously obtained a positive HIV test result. It is plausible that individuals who have received multiple positive test results may be more likely to seek HIV care services than those diagnosed for the first time. Second, the method of ascertainment of linkage to care included participant self-report, which may be affected by social desirability bias.[146] Third, the meta-analysis included studies that followed individuals up to one year to identify successful links compared to 90 days in our study. However, very few individuals in our sample linked to care between 91-365 days. Fourth, our sample for facility-based PITC was restricted to ANC users, whereas users of other facilities offering PITC, such as outpatient clinics, may include sicker individuals. Finally, uptake of HIV care services in Kisesa has consistently lagged behind other eastern and southern African communities, likely due to community-level stigma and other social and structural barriers.[10-13] Notably, the systematic review found that community-based HTC accompanied by facilitated linkage to care by trained lay counsellors or health workers achieved 95% linkage within 12 months.[135] Therefore, future sero-surveys should explore including facilitated linkage to care among those diagnosed with HIV as a way to improve linkage to care.

We found that individuals who resided in villages nearer the CTC were more likely to link to care than those in villages further away, except for those living in Kisesa village. Given Kisesa village's proximity with the CTC, we hypothesise that individuals who lived in the immediate area surrounding the clinics may have been more likely to travel to obtain HIV care outside of the surveillance area following a diagnosis made within Kisesa, which would have resulted in the attenuated effect. There is some evidence for this in a previous study of ours that showed nearly half of all clinic attendees in the CTC between 2015 and 2017 were non-residents,[74] which underscores the importance for future trials and observational studies to include tracing of diagnosed individuals to capture care received outside the immediate area, where possible.

Previous studies have highlighted the success of community-based HTC to identify asymptomatic HIV-positive individuals at relatively higher CD4 counts compared to facility-based HTC.[135, 139, 147] Our findings are consistent with these previous studies in that nearly half of individuals diagnosed in the community-based sero-survey had CD4 counts >500 cells/mm^3 at care initiation compared to only 16% of those diagnosed using facility-based VCT. Of note, 74% of all newly diagnosed individuals in Kisesa who linked to care between 2014-2017 had CD4 <500 cells/mm^3 when they initiated care. National treatment guidelines in Tanzania were to initiate ART in individuals with CD4 <500 cells/mm^3 in 2015,[148] which expanded to all diagnosed individuals in 2017[71] to match current WHO guidelines.[149] Therefore, most individuals in our sample were eligible for treatment when they linked to care, which corresponds with the high level of ART initiation we observed.

We conducted a sensitivity analysis on our regression model to exclude individuals who were diagnosed and linked to care in the same day. This exclusion could be seen as providing a fairer comparison to individuals who obtained community-based HTC in that those who were diagnosed in a stationary clinic had to return to Kisesa heath centre on a subsequent day. In this restricted model, we still found a strong, albeit attenuated, association between testing modality and linkage to care. We also found the distance between an individual’s household and the CTC played a role in the likelihood of successfully linking to care when restricted to those who would have needed to travel back to the CTC on a subsequent day to achieve linkage to care. These findings further underscore the importance of providing equal access to HIV care and treatment services irrespective to the distance from the nearest stationary CTC, including transportation refunds. By the end of 2017, three village-based health posts located within the Kisesa HDSS surveillance area were offering ART directly to attendees in addition to the CTC at Kisesa health centre. We are currently assessing how best to link clinic records from these less-frequented health posts into the linked data infrastructure.

Our study had limitations. First, our analysis did not capture linkages to care that occurred outside the study area so our estimate of the proportion who linked to care is likely to be underestimated. If the decision to obtain care outside of Kisesa ward was related to modality of HIV testing or any spatial characteristic (as may have been the case for Kisesa village), the results in this paper may be subject to bias. We used all available data to control for such bias, including limiting the analytic sample to those who were resident in the study area as of 2017. Second, we lacked sufficient power because of the relatively small sample size, particularly among those diagnosed using facility-based ANC-PITC, which resulted in large standard errors of regression estimates. A larger number of repeat testers in the facility-based sample may have also allowed for a separate analysis among these individuals to identify their likelihood of linkage to care.

7.7 CONCLUSIONS

We measured linkage to care following a population-level HIV serological surveillance round, which is a form of community-based HTC not previously included in systematic reviews on the topic. This analysis was made possible and strengthened by the novel linked data infrastructure available in the Kisesa observational HIV cohort study, which includes directly observed data for HIV testing (both community- and facility-based), diagnoses, care, and treatment. We found that while overall linkage to care was low among newly diagnosed adults in this rural Tanzanian community, those diagnosed using facility-based VCT had higher uptake of HIV care services than those diagnosed using facility-based ANC-PITC or in the community-based sero-survey. However, once individuals were in care, there was no evidence of any further delays to ART initiation by testing modality. Community-based HTC is important for earlier diagnosis of HIV-positive individuals; however, these efforts should include interventions to link individuals newly diagnosed with HIV into care and provide stationary care and treatment services nearby all locations offering HTC.

8 Discussion

8.1 Overview

In the previous chapters, I described in detail i) how PIRL was implemented in Kisesa, ii) compared PIRL to automated linkage, and iii) used the emerging data infrastructure to measure directly-observed patterns of HIV service utilisation. This chapter consolidates the key findings from each paper, although detailed conclusions previously presented will not be repeated. Recommendations for programmes, policy, and future research will be provided along with a discussion of the strengths and limitations of the analyses conducted in this thesis. Finally, I will describe various efforts to disseminate my findings and provide concluding remarks.

8.2 Synthesis of findings

The primary aim of this PhD research was to augment existing computer software towards a novel approach to record linkage in a rural community in northwest Tanzania and use the emerging data infrastructure to measure patterns of HIV service utilisation.

This aim led to the following objectives:

1. To implement a locally-relevant approach to link community cohort data with medical records from three separate health facilities offering HIV services (Chapters 2 and 3)

2. To identify individual characteristics associated with successful linkage using PIRL and compare PIRL with automated probabilistic record linkage (Chapters 4 and 5)

3. To measure patterns of HIV service utilisation using the linked data infrastructure created by PIRL (Chapters 6 and 7)

Table 8.1 details the key findings from each paper presented in this thesis. In this section, I will discuss these findings in relation to the above objectives.

Table 8.1: Key findings from papers presented in this thesis

Paper	Objective	Paper title	Journal	Key findings
A (Chapter 3)	1	Point-of-contact Interactive Record Linkage (PIRL): A software tool to prospectively link demographic surveillance and health facility data	Published in *Gates Open Research*, 2018	A key advantage of this software is the ability to perform unlimited searches in the presence of the individual whose records are being linked; each search attempt took <15 seconds; excluding time spent obtaining consent. Median duration of time spent with each patient was six minutes.
B (Chapter 4)	2	Point-of-contact interactive record linkage (PIRL) between demographic surveillance and health facility data in rural Tanzania	Published in *International Journal for Population Data Science*, 2017	Matched 84% of individuals who reported residence history; qualitative (not quantitative) amendments to identifiers during repeated searches key driver of linkage; other important associations with successful linkage: experience of fieldworkers, increased age of participant, and residence in areas without access to a paved road; no association found with HIV test result (likely due to linking prior to test); automated linkage resulted in substantial linkage errors (up to 55% false-match rate).
C (Chapter 5)	2	Impact of linkage quality on inferences drawn from analyses using data with high rates of linkage errors in rural Tanzania	Under review at *BMC Medical Research Methodology*, 2018	Bias and precision in analyses using linked data are impacted by substantial linkage errors similarly to how they are impacted by more negligible linkage errors; impact of linkage errors on bias and precision found at all match score thresholds; selection bias is likely to have impacted the analyses given that selection into the linked datasets was related to both exposure and outcome, and as found, conditioning the analyses to records that were linked induced a protective relationship between the exposure and outcome.
D (Chapter 6)	3	Non-disclosure of HIV testing history in population-based surveys: implications for the estimation of the UNAIDS 90-90-90 target	Under review at *AIDS*, 2018	Using UNAIDS guidelines for measuring the 'first 90', the proportion of PLHIV in Kisesa who were diagnosed was estimated to have doubled from 34% in 2010 to 65% in 2016; Evidence of bias resulting from non-disclosure of HIV testing history: between 20-33% did not disclose testing history; more non-disclosures correlated with more bias; current UNAIDS guidelines under-estimates 'first 90' by at least 10-20%.
E (Chapter 7)	3	Linkage to care and antiretroviral therapy initiation by testing modality among individuals newly diagnosed with HIV in Tanzania, 2014-2017	Under review at *Tropical Medicine & International Health*, 2018	Linkage to care was 24% among adults newly diagnosed with HIV between 2014 and 2017; individuals who received their first HIV diagnosis using facility-based VCT had seven-fold greater linkage to care than individuals diagnosed using community-based sero-surveys (53% vs. 10%); distance to the CTC was associated with linkage to care; among those who linked to care, ART initiation rates were similar irrespective to HIV testing modality.

8.2.1 Objective 1: To implement a locally-relevant approach to link community cohort data with medical records from three separate health facilities offering HIV services

Adoption of record linkage activities has remained low in Tanzania, and more broadly in sub-Saharan Africa for a variety of reasons. First, there is an overall lack of electronic data available for linkage. Second, even where electronic data are available, linkage is inhibited by the absence of corresponding fields in all data sources and the relatively poor quality of variables that could be used by a linkage algorithm. Third, traditional, automated linkage approaches, which require minimal overhead costs, are complicated in settings without a dataset of high quality, gold standard links to guide the linkage process.

The field methods and execution of PIRL in Kisesa was specifically tailored to overcome all three common barriers to linkage activities in sub-Saharan Africa, as listed above and further described below.

First, the TAZAMA Project had a wealth of demographic and HIV serological surveillance data collected over two decades already in electronic format. In addition, the primary source for HIV care and treatment within the surveillance area had electronic medical records dating back to 2008. I obtained local approval to digitise the paper logbooks used in the remaining two clinics included in this PhD research.

Second, PIRL was not subject to requirements of common identifiers found in all data sources. The PIRL software was designed to collect and use identifiers only found in the search database, which in this case was the HDSS database. During the PIRL interview, only participants' clinic identifiers and not any other medical details were collected to allow for subsequent linkage between the community cohort data and electronic medical records in each facility. Also embedded within the PIRL software were numerous data integrity checks on all collected identifiers (i.e., clinic, personal, and residential) to ensure adequate data quality during the linkage process.

Third, Kisesa offered a suitable location to invest the resources necessary to attempt PIRL for the purposes of monitoring HIV service uptake in a population due to the limited number of health facilities offering such services within the surveillance area. The primary sources of HIV diagnoses among Kisesa residents were captured in this PhD research (i.e., HTC, ANC, sero-survey). Similarly, nearly all ART offered in the surveillance area came from the CTC in Kisesa health centre, although a few of the smaller village-based health posts began offering ART towards the end of the study period.

The PIRL software included a variety of features to make it appropriate for use in Kisesa. Due to concerns of data quality and the dynamic nature of the identifiers used in the linkage algorithm, the PIRL software allowed for a limitless number of searches for a participant's HDSS records. This feature was found to be a key driver of successful linkage (Chapter 4). Since most PIRL interviews were conducted while a clinic attendee was awaiting their visit with a CTC clinician, HTC counsellor, or ANC nurse, I ensured the amount of time we spent with each individual was minimal. For example, I reduced the time the PIRL software took to perform a single search to less than 15 seconds. I also trained the field team to use the same process for each interaction with a participant as a way to streamline the PIRL process. Excluding time spent obtaining consent, the median duration of time spent with each patient was six minutes during an initial visit. Since the PIRL software retrieves information collected and links made during all previous sessions, most repeat visits were completed in less than one minute.

The multilingual skills of the field team and translations of information given in the PIRL software allowed for smooth operations in Kisesa. The fieldworkers were able to communicate in both local languages spoken in Kisesa, KiSwahili and KiSukuma. All were fluent in English. I worked with a local experienced researcher at NIMR to translate instructions and messages that appeared during normal use of the PIRL software to KiSwahili.

It is important to also attribute the acceptability of PIRL in Kisesa principally to the long-standing TAZAMA Project, which is well recognised and respected in the community. As part of a fieldworker's introduction to a given clinic attendee, they conveyed this study's association with TAZAMA. During my experience sitting in on many PIRL sessions over the first several months of operations, I witnessed the positive familiarity that patients had with TAZAMA. Likely due to the experience Kisesa residents have with the repeated HDSS rounds and sero-surveys, there were no active refusals to participate in this PhD research among those who agreed to sit down with a fieldworker. It should be noted, however, that it was likely some clinic attendees passively refused to take part in this research by not agreeing to meet with a fieldworker. During high-volume clinic days, the number of clinic attendees exceeded the number of individuals we could enrol in the PIRL study, and patients who were not willing to participate may have self-selected to not be in the waiting room or take a number card. I conducted an analysis of coverage of PIRL in each of the three clinics by calculating the proportion of patients listed in a logbook who consented to the PIRL study. In the CTC, 78% of all patients who visited during PIRL data collection had consented to PIRL. There was higher coverage in the HTC (89%) and ANC (91%), potentially due to the lower average daily number of clinic attendees in these clinics. Nonetheless, after two years of PIRL data collection, there was high coverage and therefore potentially representativeness among individuals who attended each of the three clinics.

Direct comparisons of PIRL to previous record linkage approaches conducted in Kisesa are made difficult by the differing methodologies and available published data. Linkage rates were higher when using PIRL (84%) than previous approaches without patient interaction (37% in the HTC[11] and 75% in the ANC[39]). However, the previous approaches used automated approaches with manual review, which may potentially be used to complement PIRL as a way to increase coverage and representativeness of linked records of individuals, particularly among those who do not attend any of the three PIRL clinics (discussed further in Section 8.5.4).

The source code for the PIRL software along with a full user guide detailing the implementation in Kisesa has been published online in an open-source repository.[72] The software was published with an MIT license, which allows others to download, edit, and use the software in any way to make it appropriate in their setting. Objective 1 identified a locally-relevant approach to perform record linkage in Kisesa. Understanding factors associated with successful linkage using PIRL and benchmarking PIRL against a less resource-intensive approach to record linkage was the remit of Objective 2.

8.2.2 Objective 2: To identify individual characteristics associated with successful linkage using PIRL and compare PIRL with automated probabilistic record linkage

After two years of PIRL in Kisesa, 84% of individuals who reported residence history in the surveillance area, and therefore were likely to have an HDSS record, were successfully matched. This match percentage did not significantly differ between the three clinics. Due to the novelty of the PIRL approach, it is difficult to benchmark this match percentage. However, there are some previous linkage studies that share some similarities to this PhD research. Compared to automated linkage conducted in Kisesa, the match percentage attained in this PhD outperformed both previous attempts, which had match percentages of 75% and 37%.[11, 39] One linkage study at another ALPHA Network site in Agincourt, South Africa used a combined approach of probabilistic and deterministic methods (South Africa has a national identification number system) to achieve a match percentage of 88%.[38] Given these comparisons, PIRL has been demonstrated as a successful alternative to automated methods for the purposes of linking community and clinic data in Kisesa.

Several factors were associated with successful linkage using PIRL. As discussed in Chapter 4, increased age of the participant and residence in areas without access to a paved road were associated with being matched, suggesting that those with longer and more stable residency history were more likely to be matched. Importantly, factors related to field operations were among the strongest associations with successful linkage. Fieldworkers with more previous experience with HDSS data had higher match percentages than those with more limited experience. In addition, compared to the first three months of PIRL operations in Kisesa, match percentages were significantly higher in all other time periods during the study period. These findings highlight the importance of repeated trainings of the field team to increase the understanding of the HDSS database being searched and the PIRL process. During the entire study period, the field team held daily debriefings in which experiences from the day, particularly unusual or difficult sessions, were shared among the group.

As mentioned in Chapter 3, a key advantage of the PIRL approach over automated linkage was the ability to perform multiple search attempts on the same individual until all relevant HDSS records were found. The analysis conducted in Chapter 4 underscored this point by providing evidence that qualitative amendments to identifiers collected during the PIRL interview was a key driver of successful linkage. These amendments often included updating spelling of names, dates of birth, and residence details. The interactive nature of PIRL was a vital feature for record linkage in this setting.

PIRL was shown to be superior to automated linkage in Kisesa in two ways. First, an automated approach using the same linkage algorithm embedded in the PIRL software would have only correctly identified half of the links made using PIRL. These results from Chapter 4 led to the hypothesis that analyses using data from automated linkage would be impacted by the substantial linkage errors found in this setting. Chapter 5 provided original evidence that bias and precision in analyses using linked data were impacted by substantial linkage errors similarly to how they were impacted by more negligible linkage errors. Selection bias was likely to have impacted the analyses given that selection into the linked datasets was related to both exposure and outcome, and as found, restricting the analyses to records that were linked under-estimated the strength of the association between the exposure and outcome. Importantly, there was no match score threshold that removed bias from the analyses using automated linked data. Until future analyses investigate methods to adjust for these biases and provide more robust results using data with considerable linkage errors, our results suggest that other researchers in similar settings who wish to perform probabilistic record linkage should allocate resources toward PIRL or similar system.

8.2.3 Objective 3: To measure patterns of HIV service utilisation using the linked data infrastructure created by PIRL

Substantive analyses investigating how PLHIV progress through HIV services from diagnosis to care were important to demonstrate the value of the linked data infrastructure created by PIRL. Estimates of the proportion of PLHIV in Kisesa who were diagnosed were shown to have doubled from 34% in 2010 to 65% in 2016 (Chapter 6). Among adults newly diagnosed with HIV between 2014 and 2017, only 24% successfully linked to care within 90 days after diagnosis (Chapter 7). Among those who had linked to care, however, approximately 90% initiated ART (the 'second 90') within 365 days of their first HIV positive test, with no statistically significant differences by modality of diagnosis.

The PIRL data infrastructure further allowed for the quantification of bias in the measurement of the 'first 90.' Current UNAIDS guidelines on how to estimate the 'first 90' from population-based surveys were found to be influenced by individuals who did not disclose their HIV testing history, which is a principal component in the estimator. Up to one-third of sero-survey participants did not accurately report their HIV testing history, whether intentionally or unintentionally. Thus, estimates were found to be biased by a relative factor of up to 20% after controlling for directly-observed HIV testing history captured in the PIRL data infrastructure. As detailed in Chapter 6, this may still likely be an under-estimate. The strongest association with non-disclosure of HIV testing history was refusing to answer other sensitive questions (e.g., number of sex partners in last 12 months or condom use at last sex). UNAIDS guidelines could potentially incorporate this finding into updated estimates of the 'first 90' using population-based surveys, particularly those without linked and directly-observed HIV testing history data.

The linked data created by PIRL also allowed for a comparison of linkage to care rates by various testing modalities. Before PIRL, a paper-based tracking system was developed to link patients testing HIV-positive in the HTC to the CTC;[13] however, this method fell into disuse. Linkage to care among sero-survey participants who tested HIV-positive was based on self-reported HIV service use. PIRL offers a more robust form of record linkage between the locations where most diagnoses occur within the surveillance area (i.e., HTC, ANC, and sero-survey) and the local CTC.

While the ‘third 90’ cannot currently be measured in this setting due to the lack of HIV viral load testing, if, or when, such tests become routine care in Kisesa, they will be automatically available in the linked data infrastructure created by PIRL. An alternative criterion to assess treatment success could be to investigate retention in care at various time points after ART initiation.

8.3 Programme and policy recommendations

Findings from this PhD research have the potential to inform improvements to the design of local health systems for better monitoring and utilisation of HIV services. Drawing from the key findings presented above, this section presents programme and policy recommendations in three broad areas: i) continued enhancements of the linked data infrastructure, ii) improving access to HIV care services in Kisesa and more generally throughout Tanzania, and iii) funding stability.

8.3.1 Continued enhancements of the linked data infrastructure

Continuing to add data from facilities offering medical services, even non-HIV-related services, is important to monitor service uptake and retention in the entire community for myriad health conditions without solely relying on self-reported data. Chapter 6 demonstrated the bias that was present in a commonly used indicator for monitoring HIV service uptake globally when relying only on self-reported survey data. The data infrastructure used in the TAZAMA Project and also that which was created during this PhD research allowed for the augmentation of estimates based on self-reported HIV testing history in two important ways. First, all HIV sero-survey rounds conducted by the TAZAMA Project have used a unique identification system that enables the linkage of participants’ records across multiple rounds. This feature allowed for the verification of a participant’s self-reported HIV testing history during a sero-survey round. Second, PIRL supplements this existing data infrastructure with directly-observed HIV testing history data from two more sources of HIV testing in Kisesa (the HTC and ANC). These additions led to the first evidence of bias in the 'first 90' estimate at a sub-national level in Tanzania and one of the first in sub-Saharan Africa (Paper D).

I travelled to two health posts within the surveillance area to assess the feasibility of introducing PIRL in these facilities. There are several conditions that need to be considered prior to expanding PIRL into the village-based health posts. First, the number of patients who receive HIV care in these clinics is relatively low. In the health posts that I visited, there was less than one HIV test given per day, on average. However, the number of people who seek HIV services may increase since ART is now indicated for all HIV-positive individuals throughout Tanzania.[71] To make record linkage activities more cost-beneficial in these smaller health posts, all individuals who attend these clinics should be approached irrespective of services sought. This would also preclude the potential for ethical issues relating to only identifying and approaching individuals seeking HIV services in these more rural settings.

The recommendation to add medical records from more health facilities using PIRL in Kisesa, however, would implicate the need for a linkage fieldworker to be stationed at every health facility that participates in the proposed linkage activities, which may not be cost-effective. An alternative solution for incorporating the records from these clinics is to digitise their paper logbooks at regular intervals and link these records retrospectively. There are several considerations that need to be made for the proposal to be viable. First, not all variables currently included in the linkage algorithm (or the limited algorithm identified in Chapter 4) are collected in the paper logbooks of the health posts, particularly the names of an additional household member. All data fields used by the linkage algorithm in the PIRL software should be introduced into the registers as soon as possible to permit future linkage activities. Second, enhancements to PIRL activities in Kisesa would require the same level of commitment from clinic staff as given in the clinics included in this PhD research. The clinic staff I met during my visits to two of the smaller health posts were receptive to allowing TAZAMA to routinely digitise their logbooks. Clinic staff in the remaining health posts should be visited to determine their likelihood of acceptance to record linkage being conducted in these sites. Third, a highlight of PIRL is the ability to obtain informed consent from participants. Ethical issues surrounding collecting and linking medical records with community data without obtaining informed consent would need to be resolved. One potential solution would be to obtain consent during an upcoming HDSS round.

As presented in Chapter 4, about one-third of clinic attendees approached in Kisesa health centre reported residence outside of Kisesa ward. This finding indicates that some Kisesa residents may obtain HIV testing or care outside of the surveillance area. A first step to assess the proportion of Kisesa residents who obtain medical services outside of the surveillance area would be to partner with nearby health centres, five of which are situated within 20 km of Kisesa health centre. In the absence of a nationwide identification system and electronic medical records, which are likely not to be implemented in the short-term, expanding linkage activities to adjacent facilities would provide more complete histories of HIV (or non-HIV) services obtained by an individual who resides in the surveillance area.

The distinct identification systems used in each clinic involved in this PhD research made linking medical records across clinics nearly impossible without the use of PIRL. A possible solution is to construct a central registration office in which all Kisesa health centre attendees register their arrival and obtain a unique Kisesa health ID number that can be recorded in medical records irrespective of the clinic (or clinics) where they receive care. During fieldwork, I met with the District Medical Officer in Magu to discuss plans to construct a central registration office in Kisesa health centre to act as both an entry point to receive care and a single location to perform PIRL for the entire health centre. The District Medical Officer was receptive to this idea – he was familiar with similar systems that are used in many private hospitals in the area – and committed to get it approved by the relevant decision-makers. However, there are many conditions, such as time, resources, and costs of the development, training, staffing, and integration of the new office that need to be considered before pursuing a central registration office.

Beyond Kisesa, a health identification numbering system could be beneficial if implemented at the national level similar to the "health passports" issued in neighbouring Malawi. In addition to records kept at clinics, these health passports document immunisations, public health interactions, diagnoses, and visits for each person. These health passports are also popular; over 90% of the Malawian population possess a health passport.[150] A national health identification system could potentially leverage the numbering system used in all CTC nationally. Benefits arising from such a system would be plentiful, including linked medical histories, enhanced continuation of care, and the ability to track service utilisation.

Another complication that arose from individuals having a different HDSS identifier for each residency episode they had. The identification system used by the TAZAMA Project could be substantially improved by introducing full-scale migration reconciliation. As detailed in Chapter 4 and Appendix 10.5, approximately one out of four participants in this PhD research had multiple HDSS identifiers. I created a unique identifier that strings together each participant's set of HDSS identifiers in the PIRL database and have sent this table to LSHTM and NIMR researchers to attempt migration reconciliation on the back-end databases. Alternatively, it would be feasible to leverage the field methods from this PhD research described in Section 2.2 in an upcoming HDSS round whereby the PIRL software can be used to identify all HDSS records for every resident based on their residence history. In this event, it may be important to devise an updated identification system that would chain together all HDSS identifiers to an individual, and importantly, be given to new residents going forward.

8.3.2 Improving access to HIV care services

The low level of linkage to care identified in Chapter 7 is concerning, particularly among those newly diagnosed during the most recent sero-survey in Kisesa. This finding highlights the need for interventions to accompany sero-surveys, which are important for expanding testing coverage and identifying more recent infections, to help link individuals who are diagnosed with HIV into care. The ongoing HIV Prevention Trials Network (HPTN) 071 (Population Effects of Antiretroviral Therapy to Reduce HIV Transmission [PopART]) trial is evaluating the impact of a combination HIV prevention package, including repeated home-based testing with facilitated linkage to care and ART adherence offered by community HIV-care providers, on HIV incidence at the population level in Zambia and South Africa.[151] However, the rates of linkage to care were similarly low over the first year of the trial as they were in Kisesa.[145] While uptake of HIV testing with the community HIV-care providers in the PopART trial was high (as has been found elsewhere[152]), only 41% of individuals who were diagnosed with HIV and reported they had never previously registered for HIV care had linked to care within 90 days of referral.[145] The PopART investigators provided updated estimates at the 2018 International AIDS Conference, in which they reported that second 90 targets were reached overall among women and almost reached among men.[153] Moreover, a systematic review found that community-based HTC with facilitated linkage resulted in a linkage to care rate of 95%,[135] indicating the potential benefits arising from the WHO recommendations. The TAZAMA Project has previously offered transportation allowances and volunteer escorts to mitigate any barriers to access HIV care services;[13] these efforts and others should be explored again.

The distance between an individual's residence and the CTC was found to be associated with linkage to care in Kisesa, as detailed in Chapter 7. Offering HIV care services, including ART, at more points of care throughout Kisesa could potentially improve access to these services, as decentralisation in other settings has shown.[154, 155] By 2017, near the end of this PhD research, Tanzanian government policy stipulated that every health facility in Tanzania should offer HIV care and treatment services. At the time of PhD submission, three health posts within Kisesa (Welamasonga, Igekamaja, and Ihayabuyaga) had started to offer ART, although with no computers available in these smaller health posts, records for HIV care and treatment are kept in paper logbooks. Continued expansion of CTC services is warranted, and records should be captured in future expansion of linkage activities, as detailed in Section 8.4.

8.3.3 Funding stability

As mentioned in Section 1.8, funding was depleted and linkage activities were halted on 31 May 2017, exactly two years after PIRL activities commenced. The original intention was for linkage activities beyond two years to be included in the larger funding package obtained for repeated HDSS rounds; however, such support was not obtained for the remainder of this PhD research. The LSHTM and NIMR teams are currently exploring options to allow PIRL to continue alongside the TAZAMA Project. Gaps between funding periods and linkage activities have the potential to induce missing data issues in the linked data infrastructure, particularly by not capturing paper logbooks completed by the clinic staff during the break and should be avoided when possible.

8.4 Recommendations for future research

The findings presented in this thesis answered some questions but raised others that warrant further investigation. In no particular order, recommendations for future research include:

- **Test PIRL in other settings.** While effort was made in this PhD research to tailor PIRL specifically to Kisesa, the fundamentals of the PIRL approach overcome three barriers that are common to other settings, including i) lack of electronically available data, ii) poor data quality, and iii) dynamic personal identifiers. This thesis found PIRL to be successful in Kisesa, which in itself was an augmentation of a system developed in Agincourt, South Africa; however, the utility of PIRL in other settings that lack national personal identifiers remains unknown. The positive familiarity of the TAZAMA Project among Kisesa residents, which I previously cited as a principal reason for high participation rates found in this research, may not translate to other settings. However, the PIRL approach was specifically developed to be minimally intrusive (e.g., interviews conducted in a private area and no health information collected in the PIRL software) and allows the patient to remain in control of their participation (e.g., informed consent/assent). Thus, I hypothesise that PIRL may be an effective record linkage solution for smaller-scale research projects where data quality is a principal concern, even in settings without long-standing and renowned demographic surveillance.

- **Assess the effectiveness of PIRL without participant-interaction.** As detailed in Section 8.3.1, many of the smaller health posts that offer medical services in the surveillance area do not have large enough patient populations to make PIRL, as implemented in this PhD research, a cost-effective tool. One potential solution is to ensure the linkage variables from the limited algorithm are captured in the paper logbooks kept at these facilities and digitise these records regularly. However, the linkage quality of this approach needs to be formally assessed.

- **Update the probabilistic linkage algorithm.** Paper B found that an algorithm limited to a smaller set of identifiers performed similarly to the full algorithm. However, both the full and limited algorithms performed poorly using automated linkage, which impacted secondary analyses as evidenced in Paper C. Other identifiers should be tested for inclusion into the linkage algorithm used in Kisesa as a way to potentially improve the algorithm's performance when applied retrospectively. Suggestions for identifiers include mobile telephone numbers and names of parents. Any additional identifiers to be tested in the algorithm would also need to be collected during an upcoming HDSS round. The findings in Paper B also suggest that the PIRL software could potentially remove the identifiers not included in the limited algorithm. In addition, the size of the HDSS database and computing power of the machines used in this PhD research were sufficient to assess each record-pair for the likelihood of being a match. There is potentially a need to incorporate efficiencies into the algorithm if used on larger-scale data or less powerful machines (i.e., tablets), such as blocking on sex, which would reduce computing time by roughly half given the even distribution of males and females in Kisesa.

- **Continued research on the impact of substantial linkage errors on secondary analyses.** The analysis presented in Chapter 5 assessed the impact of linkage errors in a Cox regression analysis in which the exposure was highly associated with the outcome and successful linkage was associated with both the exposure and outcome. There was some evidence that a study intending only to assess the prevalence of the outcome used in Paper C would have also been severely impacted by linkage errors. Further research is necessary to assess other exposure, outcomes, and types of analyses using data with substantial linkage errors. Importantly, future analyses should investigate methods to adjust for these biases and provide more robust results using data with considerable levels of linkage errors.

- **Quantify bias in the 'first 90' by testing modality.** Paper D found that non-disclosure of HIV testing history resulted in up to 20% relative bias in the estimate of the 'first 90' among sero-survey participants. However, a 2018 publication found that people may not disclose their HIV testing history differentially by testing modality.[156] Future investigations are warranted to measure the potentially diverse levels of bias in the UNAIDS indicator across multiple HIV testing options.

- **Comparative research on bias in the 'first 90' across settings.** Comparative research should be employed in other HDSS sites that benefit from linked HIV testing and clinical data to investigate how the corrective factor may vary between settings.

- **Continued monitoring of linkage to care among newly diagnosed individuals.** Paper E found that overall linkage to care was low among newly diagnosed individuals since 2014, which corresponded with findings from recent and ongoing clinical trials investigating ways to improve linkage to care.[144, 145] Future quantitative and qualitative research should identify locally-appropriate interventions to improve linkage to care with the ultimate goal of reducing HIV-related incidence, morbidity, and mortality in the population.

- **Estimate the other UNAIDS 90-90-90 targets in Kisesa.** Paper D estimated the 'first 90' in Kisesa over a six-year period. Continued work should estimate the 'second 90' and 'third 90'. The former is estimated using programme data and not self-reported ART status; thus, expanding linkage activities to capture CTC records in the health posts will be necessary to provide an accurate estimate of the 'second 90'. Longitudinal sub-national estimates of the 'second 90' are also necessary to provide information towards bias in the 'first 90', as described in Chapter 6. A wide range of sensitivity analyses were performed to gauge the robustness of the bias estimates; however, longitudinal estimates specific to Kisesa would improve the precision and therefore confidence in these bias estimates. The 'third 90' will require CTC services to provide viral load measurements, which may not happen in the short-term. However, CTC records that are linked before these data become available will automatically include viral load tests once they are available.

- **Assess the acceptability of CTC services provided in health posts.** As previously mentioned, Tanzanian government policy shifted toward the end of this PhD research to include the offer of HIV care and treatment services in all government-run health facilities. At least three smaller health posts in Kisesa have begun offering such services. Interviews with PLHIV and health workers should be conducted to understand the acceptability of receiving HIV care closer to people's homes. In addition, a feasibility assessment for expanding PIRL to these facilities should include regular monitoring of the number of patients who receive HIV care and treatment from these clinics.

8.5 STRENGTHS AND LIMITATIONS

In this section, I will discuss the strengths and limitations of PIRL separately from those that affect the overall conclusions of this thesis. Specific limitations to each research paper can be found at the end those chapters. Attempts have been made to minimise overlap.

8.5.1 Strengths of PIRL

The implementation of PIRL in Kisesa described in this thesis is a unique approach to record linkage that incorporates multiple methods currently used in this area of research. The novelty of prospectively linking records by incorporating brief interactions with those whose records are being linked provides an important contribution to the ever-growing field of record linkage, most notably in an African context. Record linkage methods are rarely used in sub-Saharan Africa most likely due to poor data quality and the general lack of electronic data. The embedded features in PIRL begin to overcome those challenges. A principal feature of PIRL over other traditional methods is the ability to perform multiple searches for the same individual by modifying information originally collected on the individual. Other sub-Saharan African settings where record linkage is being considered should consider the PIRL approach.

A key feature of the implementation of PIRL in the clinics was that fieldworkers did not seek out or pressure any clinic attendee to participate in PIRL. The rationale for this approach was that many clinic attendees were likely to have repeated visits to the clinic, and they may feel more inclined to participate after several visits and increased familiarity with the PIRL fieldworker. After two years of PIRL data collection, coverage in each of the clinics was 78% in the CTC, 89% in the HTC, and 91% in the ANC.

Several features of the data collection process used in this PhD research ensured quality and integrity in the data. Other than names, all identifiers collected in the PIRL software required one or more of the following: double-entry, check digits, specific format, or drop-down selection box (rather than free text fields). The data entry systems used to capture clinic logbooks also ensured data quality and integrity using the same methods as the PIRL software. In addition, all clinic logbooks were captured twice by two independent fieldworkers, and any discrepancies were resolved in a third round (or more, if the third entry did not match either the first two entries). Although all fieldworkers were fluent in English, multilingual message boxes appeared throughout the PIRL software, including warnings when an attempted matched record had a significant inconsistency with the collected information (e.g., difference in birth year >10 years). While the fieldworker could simply override such messages, all matches made in the field were validated through monthly inspections of the back-end data. Only eight (0.2%) of the 3,456 matches made during this PhD research were deemed unlikely and were deleted from the back-end database.

In the field of computer science, software is often evaluated in terms of its computational complexity. Automated record linkage approaches often have issues of scalability, particularly as the size of the two input datasets increase, and therefore use approaches to limit the number of pairwise comparisons, such as blocking on sex or race. For example, if database A had 1,000 records (500 male and 500 female) and database B had 1,000 records (400 male and 600 female), an all pairwise comparisons approach would require computing a match score for 1,000,000 record-pairs (1,000*1,000). Whereas if sex was used for blocking, a match score would only be computed for 200,000 record-pairs (500*400) for males and 300,000 record-pairs (500*600) for females. The PIRL software does not suffer from this issue as only one record of matching variables are passed through the algorithm at a time. Therefore, a match score is computed on *n* number of records found in the back-end database. In Kisesa, the search database included approximately 100,000 records; thus, only 100,000 match scores are computed for each search, which takes 10-15 seconds per search. If PIRL continues in Kisesa or is used in a different setting, future users may need to introduce techniques, such as blocking, if they wish to decrease time per search or if the size of the back-end database increases.

Another strength of this research was the rapport within the field team. Through daily debriefings, the team, including myself, quickly began recognising areas of improvement in the field methodology. One example included updating our initial question, “Do you live in Kisesa?” to, “Where do you live?” as some individuals misclassified themselves as non-residents. These daily meetings also increased the accountability of each fieldworker to the data they collected. On several occasions, I was asked to check on specific matches they had made on a given day to verify an assumption they made during the linkage process. These conversations continued even when I was not in the field. I held weekly Skype calls with the entire team for the first year of PIRL activities before switching to monthly calls for the second year of data collection.

8.5.2 Limitations of PIRL

Data missingness is a key limitation of the PIRL approach, and it is present in the PIRL-created data infrastructure in three ways. First, PIRL does not link records of individuals who do not attend the three clinics used in this PhD research. Expanding PIRL activities to capture data in smaller health posts or clinics neighbouring the surveillance area (as discussed in Section 8.3.1) would help mitigate this issue. Second, the number of clinic attendees were sometimes too many to consent into PIRL on a given day. Therefore, some individuals may have passively refused to participate by not agreeing to meet with a fieldworker without first hearing about the study. At the end of PIRL data collection, I showed that there was good coverage in each of the clinics (79% for CTC, 89% in HTC, and 91% in ANC). However, if individuals who self-select not to participate differ on any characteristics examined in this thesis, there is the potential for selection bias. Third, among PIRL participants who reported a residency history in the surveillance area, 16% were not matched to an HDSS record. Session-specific notes stored in the software and discussions with fieldworkers suggested likely reasons (usually in combination with each other) why an HDSS record was not found for these individuals. First, the chance an HDSS enumerator contacted any respondent in a household was reduced as the household size decreased, particularly in households with one or two members. Second, HDSS rounds were usually conducted during the work day and may fail to capture individuals whose employment requires them away from home for extended periods of time. There is a chance they had an HDSS record, but it was simply not found. It is also possible that these individuals were truly not captured in the HDSS system, although that is considered unlikely as response rates during HDSS rounds are approximately 98%.[68] If unsuccessful record linkage among true residents was associated with any of the outcomes or variables assessed throughout this thesis, there would be a potential for systematic bias in the results. Further analyses of these three levels of missingness and if or how they bias analyses using only PIRL-linked records are warranted.

Of note, the percentage of participants who reported residency history but were not matched did not differ significantly by clinic. Therefore, analyses comparing patients between clinics will be prone to less selection bias than analyses among patients in any one particular clinic. The rationale for the conclusion is that the same bias that may affect the selection of patients in one clinic is likely the same as another clinic, and that these biases will cancel out in any analysis comparing patients across clinics.

A limitation of the search database in the existing implementation of the PIRL software is that it can only be as current as the most recently completed HDSS round. Therefore, newer residents to Kisesa, such as children and adults who first move into the HDSS area or infants born after the last HDSS round, will not have an HDSS record. The HDSS database that was searched by the PIRL software only extended through round 29, which ended in late 2014. The software required fieldworkers to input the year of first residence in the HDSS area, so that individuals arriving after round 29 could be flagged for subsequent analyses. During the study period, about one in four participants reported first residence after 2014, of whom were mostly newborns in the ANC and individuals obtaining HIV testing and counselling in the HTC (Table 4.1 on page 78). If PIRL activities recommence in Kisesa, it will be imperative to update the HDSS database to capture more recent residency episodes.

The HIV services offered in Kisesa health centre are not limited to Kisesa residents. About one in three participants approached for this PhD research reported no residence history in the Kisesa HDSS surveillance area. Interestingly, HDSS records were found for approximately 10% of these participants. This finding is potentially related to participants not aware of the surveillance area bounds, not answering truthfully through fear of being identified as having used HIV services, or it may be indicative of the HDSS survey design. In an HDSS round, an enumerator visits each household in the surveillance area. Households are self-defined as ‘a group of people living together in the same compound and who regularly eat together from the same pot.’[68] At each household, only one respondent reports information of all residents in the household. The respondents are normally heads of household though, on some occasions, the respondent is another adult household member who is well informed about the household.[68] Therefore, participants who may casually stay at a household in Kisesa may consider themselves a non-resident whereas the head of household may have reported their residence during an HDSS round. Further characterisation of these 10% of individuals who report no residence, but for whom a record is found, is warranted.

Given the novelty of probabilistic linkage and the lack of high quality, gold standard links in Kisesa, the software was originally developed to include match probabilities calculated from Agincourt data for each identifier. Concerns about not having match probabilities specific to Kisesa were alleviated after the match percentage attained by the most experienced fieldworker in the first month of PIRL was 85%. In addition, as high-quality links made by PIRL began accumulating in the back-end database, I regularly monitored the match probabilities for each identifier. There were no meaningful differences identified. I further explored a multitude of linkage algorithms by varying match probabilities or number of identifiers, and the largest difference between the originally coded algorithm and a tested algorithm was the one with a limited set of identifiers, as presented in Chapter 4.

A central registration office could resolve issues with the complicated identification systems used in Kisesa health centre. Each clinic in the study used a different identification system (Figure 2.6 on page 44). In addition, the HTC and ANC identifiers do not link together all medical records for a given individual. The PIRL software was designed to capture all identifiers from each of these clinics, and our approach was to ask participants for all clinic identifiers they had available to increase the chances of linking records across clinics. A more ideal solution would be to create a unique identification system for individuals who access the Tanzanian healthcare system and issue these numbers on durable cards that could be kept by each attendee and used in future. The medical record systems in each clinic could capture these identifiers, although clinic staff would need to commit to these changes and be trained to collect them. A central registration office in Kisesa health centre, as suggested in Section 8.3.1, would help distribute PIRL responsibilities more evenly among the fieldworkers so that all clinic attendees could be offered the chance to participate and any refusal could be counted.

The stability of funding for PIRL activities in Kisesa remains a key concern for linkage between the community and clinic data. The original intention was that funding for PIRL would be subsumed by larger grants obtained for HDSS rounds. However, future HDSS rounds have yet to secure stable funding. Gaps in funding cause gaps in data collection, primarily in the ANC logbooks. Nurses in the ANC do not store complete logbooks within the clinic, rather completed books are shipped to larger government facilities for storage. The hope is that this PhD thesis acts as a catalyst to obtain future funds for continued and expanded linkage operations in Kisesa.

8.5.3 Strengths of this thesis

The primary strength of this thesis was the multifaceted approach taken to investigate the effectiveness of PIRL and its utility for monitoring HIV service uptake in a rural

Tanzanian community. The methodological component of this thesis demonstrated the success of PIRL and its superiority to automated linkage, the latter of which may have been an attractive, less resource-intensive approach to linkage in rural settings like Kisesa.

The utility of the linked data infrastructure created by PIRL was further demonstrated by substantive analyses investigating how PLHIV progress through HIV services from diagnosis to care. A key strength of these substantive analyses was the availability of multiple sources of data that held complementary information that could be used for validation or comparative purposes. First, this thesis provided a novel analysis that validated self-reported HIV testing histories against directly-observed HIV testing histories as captured in the three PIRL clinics to highlight bias in a widely used indicator for the global HIV response. Second, this thesis was the first to calculate linkage to care rates following a new HIV diagnosis among those diagnosed in a community-based sero-survey and compared these rates to those from facility-based HTC. Findings generated from these two substantive analyses promote expansion of PIRL activities throughout Kisesa and stimulate other research projects that could not have been conducted without the data infrastructure created in this thesis.

The use of secondary data provided by the TAZAMA Project was another strength of this thesis. The analyses conducted in this thesis could not have been conducted without the longitudinal HDSS and sero-survey data collected by the TAZAMA Project. The HDSS data was the backbone in the PIRL software, and the sero-survey data was prominently used in both substantive analyses. This PhD research also benefitted from its association with TAZAMA – a well-known and respected organisation in Kisesa – as there were no active refusals to participate in this PhD research.

8.5.4 Limitations of this thesis

Careful consideration should be given to analyses using the emerging linked data infrastructure so as to not introduce survival bias into any results. The analyses presented in Chapter 7 were limited to individuals who were first diagnosed with HIV during PIRL activities since linkage to care rates among those with historical HIV diagnoses would have been under-estimated due to individuals who may have died or out-migrated before having the chance to participate in this research. It is likely that most analyses using the linked data infrastructure will also be limited to being prospective in nature, particularly longitudinal analyses like analysing the HIV care continuum. However, as mentioned in Section 8.4, identifying an automated linkage algorithm that could be employed to link records of individuals who either did not attend any of the PIRL clinics or were not successfully linked. While this thesis tested automated linkage algorithms that did not perform well in terms of sensitivity and PPV, adding variables into the linkage algorithm (and collecting them in the HDSS and clinics) that are stable over time (e.g., mother/father names, mobile telephone number) could improve the performance of automated linkage in this setting. Using both PIRL and automated linkage approaches could derive a more holistic data infrastructure that could potentially be used for retrospective analyses. However, further analysis on the impact of any linkage errors associated with the automate linkage would need to be conducted. In addition, the issue of obtaining informed consent from individuals who do not attend clinics, passively or actively refuse participation in the clinics, or those who have out-migrated or died would need to be addressed.

The potential for selection bias was present in the sero-survey data. Sero-survey participation rates have fluctuated over time (43-86% of eligible adults). If participation was associated with duration of HIV infection whereby sicker residents eligible for the sero-survey did not participate (i.e., they sought HIV testing elsewhere rather than wait for the next sero-survey round), the proportion of PLHIV who linked to care after a HIV-positive diagnosis in a sero-survey could have been underestimated (Chapters 5 and 7). Therefore, comparisons of linkage to care between individuals newly diagnosed in a sero-survey and facility-based VCT could potentially be over-estimated. In addition, the associations found between linkage to care and potential risk factors, including sex and distance to clinic, could be biased if participation differed in terms of these characteristics.

Similar considerations need to be made for the analysis presented in Chapter 6 as it relied on three rounds of sero-survey data. If participation was associated with non-disclosure of HIV testing history, the results may have been biased. For example, if participants were more likely to disclose their HIV testing history than non-participants, the proportion of non-disclosures would have been under-estimated. In this case, the estimates of bias in the ‘first 90’ would have also been under-estimated. Further, the many associations found with non-disclosure of HIV testing history may be systematically biased if any of these factors was associated with sero-survey participation.

Several of the analyses presented in this thesis suffered from small sample sizes and therefore were underpowered. In Chapter 5, the small number of newly diagnosed individuals linked by PIRL did not allow for linkage bias to be assessed at match score thresholds higher than the 75th percentile. However, the match score thresholds that could be tested provided sufficient evidence that linkage was associated with both the exposure and outcome. Therefore, conditioning or limiting the analyses to records that were linked could therefore induce a protective relationship between the exposure and outcome, as was found in this analysis. Further exploration is needed to determine if multiple imputation to handle missing values due to unlinked records could potentially correct for this bias.

The substantive analyses also had issues with small sample sizes. In Chapter 6, the limited number of PLHIV resulted in insufficient power detect significant associations with non-disclosure of HIV testing history in this group. In Chapter 7, small numbers of people newly diagnosed with HIV in the ANC resulted in large standard errors of regression estimates. Similarly, there were too few repeat testers during the study period to allow for a separate analysis; therefore, they were dropped from the analysis. Continuing (and expanding) PIRL in Kisesa would continue to increase the sample size of PIRL-derived links and therefore support eligibility for these analyses.

The linkage activities conducted for this thesis did not capture HIV services outside of the surveillance area or at smaller health posts within Kisesa. The number of HIV tests delivered at health posts were minimal and ART was not made available until after the study period, so estimates were likely unaffected. However, it is likely that some Kisesa residents were diagnosed or obtained HIV care and treatment outside of Kisesa. If the decision to obtain HIV tests or care outside of Kisesa ward was related to any of the outcomes in this thesis (e.g., modality of HIV testing, non-disclosure of HIV testing history) or any other examined factor, the results in this paper may be subject to bias. In particular, the proportion who linked to care is likely to be underestimated. Expanding linkage to popular clinics adjacent to Kisesa most likely to be used by Kisesa residents would minimise this concern and strengthen the findings.

8.6 DISSEMINATION

8.6.1 Local investigators and healthcare authorities

As part of data monitoring for quality and reliability issues, I distributed a monthly record linkage report (Appendix 10.5) to all TAZAMA, NIMR, and LSHTM investigators and staff involved in this PhD research, including the fieldworkers. All published manuscripts have been sent to the Magu District Medical Officer for review prior to and after publication. I will also prepare a brief report highlighting the key findings from this thesis and the programme and policy recommendations detailed in Section 8.3 to TAZAMA and NIMR investigators who will distribute to local policymakers, such as the National AIDS Control Program in the Ministry of Health or the Tanzania Commission for AIDS in the Prime Minister’s Office. I will also create materials to be posted on the ALPHA Network website to reach other eastern and southern Africa HDSS sites that may wish to initiate record linkage.

8.6.2 Researchers – publications and conference presentations

All five research papers presented in this PhD thesis have either been published or submitted to journals for peer-review.

I have given eight presentations detailing various aspects of this PhD research at domestic and international conferences. Select presentations can be found in Appendix 10.1. Presentation #5 was originally accepted as a poster presentation, but I was subsequently given the opportunity to present the work as both a poster and an oral presentation. All conferences presentations are listed in chronological order in Table 8.2.

Following my presentation at UAPS (Presentation 1), I was approached by a representative of the International Union for the Scientific Study of Population (IUSSP) Scientific Panel on Innovations in Strengthening Civil Registration and Vital Statistics (CRVS) Systems. I accepted his invitation to present my PhD research at an expert group meeting co-organised by the Population Association of America (PAA) and held at The World Bank in Washington, D.C. in April 2016 titled, “Towards the next generation of record-linkage studies to advance data quality assessment of CRVS systems in low- and middle-income countries.” More information about the meeting can be found at: http://iussp.org/en/towards-next-generation-record-linkage-studies. Additionally, slides from my presentation can be found in Appendix 10.1.3.

I have also had the opportunity to present my PhD research at various informal meetings at LSHTM and around Tanzania. A highlight was being awarded the ‘Best Poster Award’ at the 2016 LSHTM Research Degree Poster Day. All non-conference presentations are listed in chronological order in Table 8.3. The presentation delivered at the Measurement & Surveillance of HIV Epidemics (MeSH) Consortium International Scientific Symposium was presented by my primary supervisor, Basia Żaba, on my behalf because she was already planning on attending and I had limited travel funds available.

Table 8.2: Conference presentations of this PhD research

Presentation	Title	Year	Conference	Type	Location	Appendix
1	Real-time record linkage between demographic surveillance and health facility data for monitoring access and utilization of services in rural Tanzania	2015	Union for African Population Studies (UAPS)	Oral	Pretoria, South Africa	10.1.1
2	Real-time record linkage between HDSS and health facility data in rural Tanzania	2017	British Society for Population Studies (BSPS)	Oral	Liverpool, UK	
3	Bias in the 'first 90' of the UNAIDS 90-90-90 target: evidence from a community cohort study with linked clinical data	2017	British Society for Population Studies (BSPS)	Oral	Liverpool, UK	10.1.4
4	Bias in the 'first 90' of the UNAIDS 90-90-90 target: evidence from a community cohort study with linked clinical data	2017	British Society for Population Studies (BSPS)	Poster	Liverpool, UK	
5	Point-of-contact interactive record linkage (PIRL) between demographic surveillance and healthy facility data in rural Tanzania	2017	International Population Conference of the International Union for the Scientific Study of Population (IUSSP)	Oral	Cape Town, South Africa	10.1.5
6	Underreporting of HIV positive diagnosis and its implications for measuring progress along the HIV treatment cascade: evidence from a community cohort study with linked clinical data	2017	International Population Conference of the International Union for the Scientific Study of Population (IUSSP)	Poster	Cape Town, South Africa	
7	Time from HIV diagnosis to care by testing modality in a rural Tanzanian community	2018	International Workshop on HIV Observational Databases (IWHOD)	Poster	Fuengirola, Spain	10.1.6
8	Impact of linkage quality on inferences drawn from analyses using imperfectly matched data with high rates of linkage errors	2018	International Population Data Linkage Network (IPDLN)	Oral	Banff, Canada	

Table 8.3: Non-conference presentations of this PhD research

Presentation	Title	Year	Conference	Type	Location	Appendix
1	Introducing Record Linkage	2015	LSHTM-Tanzania Network Launch Meeting	Oral	Dar es Salaam, Tanzania	
2	Real-time record linkage in Kisesa ward, Tanzania	2015	LSHTM-Tanzania Network Meeting during LSHTM Week	Oral	London, UK	
3	Monitoring access to HIV services in Kisesa ward, Taznania: Real-time record linkage	2015	Mwanza Intervention Trials Unit Board of Directors Meeting	Oral	Mwanza, Tanzania	
4	Real-time record linkage in rural Tanzania	2015	Population Studies Group Seminar	Oral	London, UK	
5	Using record linkage to improve engagement in HIV care	2016	LSHTM Research Degree Poster Day	Poster	London, UK	
6	Real-time record linkage between HDSS and health facility data in rural Tanzania	2016	Expert meeting co-organised by IUSSP and PAA	Oral	Washington, DC	
7	Underreporting of HIV test history in population-based surveys: implications for estimating the UNAIDS 90-90-90 target	2017	MeSH Consortium International Scientific Symposium	Oral	Muldersdrift, South Africa	
8	Developing and implementing point-of-contact interactive record linkage (PIRL) to measure patterns of HIV service utilisation in Tanzania	2018	LSHTM-Tanzania Network Meeting	Oral	London, UK	

8.6.3 Researchers – other

The findings of relative bias up to 20% in the estimation of the first UNAIDS 90-90-90 target detailed in Chapter 6 were communicated informally to the UNAIDS Reference Group on Estimates, Modelling and Projections. I am currently formulating plans with the group on the best strategy to formally present my findings to inform their national model-based estimates of the 'first 90'.

8.6.4 Feedback to funders

Annual reports were provided to ESRC via the ResearchFish platform now used by all major UK Research Councils. Continued reporting on any output resulting from this PhD research will occur up to three years after the award of the PhD. In addition, a separate report detailing the Advanced Quantitative Methods training I undertook during the PhD will be prepared and submitted to the ESRC.

8.6.5 Data sharing

The entire linked data infrastructure created by this PhD research will be documented and securely archived on the TAZAMA server held on the NIMR Mwanza campus.

As detailed in Section 3.7, identifiable data captured in this PhD thesis are unable to be shared with anyone outside the immediate study team due to ethical clearances. However, applications to access the anonymised data for collaborative analysis are encouraged and can be made by contacting the project coordinator for the Kisesa HDSS, Mark Urassa (urassamark@yahoo.co.uk), or by contacting the ALPHA Network team (alpha@lshtm.ac.uk).

8.7 Conclusions

This PhD research introduced PIRL in a rural Tanzanian population to link community data with medical records from a health centre that serves the population with the goal of producing an emerging data source that could be used to monitor utilisation of HIV care and treatment services. This implementation of PIRL was made locally-relevant to Kisesa and had no active refusals by those approached to participate in the study. PIRL performed well in this setting that lacks unique identifiers for individuals in the population; the proportion of residents whose records were linked rivalled other sub-Saharan African linkage studies in settings with national personal identifiers.

The methodological analyses in this thesis described the PIRL ecosystem and compared PIRL to less resource-intensive alternatives. Until future analyses investigate methods to provide more robust results using data with considerable linkage errors, the findings presented in this thesis suggest that researchers in similar settings who wish to perform probabilistic record linkage should allocate resources toward PIRL or a similar system. Further, support should be sought to continue and expand PIRL activities in Kisesa.

The utility of the linked data infrastructure created by PIRL was demonstrated by substantive analyses investigating how PLHIV progress through HIV services from diagnosis to care. The linked data were also used to substitute self-reported health service use collected in population-based surveys to identify bias in a commonly used indicator for monitoring progression toward a global UNAIDS target. These results can be used to formulate updated algorithms to more accurately estimate HIV service uptake.

The multifaceted approach undertaken in this PhD research allowed for the synthesis of findings from both methodological and substantive research, which was helpful to provide programme and policy recommendations and to inform future research. This thesis provided evidence that can help to improve measurement of HIV service use in this community. This thesis also stands as evidence that a linked data source allows for novel and important analyses of HIV service use that promotes the continuation and expansion of PIRL within Kisesa, and exploration of PIRL in other HDSS sites and beyond.

9 References

1. UNAIDS. 90-90-90: an ambitious treatment target to help end the AIDS epidemic. 2014.

2. Haber N, Pillay D, Porter K, Barnighausen T. Constructing the cascade of HIV care: methods for measurement. Curr Opin HIV AIDS. 2016;11(1):102-8.

3. Haber N, Naidu K, Pillay D, Barnighausen T. HIV System Assessment with Longitudinal Treatment Cascade in Kwazulu-Natal, South Africa. Population Association of America; San Diego, CA2015.

4. Cohen MS, Smith MK, Muessig KE, Hallett TB, Powers KA, Kashuba AD. Antiretroviral treatment of HIV-1 prevents transmission of HIV-1: where do we go from here? The Lancet. 2013;382(9903):1515-24.

5. Granich RM, Gilks CF, Dye C, de Cock KM, Williams BG. Universal voluntary HIV testing with immediate antiretroviral therapy as a strategy for elimination of HIV transmission: a mathematical model. Lancet. 2009;373:48-57.

6. Montaner JS, Lima VD, Harrigan PR, Lourenco L, Yip B, Nosyk B, et al. Expansion of HAART Coverage Is Associated with Sustained Decreases in HIV/AIDS Morbidity, Mortality and HIV Transmission: The ''HIV Treatment as Prevention'' Experience in a Canadian Setting. PLoS ONE. 2014;9(2).

7. UNAIDS. The gap report. Geneva: Joint United Nations Programme on HIV/AIDS; 2014.

8. UNAIDS. Ending AIDS: Progress towards the 90-90-90 targets. 2017.

9. UNAIDS. AIDSinfo 2017 [Available from: http://aidsinfo.unaids.org].

10. Wringe A, Floyd S, Kazooba P, Mushati P, Baisley K, Urassa M, et al. Antiretroviral therapy uptake and coverage in four HIV community cohort studies in sub-Saharan Africa. Tropical Medicine and International Health. 2012;17(8):e38-48.

11. Cawley C, Wringe A, Todd J, Gourlay A, Clark B, Masesa C, et al. Risk factors for service use and trends in coverage of different HIV testing and counselling models in northwest Tanzania between 2003 and 2010. Tropical medicine & international health : TM & IH. 2015.

12. Isingo R, Wringe A, Todd J, Urassa M, Mbata D, Maiseli G, et al. Trends in the uptake of voluntary counselling and testing for HIV in rural Tanzania in the context of the scale up of antiretroviral therapy. Tropical medicine & international health : TM & IH. 2012;17(8):e15-25.

13. Nsigaye R, Wringe A, Roura M, Kalluvya S, Urassa M, Busza J, et al. From HIV diagnosis to treatment: evaluation of a referral system to promote and monitor access to antiretroviral therapy in rural Tanzania. Journal of the International AIDS Society. 2009;12(1):31.

14. Sankoh O, Network I. CHESS: an innovative concept for a new generation of population surveillance. Lancet Glob Health. 2015;3(12):e742.

15. Wellcome Trust. Enabling Data Linkage to Maximise the Value of Public Health Research Data: full report. 2015.

16. Herrett E, Gallagher AM, Bhaskaran K, Forbes H, Mathur R, van Staa T, et al. Data Resource Profile: Clinical Practice Research Datalink (CPRD). Int J Epidemiol. 2015;44(3):827-36.

17. Smeeth L, Cook C, Fombonne E, Heavey L, Rodrigues LC, Smith PG, et al. MMR vaccination and pervasive developmental disorders: a case-control study. Lancet. 2004;364(9438):963-9.

18. Smeeth L, Thomas SL, Hall AJ, Hubbard R, Farrington P, Vallance P. Risk of myocardial infarction and stroke after acute infection or vaccination. N Engl J Med. 2004;351(25):2611-8.

19. Bhaskaran K, Douglas I, Forbes H, dos-Santos-Silva I, Leon DA, Smeeth L. Body-mass index and risk of 22 specific cancers: a population-based cohort study of 5.24 million UK adults. Lancet. 2014;384(9945):755-65.

20. Dankovchik J, Hoopes MJ, Warren-Mears V, Knaster E. Disparities in Life Expectancy of Pacific Northwest American Indians and Alaska Natives: Analysis of Linkage-Corrected Life Tables. Public Health Reports. 2015;130(1):71-80.

21. Joubert J, Bradshaw D, Kabudula C, Rao C, Kahn K, Mee P, et al. Record-linkage comparison of verbal autopsy and routine civil registration death certification in rural north-east South Africa: 2006-09. International Journal of Epidemiology. 2014;43(6):1945-58.

22. Han J, Kamber M. Data mining: concepts and techniques. 2 ed. Morgan Kaufmann, editor2006.

23. Wilson DR, editor Beyond Probabilistic Record Linkage: Using Neural Networks and Complex Features to Improve Genealogical Record Linkage. International Joint Conference on Neural Networks; 2011; San Jose, CA.

24. Roos LL, Wajda A, Nicol JP. The Art and Science of Record Linkage: Methods that Work with Few Identifiers. Comput Biol Med. 1986;16(1):45-57.

25. Jaro MA. Probabilistic linkage of large public health data files. Stat Med. 1995;14:491-8.

26. Meray N, Reitsma JB, Ravelli ACJ, Bonsel GJ. Probabilistic record linkage is a valid and transparent tool to combine databases without a patient identification number. Journal of Clinical Epidemiology. 2007;60(9):883-91.

27. Sayers A, Ben-Shlomo Y, Blom AW, Steele F. Probabilistic record linkage. Int J Epidemiol. 2015.

28. Newcombe H, Kennedy J, Axford S, James A. Automatic Linkage of Vital Records. Science. 1959;130(3381):954-9.

29. Fellegi IP, Sunter AB. A Theory for Record Linkage. J Am Stat Assoc. 1969;64(328):1183-210.

30. Baldwin E, Johnson K, Berthoud H, Dublin S. Linking mothers and infants within electronic health records: a comparison of deterministic and probabilistic algorithms. Pharmacoepidemiology and Drug Safety. 2015;24(1):45-51.

31. Gill L, Goldacre M, Simmons H, Bettley G, Griffith M. Computerized linking of medical records - methodological guidelines. Journal of Epidemiology and Community Health. 1993;47(4):316-9.

32. Gomatam S, Carter R, Ariet M, Mitchell G. An empirical comparison of record linkage procedures. Stat Med. 2002;21(10):1485-96.

33. Zhu Y, Matsuyama Y, Ohashi Y, Setoguchi S. When to conduct probabilistic linkage vs. deterministic linkage? A simulation study. Journal of Biomedical Informatics. 2015;56:80-6.

34. Newcombe H. Strategy and art in automated death searches. American Journal of Public Health. 1984;74(12).

35. Nakhaee F, McDonald A, Black D, Law M. A feasible method for linkage studies avoiding clerical review: linkage of the national HIV/AIDS surveillance databases with the National Death Index in Australia. Australian and New Zealand Journal of Public Health. 2007;31(4):308-12.

36. Quantin C, Binquet C, Allaert FA, Cornet B, Pattisina R, Leteuff G, et al. Decision analysis for the assessment of a record linkage procedure - Application to a perinatal network. Methods of Information in Medicine. 2005;44(1):72-9.

37. Corbell C, Katjitae I, Mengistu A, Kalemeera F, Sagwa E, Mabirizi D, et al. Records linkage of electronic databases for the assessment of adverse effects of antiretroviral therapy in sub-Saharan Africa. Pharmacoepidemiology and Drug Safety. 2012;21(4):407-14.

38. Kabudula CW, Clark BD, Gómez-Olivé FX, Tollman S, Menken J, Reniers G. The promise of record linkage for assessing the uptake of health services in resource constrained settings: a pilot study from South Africa. BMC Med Res Methodol. 2014;14(71).

39. Gourlay A, Wringe A, Todd J, Cawley C, Michael D, Machemba R, et al. Factors associated with uptake of services to prevent mother-to-child transmission of HIV in a community cohort in rural Tanzania. Sex Transm Infect. 2015.

40. Fure E. Interactive Record Linkage. Demographic Research. 2000;3(11).

41. Kum HC, Krishnamurthy A, Machanavajjhala A, Reiter MK, Ahalt S. Privacy preserving interactive record linkage (PPIRL). J Am Med Inform Assoc. 2014;21(2):212-20.

42. Randall SM, Ferrante AM, Boyd JH, Bauer JK, Semmens JB. Privacy-preserving record linkage on large real world datasets. Journal of Biomedical Informatics. 2014;50:205-12.

43. RumeauRouquette C. Linking individual data: Methods of record linkage. Revue D Epidemiologie Et De Sante Publique. 1997;45(3):248-56.

44. Schmidlin K, Clough-Gorr KM, Spoerri A, Grp SNCS. Privacy Preserving Probabilistic Record Linkage (P3RL): a novel method for linking existing health-related data and maintaining participant confidentiality. BMC Medical Research Methodology. 2015;15.

45. Baldi I, Ponti A, Zanetti R, Ciccone G, Merletti F, Gregori D. The impact of record-linkage bias in the Cox model. J Eval Clin Pract. 2010;16(1):92-6.

46. Moore CL, Amin J, Gidding HF, Law MG. A new method for assessing how sensitivity and specificity of linkage studies affects estimation. PLoS One. 2014;9(7):e103690.

47. Bohensky MA, Jolley D, Sundararajan V, Evans S, Pilcher DV, Scott I, et al. Data linkage: a powerful research tool with potential problems. BMC health services research. 2010;10:346.

48. Brinkhof MW, Dabis F, Myer L, Bangsberg DR, Boulle A, Nash D, et al. Early loss of HIV-infected patients on potent antiretroviral therapy programmes in lower-income countries. Bull World Health Organ. 2008;86(7):559-67.

49. Coetzee D, Hildebrand K, Boulle A, Maartens G, Louis F, Labatala V, et al. Outcomes after two years of providing antiretroviral treatment in Khayelitsha, South Africa. AIDS. 2004;18(6):887-95.

50. Ford N, Kranzer K, Hilderbrand K, Jouquet G, Goemaere E, Vlahakis N, et al. Early initiation of antiretroviral therapy and associated reduction in mortality, morbidity and defaulting in a nurse-managed, community cohort in Lesotho. AIDS. 2010;24(17):2645-50.

51. Fox MP, Sanne IM, Conradie F, Zeinecker J, Orrell C, Ive P, et al. Initiating patients on antiretroviral therapy at CD4 cell counts above 200 cells/microl is associated with improved treatment outcomes in South Africa. AIDS. 2010;24(13):2041-50.

52. Kigozi BK, Sumba S, Mudyope P, Namuddu B, Kalyango J, Karamagi C, et al. The effect of AIDS defining conditions on immunological recovery among patients initiating antiretroviral therapy at Joint Clinical Research Centre, Uganda. AIDS Res Ther. 2009;6:17.

53. Kitahata MM, Gange SJ, Abraham AG, Merriman B, Saag MS, Justice AC, et al. Effect of early versus deferred antiretroviral therapy for HIV on survival. N Engl J Med. 2009;360(18):1815-26.

54. Krentz HB, Auld MC, Gill MJ. The high cost of medical care for patients who present late (CD4 <200 cells/microL) with HIV infection. HIV Med. 2004;5(2):93-8.

55. Lawn SD, Harries AD, Anglaret X, Myer L, Wood R. Early mortality among adults accessing antiretroviral treatment programmes in sub-Saharan Africa. AIDS. 2008;22(15):1897-908.

56. Murphy RA, Sunpath H, Taha B, Kappagoda S, Maphasa KT, Kuritzkes DR, et al. Low uptake of antiretroviral therapy after admission with human immunodeficiency virus and tuberculosis in KwaZulu-Natal, South Africa. Int J Tuberc Lung Dis. 2010;14(7):903-8.

57. Samet JH, Freedberg KA, Savetsky JB, Sullivan LM, Stein MD. Understanding delay to medical care for HIV infection: the long-term non-presenter. Aids. 2001;15(1):77-85.

58. Stangl AL, Wamai N, Mermin J, Awor AC, Bunnell RE. Trends and predictors of quality of life among HIV-infected adults taking highly active antiretroviral therapy in rural Uganda. AIDS care. 2007;19(5):626-36.

59. Toure S, Kouadio B, Seyler C, Traore M, Dakoury-Dogbo N, Duvignac J, et al. Rapid scaling-up of antiretroviral therapy in 10,000 adults in Cote d'Ivoire: 2-year outcomes and determinants. AIDS. 2008;22(7):873-82.

60. Attia S, Egger M, Muller M, Zwahlen M, Low N. Sexual transmission of HIV according to viral load and antiretroviral therapy: systematic review and meta-analysis. AIDS. 2009;23(11):1397-404.

61. Bunnell R, Mermin J, De Cock KM. HIV prevention for a threatened continent: implementing positive prevention in Africa. JAMA. 2006;296(7):855-8.

62. Castilla J, Del Romero J, Hernando V, Marincovich B, Garcia S, Rodriguez C. Effectiveness of highly active antiretroviral therapy in reducing heterosexual transmission of HIV. J Acquir Immune Defic Syndr. 2005;40(1):96-101.

63. Girardi E, Sabin CA, Monforte AD. Late diagnosis of HIV infection: epidemiological features, consequences and strategies to encourage earlier testing. J Acquir Immune Defic Syndr. 2007;46 Suppl 1:S3-8.

64. Vernazza P, Hirschel B, Bernasconi E, Flepp M. HIV transmission under highly active antiretroviral therapy. Lancet. 2008;372(9652):1806-7; author reply 7.

65. World Health Organization. Consolidated Guidelines on HIV Testing Services: 5Cs: consent, confidentiality, counselling, correct results and connection. 2015.

66. Mabuto T, Latka MH, Kuwane B, Churchyard GJ, Charalambous S, Hoffmann CJ. Four models of HIV counseling and testing: utilization and test results in South Africa. PLoS One. 2014;9(7):e102267.

67. UNAIDS. Global AIDS Monitoring 2017: Indicators for monitoring the 2016 United Nations Political Declaration on HIV and AIDS. 2016.

68. Kishamawe C, Isingo R, Mtenga B, Zaba B, Todd J, Clark B, et al. Health & Demographic Surveillance System Profile: The Magu Health and Demographic Surveillance System (Magu HDSS). Int J Epidemiol. 2015;44(6):1851-61.

69. Ministry of Finance. National Accounts of Tanzania Mainland 2007-2016. In: Ministry of Finance, editor. Dar es Salaam2016.

70. National AIDS Control Programme (NACP). National Guidelines for the Management of HIV and AIDS. In: Ministry of Health and Social Welfare, editor. 4 ed. Tanzania2012.

71. National AIDS Control Programme (NACP). National Guidelines for the Management of HIV and AIDS. In: Ministry of Health and Social Welfare, editor. 6 ed. Tanzania2017.

72. Kabudula C, Rentsch CT, Catlett J, Beckles D, Masilela N, Żaba B, et al. PIRL - Point-of-contact Interactive Record Linkage software. https://doi.org/10.5281/zenodo.998867; 2017.

73. Rentsch CT, Kabudula CW, Catlett J, Beckles D, Machemba R, Mtenga B, et al. Point-of-contact Interactive Record Linkage (PIRL): A software tool to prospectively link demographic surveillance and health facility data [version 2; referees: 2 approved]. Gates Open Res. 2018;1(8).

74. Rentsch CT, Reniers G, Kabudula C, Machemba R, Mtenga B, Harron K, et al. Point-of-contact interactive record linkage (PIRL) between demographic surveillance and health facility data in rural Tanzania. International Journal for Population Data Science. 2017;2(1).

75. Kabudula CW, Joubert JD, Tuoane-Nkhasi M, Kahn K, Rao C, Gomez-Olive FX, et al. Evaluation of record linkage of mortality data between a health and demographic surveillance system and national civil registration system in South Africa. Population health metrics. 2014;12.

76. Christen P. Data matching: concepts and techniques for record linkage, entity resolution, and duplicate detection: Springer Science & Business Media; 2012.

77. Harron K, Goldstein H, Dibben C. Methodological developments in data linkage: John Wiley & Sons; 2015.

78. Herzog TN, Scheuren FJ, Winkler WE. Data quality and record linkage techniques: Springer Science & Business Media; 2007.

79. Winkler WE. Overview of Record Linkage and Current Research Directions. Washington, DC: US Bureau of the Census; 2006.

80. Barker M. STRDIST: Stata module to calculate the Levenshtein distance, or edit distance, between strings. Boston College Department of Economics; 2012.

81. Cohen WW, Ravikumar P, Fienberg SE, editors. A comparison of string metrics for matching names and records. KDD workshop on data cleaning and object consolidation; 2003.

82. Levenshtein V. Binary codes capable of correcting deletions, insertions, and reversals. Soviet Physics-Doklady. 1966;10(8):707-10.

83. Monge E, Elkan C. The field matching problem: algorithms and applications. The Association for Computing Machinery's Special Interest Group on Knowledge Discovery and Data Mining (SigKDD); Portland, OR1996.

84. Navarro G. A guided tour to approximate string matching. ACM Computing Surveys. 2001;33(1):31-88.

85. Keskustalo H, Pirkola A, Visala K, Leppanen E, Jarvelin K, editors. Non-adjacent digrams improve mathcing of cross-lingual spelling variants. International Conference on Extending Database Technology; 2003; Manaus, Brazil.

86. Kukich K. Techniques for automatically correcting words in text. ACM Computing Surveys. 1992;24(4):377-439.

87. van Berkel B, De Smedt K, editors. Triphone analysis: a combined method for the correction of orthographical and typographical errors. Second Conference on Applied Natural Language Processing; 1988; Austin, TX.

88. Bartolini I, Ciaccia P, Patella M, editors. String Matching with Metric Trees Using an Approximate Distance. String Processing and Information Retrival; 2002: Lisbon, Portugal.

89. Borgman CL, Siegfried SL. Getty's Synoname™ and its cousins: A survey of applications of personal name-matching algorithms. Journal of the American Society for Information Science. 1992;43(7):459-76.

90. Cilibrasi R, Vitanya P. Clustering by compression. IEEE Transactions on Information Theory. 2005;51(4):1523-45.

91. Friedman C, Sideli R. Tolerating spelling errors during patient validation. Computers and Biomedical Research. 1992;25(5):486-509.

92. Holmes D, McCabe MC, editors. Improving precision and recall for Soundex retrival. Proceedings of the IEEE International Conference on Information Technology-Coding and Computing; 2002; Las Vegas, NV.

93. Lait A, Randell B. An assessment of name matching algorithms. Department of Computer Science, University of Newcastle upon Tyne; 1993.

94. Moreau E, Yvon F, Cappe O, editors. Robust similarity measures for names entities matching. 22nd International Conference on Computational Linguistics-Volume 1; 2008.

95. Naumann F, Herschel M. An introduction to duplicate detection. Synthesis Lectures on Data Management. 2010;2(1):1-87.

96. Odell M. The profit in records management. Systems (New York). 1956;20:20.

97. Ruibin G, Tony K. Syllable alignment: A novel model for phonetic string search. IEICE transactions on information and systems. 2006;89(1):332-9.

98. Sayers A. NYSIIS: Stata module to calculate nysiis codes from string variables. Statistical Software Components. 2014.

99. Taft R. Name search techniques: New york state identification and intelligence system. Albany, NY, Special Rep. 1970(1).

100. Zobel J, Dart P, editors. Phonetic string matching: lessons from information retrieval. ACM SIGIR; 1996; Zurich, Switzerland.

101. Winkler WE. String Comparator Metrics and Enhanced Decision Rules in the Fellegi-Sunter Model of Record Linkage. 1990.

102. Elmagarmid AK, Ipeirotis PG, Verykios VS. Duplicate record detection: a survey. IEEE Transactions on Knowledge and Data Engineering. 2007;19(1).

103. Grannis SJ, Overhage JM, McDonald CJ. Real world performance of approximate string comparators for use in patient matching. Medinfo. 2004:43-7.

104. Yancey WE. Evaluating string comparator performance for record linkage. In: Bureau UC, editor. Washington, DC2005.

105. Day C. Record linkage i: evaluation of commercially available record linkage software for use in NASS: US Department of Agriculture, National Agricultural Statistics Service, Research Division; 1995.

106. Christen P, Churches T, Hegland M. Febrl-a parallel open source data linkage system. Lecture notes in computer science. 2004:638-47.

107. Jurczyk P, Lu JJ, Xiong L, Cragan JD, Correa A. Fine-grained record integration and linkage tool. Birth Defects Research Part A: Clinical and Molecular Teratology. 2008;82(11):822-9.

108. Harron K, Goldstein H, Wade A, Muller-Pebody B, Parslow R, Gilbert R. Linkage, Evaluation and Analysis of National Electronic Healthcare Data: Application to Providing Enhanced Blood-Stream Infection Surveillance in Paediatric Intensive Care. PLoS ONE. 2013;8(12).

109. Schmidlin K, Clough-Gorr KM, Spoerri A, Egger M, Zwahlen M, Swiss National C. Impact of unlinked deaths and coding changes on mortality trends in the Swiss National Cohort. BMC Med Inform Decis Mak. 2013;13:1.

110. Boyd JH, Ferrante AM, Irvine K, Smith M, Moore E, Brown A, et al. Understanding the origins of record linkage errors and how they affect research outcomes. Aust N Z J Public Health. 2016.

111. Jorm L. Routinely collected data as a strategic resource for research: priorities for methods and workforce. Public Health Res Pract. 2015;25(4):e2541540.

112. Bentley JP, Ford JB, Taylor LK, Irvine KA, Roberts CL. Investigating linkage rates among probabilistically linked birth and hospitalization records. BMC Med Res Methodol. 2012;12:149.

113. Austin PC. Balance diagnostics for comparing the distribution of baseline covariates between treatment groups in propensity-score matched samples. Stat Med. 2009;28(25):3083-107.

114. Cohen J. Statistical power analysis for the behavioral sciences. 2 ed: Erlbaum Associates, Hillsdale; 1988.

115. Harron K, Wade A, Gilbert R, Muller-Pebody B, Goldstein H. Evaluating bias due to linkage error in electronic healthcare records. BMC Medical Research Methodology. 2014;14(36).

116. Ford JB, Roberts CL, Taylor LK. Characteristics of unmatched maternal and baby records in linked birth records and hospital discharge data. Paediatr Perinat Epidemiol. 2006;20(4):329-37.

117. Harron KL, Doidge JC, Knight HE, Gilbert RE, Goldstein H, Cromwell DA, et al. A guide to evaluating linkage quality for the analysis of linked data. Int J Epidemiol. 2017;46(5):1699-710.

118. Cole SR, Platt RW, Schisterman EF, Chu H, Westreich D, Richardson D, et al. Illustrating bias due to conditioning on a collider. Int J Epidemiol. 2010;39(2):417-20.

119. Hernan MA, Hernandez-Diaz S, Robins JM. A structural approach to selection bias. Epidemiology. 2004;15(5):615-25.

120. Goldstein H, Harron K, Wade A. The analysis of record-linked data using multiple imputation with data value priors. Stat Med. 2012;31(28):3481-93.

121. Harron K, Goldstein H, Dibben C. Record linkage: a missing data problem. In: Harron K, Dibben C, Goldstein H, editors. Methodological developments in data linkage. London: John Wiley & Sons; 2015.

122. Wartenberg D, Thompson WD. Privacy versus public health: the impact of current confidentiality rules. Am J Public Health. 2010;100(3):407-12.

123. Boyd JH, Guiver T, Randall SM, Ferrante AM, Semmens JB, Anderson P, et al. A Simple Sampling Method for Estimating the Accuracy of Large Scale Record Linkage Projects. Methods Inf Med. 2016;55(3):276-83.

124. World Health Organization. Consolidated strategic information guidelines for HIV in the health sector. Geneva, Switzerland; 2015.

125. Corsi DJ, Neuman M, Finlay JE, Subramanian SV. Demographic and health surveys: a profile. Int J Epidemiol. 2012;41(6):1602-13.

126. Fishel JD, Barrere B, Kishor S. Validity of data on self-reported HIV status and implications for measurement of ARV coverage in Malawi. Calverton, Maryland, USA: ICF International; 2012.

127. Kim AA, Mukui I, Young PW, Mirjahangir J, Mwanyumba S, Wamicwe J, et al. Undisclosed HIV infection and antiretroviral therapy use in the Kenya AIDS indicator survey 2012. Aids. 2016;30(17):2685-95.

128. Tanzania Commission for AIDS (TACAIDS), Zanzibar AIDS Commission (ZAC), National Bureau of Statistics (NBS), Office of the Chief Government Statistician (OCGS), ICF International. Tanzania HIV/AIDS and Malaria Indicator Survey 2011-12. Dar es Salaam, Tanzania: TACAIDS, ZAC, NBS, and ICF International; 2013.

129. Amuyunzu-Nyamongo M, Okeng'o L, Wagura A, Mwenzwa E. Putting on a brave face: the experiences of women living with HIV and AIDS in informal settlements of Nairobi, Kenya. AIDS care. 2007;19 Suppl 1:S25-34.

130. Antelman G, Smith Fawzi MC, Kaaya S, Mbwambo J, Msamanga GI, Hunter DJ, et al. Predictors of HIV-1 serostatus disclosure: a prospective study among HIV-infected pregnant women in Dar es Salaam, Tanzania. AIDS. 2001;15(14):1865-74.

131. Mbonu NC, van den Borne B, De Vries NK. Stigma of People with HIV/AIDS in Sub-Saharan Africa: A Literature Review. J Trop Med. 2009;2009:145891.

132. Mill JE. Shrouded in secrecy: breaking the news of HIV infection to Ghanaian women. J Transcult Nurs. 2003;14(1):6-16.

133. Lyimo RA, Stutterheim SE, Hospers HJ, de Glee T, van der Ven A, de Bruin M. Stigma, disclosure, coping, and medication adherence among people living with HIV/AIDS in Northern Tanzania. AIDS Patient Care STDS. 2014;28(2):98-105.

134. Smith R, Rossetto K, Peterson BL. A meta-analysis of disclosure of one's HIV-positive status, stigma and social support. AIDS care. 2008;20(10):1266-75.

135. Sharma M, Ying R, Tarr G, Barnabas R. Systematic review and meta-analysis of community and facility-based HIV testing to address linkage to care gaps in sub-Saharan Africa. Nature. 2015;528(7580):S77-85.

136. Matovu JK, Makumbi FE. Expanding access to voluntary HIV counselling and testing in sub-Saharan Africa: alternative approaches for improving uptake, 2001-2007. Tropical medicine & international health : TM & IH. 2007;12(11):1315-22.

137. World Health Organization, Unicef. Towards Universal Access: Scaling up Priority HIV/AIDS Interventions in the Health Sector. WHO, Geneva; 2009.

138. Menzies N, Abang B, Wanyenze R, Nuwaha F, Mugisha B, Coutinho A, et al. The costs and effectiveness of four HIV counseling and testing strategies in Uganda. AIDS. 2009;23(3):395-401.

139. Suthar AB, Ford N, Bachanas PJ, Wong VJ, Rajan JS, Saltzman AK, et al. Towards universal voluntary HIV testing and counselling: a systematic review and meta-analysis of community-based approaches. PLoS medicine. 2013;10(8):e1001496.

140. Genberg BL, Naanyu V, Wachira J, Hogan JW, Sang E, Nyambura M, et al. Linkage to and engagement in HIV care in western Kenya: an observational study using population-based estimates from home-based counselling and testing. The Lancet HIV. 2015;2(1):e20-e6.

141. Ruzagira E, Baisley K, Kamali A, Biraro S, Grosskurth H, Working Group on Linkage to HIV Care. Linkage to HIV care after home-based HIV counselling and testing in sub-Saharan Africa: a systematic review. Tropical medicine & international health : TM & IH. 2017;22(7):807-21.

142. Slaymaker E, McLean E, Wringe A, Calvert C, Marston M, Reniers G, et al. The Network for Analysing Longitudinal Population-based HIV/AIDS data on Africa (ALPHA): Data on mortality, by HIV status and stage on the HIV care continuum, among the general population in seven longitudinal studies between 1989 and 2014. Gates Open Res. 2017;1:4.

143. Musheke M, Ntalasha H, Gari S, McKenzie O, Bond V, Martin-Hilber A, et al. A systematic review of qualitative findings on factors enabling and deterring uptake of HIV testing in Sub-Saharan Africa. BMC public health. 2013;13:220.

144. Iwuji CC, Orne-Gliemann J, Larmarange J, Balestre E, Thiebaut R, Tanser F, et al. Universal test and treat and the HIV epidemic in rural South Africa: a phase 4, open-label, community cluster randomised trial. Lancet HIV. 2018;5(3):e116-e25.

145. Hayes R, Floyd S, Schaap A, Shanaube K, Bock P, Sabapathy K, et al. A universal testing and treatment intervention to improve HIV control: One-year results from intervention communities in Zambia in the HPTN 071 (PopART) cluster-randomised trial. PLoS medicine. 2017;14(5):e1002292.

146. Althubaiti A. Information bias in health research: definition, pitfalls, and adjustment methods. J Multidiscip Healthc. 2016;9:211-7.

147. Paulin HN, Blevins M, Koethe JR, Hinton N, Vaz LM, Vergara AE, et al. HIV testing service awareness and service uptake among female heads of household in rural Mozambique: results from a province-wide survey. BMC public health. 2015;15:132.

148. National AIDS Control Programme (NACP). National Guidelines for the Management of HIV and AIDS. In: Ministry of Health and Social Welfare, editor. 5 ed. Tanzania2015.

149. World Health Organization. Guideline on when to start antiretroviral therapy and on pre-exposure prophylaxis for HIV. Geneva, Switzerland; 2015.

150. Neville R, Neville J. What can health care professionals in the United Kingdom learn from Malawi? Hum Resour Health. 2009;7:26.

151. Hayes R, Ayles H, Beyers N, Sabapathy K, Floyd S, Shanaube K, et al. HPTN 071 (PopART): rationale and design of a cluster-randomised trial of the population impact of an HIV combination prevention intervention including universal testing and treatment - a study protocol for a cluster randomised trial. Trials. 2014;15:57.

152. Sabapathy K, Van den Bergh R, Fidler S, Hayes R, Ford N. Uptake of home-based voluntary HIV testing in sub-Saharan Africa: a systematic review and meta-analysis. PLoS medicine. 2012;9(12):e1001351.

153. Hayes RJ, Floyd S, Schaap A, Shanaube K, Yang B, Griffith S, et al., editors. Achieving the first two UNAIDS 90-90-90 targets on completion of a three-year universal testing and treatment (UTT) intervention in the HPTN 071 (PopART) randomised trial in Zambia and South Africa. International AIDS Conference; 2018; Amsterdam, The Netherlands.

154. McGuire M, Pinoges L, Kanapathipillai R, Munyenyembe T, Huckabee M, Makombe S, et al. Treatment initiation, program attrition and patient treatment outcomes associated with scale-up and decentralization of HIV care in rural Malawi. PLoS One. 2012;7(10):e38044.

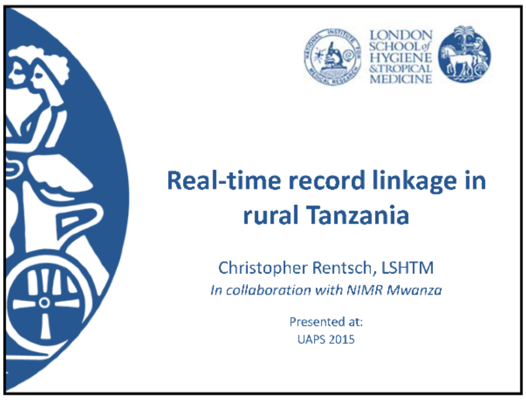
155. Reidy WJ, Sheriff M, Wang C, Hawken M, Koech E, Elul B, et al. Decentralization of HIV care and treatment services in Central Province, Kenya. J Acquir Immune Defic Syndr. 2014;67(1):e34-40.

156. Fuente-Soro L, Lopez-Varela E, Augusto O, Sacoor C, Nhacolo A, Honwana N, et al. Monitoring progress towards the first UNAIDS target: understanding the impact of people living with HIV who re-test during HIV-testing campaigns in rural Mozambique. Journal of the International AIDS Society. 2018;21(4):e25095.

10 Appendices

10.1 CONFERENCE PRESENTATIONS AND POSTERS

10.1.1 Union for African Population Studies (UAPS), 2015



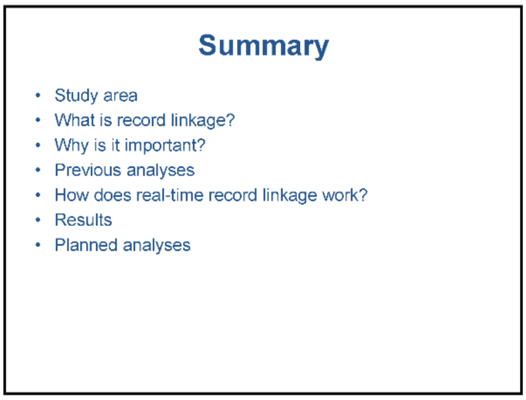
LONDON SCHOOL of HYGIENE & TROPICAL MEDICINE

Real-time record linkage in rural Tanzania

Christopher Rentsch, LSHTM

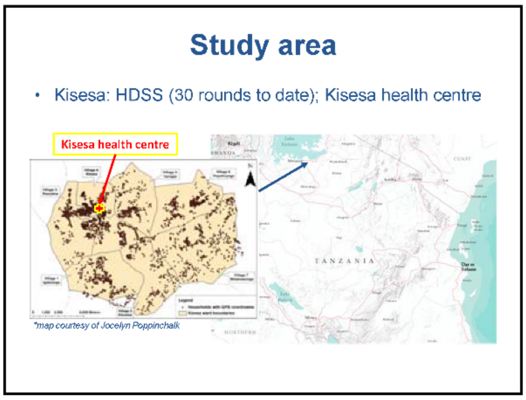
In collaboration with NIMR Mwanza

Presented at:
UAPS 2015



Summary

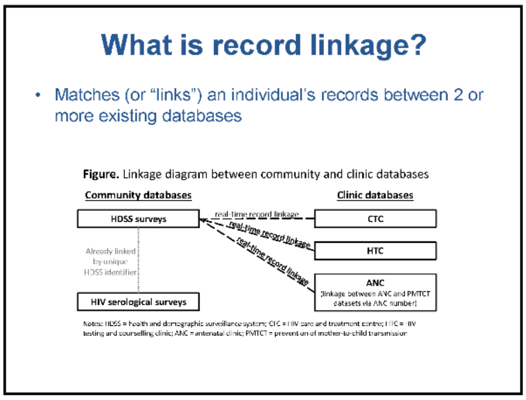
- Study area
- What is record linkage?
- Why is it important?
- Previous analyses
- How does real-time record linkage work?
- Results
- Planned analyses



Study area

- Kisesa: HDSS (30 rounds to date); Kisesa health centre

map courtesy of Jocelyn Poppinchalk



What is record linkage?

- Matches (or "links") an individual's records between 2 or more existing databases

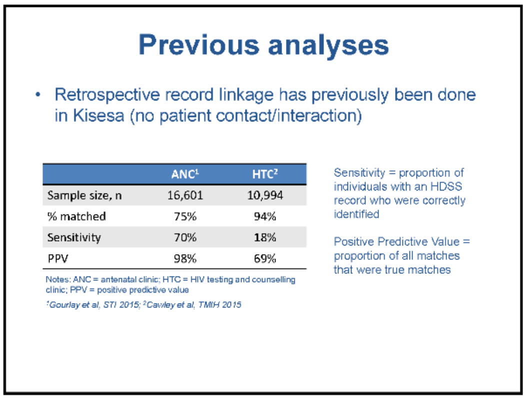
Figure. Linkage diagram between community and clinic databases

Notes: HDSS = health and demographic surveillance system; CTC = HIV care and treatment centre; HTC = HIV testing and counselling clinic; ANC = antenatal clinic; PMTCT = prevention of mother-to-child transmission

Why is record linkage important?

- Access to HIV services has risen over the years, but still remains low (~55% ever tested[1]; 60% ART coverage[2])
- Currently, we can only research those who access healthcare services
- By linking records, we can better investigate service utilization, clinical outcomes, and lost to follow-up from a population perspective
- Explore modifiable mediators that allow us to better design local health systems and simple interventions to improve HIV service access

[1] *WHO, 2013*
[2] *DHS Comp Reports, 2013*



Previous analyses

- Retrospective record linkage has previously been done in Kisesa (no patient contact/interaction)

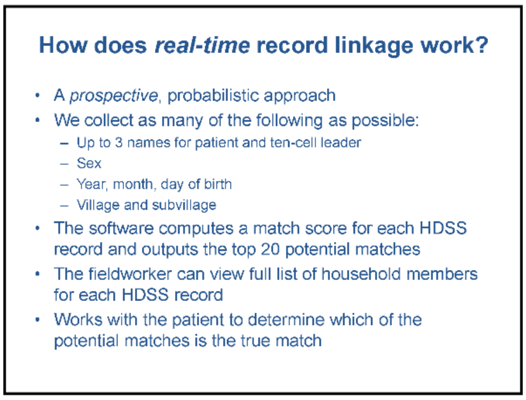
	ANC[1]	HTC[2]
Sample size, n	16,601	10,994
% matched	75%	94%
Sensitivity	70%	18%
PPV	98%	69%

Notes: ANC = antenatal clinic; HTC = HIV testing and counselling clinic; PPV = positive predictive value

[1] *Gourlay et al, STI 2015;* [2] *Cawley et al, TMIH 2015*

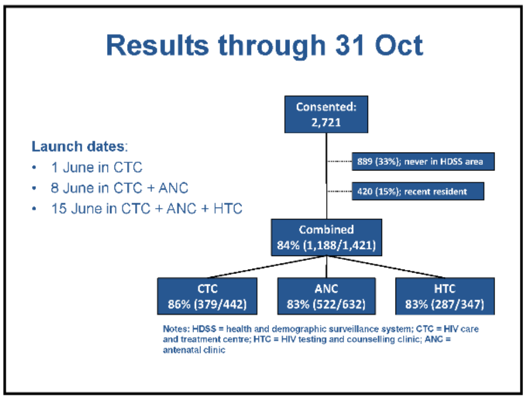
Sensitivity = proportion of individuals with an HDSS record who were correctly identified

Positive Predictive Value = proportion of all matches that were true matches



How does *real-time* record linkage work?

- A *prospective*, probabilistic approach
- We collect as many of the following as possible:
  - Up to 3 names for patient and ten-cell leader
  - Sex
  - Year, month, day of birth
  - Village and subvillage
- The software computes a match score for each HDSS record and outputs the top 20 potential matches
- The fieldworker can view full list of household members for each HDSS record
- Works with the patient to determine which of the potential matches is the true match



Results through 31 Oct

Launch dates:

- 1 June in CTC
- 8 June in CTC + ANC
- 15 June in CTC + ANC + HTC

Consented: 2,721

889 (33%); never in HDSS area

420 (15%); recent resident

Combined 84% (1,188/1,421)

CTC 86% (379/442)

ANC 83% (522/632)

HTC 83% (287/347)

Notes: HDSS = health and demographic surveillance system; CTC = HIV care and treatment centre; HTC = HIV testing and counselling clinic; ANC = antenatal clinic

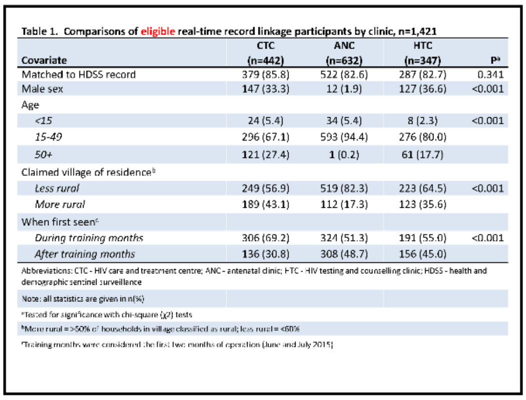


Table 1. Comparisons of eligible real-time record linkage participants by clinic, n=1,421

Covariate	CTC (n=442)	ANC (n=632)	HTC (n=347)	P[a]
Matched to HDSS record	379 (85.8)	522 (82.6)	287 (82.7)	0.341
Male sex	**147** (33.3)	12 (**1.9**)	127 (36.6)	<0.001
Age				
<15	24 (5.4)	34 (5.4)	8 (2.3)	<0.001
15-49	296 (67.1)	593 (94.4)	276 (80.0)	
50+	**121** (27.4)	**1** (0.2)	**61** (17.7)	
Claimed village of residence[b]				
Less rural	249 (56.9)	519 (82.3)	223 (64.5)	<0.001
More rural	**189** (43.1)	**112** (17.3)	123 (35.6)	
When first seen[c]				
During training months	306 (69.2)	324 (51.3)	191 (55.0)	<0.001
After training months	**136** (30.8)	308 (48.7)	156 (45.0)	

Abbreviations: CTC - HIV care and treatment centre; ANC - antenatal clinic; HTC - HIV testing and counselling clinic; HDSS - health and demographic sentinel surveillance

Note: all statistics are given in n(%)

[a] Tested for significance with chi-square (χ2) tests

[b] More rural = >50% of households in village classified as rural; less rural = <60%

[c] Training months were considered the first two months of operation (June and July 2015)

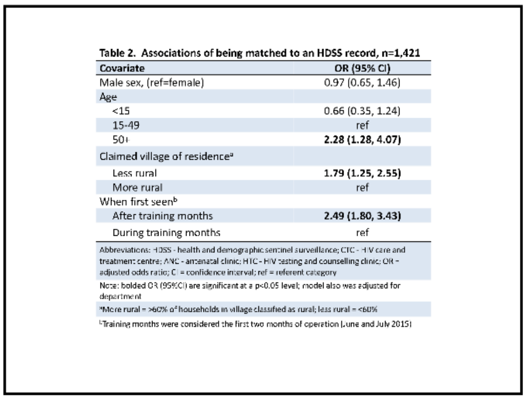


Table 2. Associations of being matched to an HDSS record, n=1,421

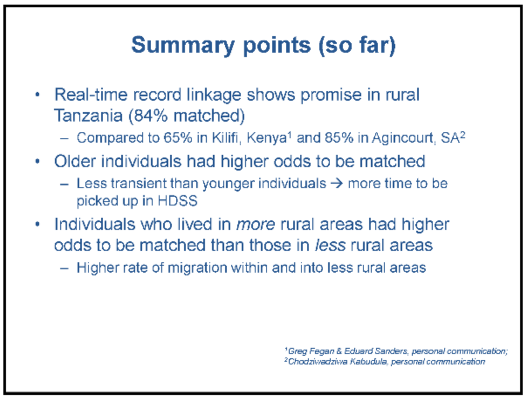
Covariate	OR (95% CI)
Male sex, (ref=female)	0.97 (0.65, 1.46)
Age	
<15	0.66 (0.35, 1.24)
15-49	ref
50+	**2.28 (1.28, 4.07)**
Claimed village of residence[a]	
Less rural	**1.79 (1.25, 2.55)**
More rural	ref
When first seen[b]	
After training months	**2.49 (1.80, 3.43)**
During training months	ref

Abbreviations: HDSS - health and demographic sentinel surveillance; CTC - HIV care and treatment centre; ANC - antenatal clinic; HTC - HIV testing and counselling clinic; OR = adjusted odds ratio; CI = confidence interval; ref = referent category

Note: bolded OR (95%CI) are significant at a p<0.05 level; model also was adjusted for department

[a] More rural = >60% of households in village classified as rural; less rural = <60%

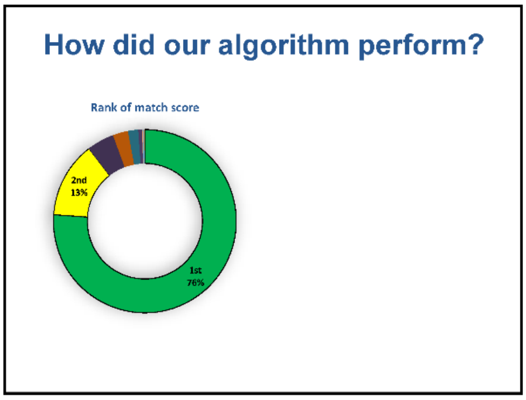
[b] Training months were considered the first two months of operation (June and July 2015)



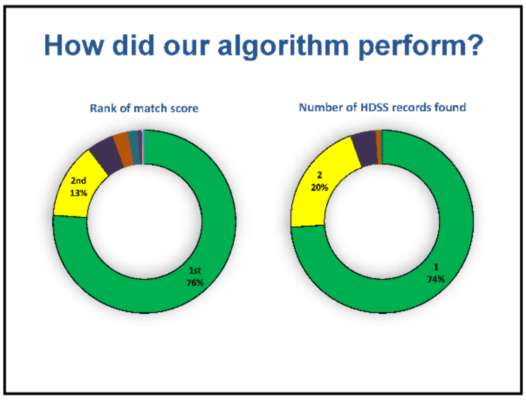
Summary points (so far)

- Real-time record linkage shows promise in rural Tanzania (84% matched)
  - Compared to 65% in Kilifi, Kenya[1] and 85% in Agincourt, SA[2]
- Older individuals had higher odds to be matched
  - Less transient than younger individuals → more time to be picked up in HDSS
- Individuals who lived in *more* rural areas had higher odds to be matched than those in *less* rural areas
  - Higher rate of migration within and into less rural areas

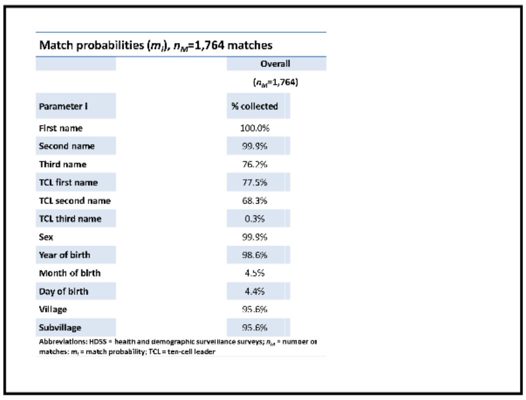
[1] *Greg Fegan & Eduard Sanders, personal communication;*
[2] *Chodziwadziwa Kabudula, personal communication*



How did our algorithm perform?



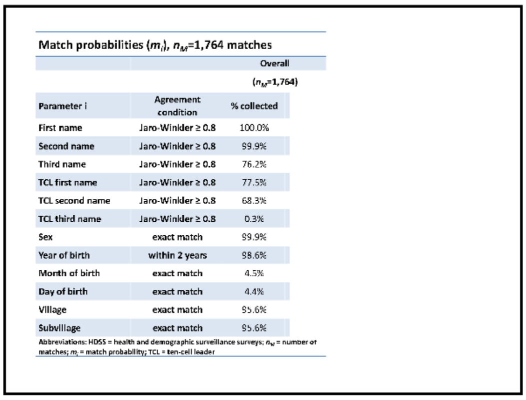
How did our algorithm perform?



Match probabilities (m_i), n_M=1,764 matches

		Overall
		(n_M=1,764)
Parameter i		% collected
First name		100.0%
Second name		99.9%
Third name		76.2%
TCL first name		77.5%
TCL second name		68.3%
TCL third name		0.3%
Sex		99.9%
Year of birth		98.6%
Month of birth		4.5%
Day of birth		4.4%
Village		95.6%
Subvillage		95.6%

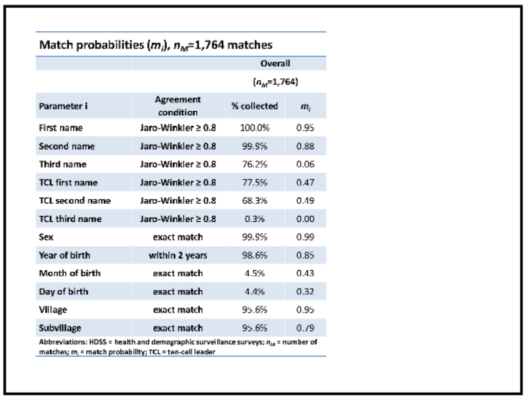
Abbreviations: HDSS = health and demographic surveillance surveys; n_M = number of matches; m_i = match probability; TCL = ten-cell leader



Match probabilities (m_i), n_M=1,764 matches

		Overall
		(n_M=1,764)
Parameter i	Agreement condition	% collected
First name	Jaro-Winkler ≥ 0.8	100.0%
Second name	Jaro-Winkler ≥ 0.8	99.9%
Third name	Jaro-Winkler ≥ 0.8	76.2%
TCL first name	Jaro-Winkler ≥ 0.8	77.5%
TCL second name	Jaro-Winkler ≥ 0.8	68.3%
TCL third name	Jaro-Winkler ≥ 0.8	0.3%
Sex	exact match	99.9%
Year of birth	within 2 years	98.6%
Month of birth	exact match	4.5%
Day of birth	exact match	4.4%
Village	exact match	95.6%
Subvillage	exact match	95.6%

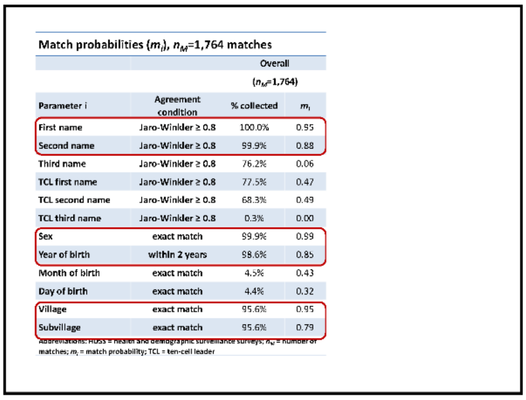
Abbreviations: HDSS = health and demographic surveillance surveys; n_M = number of matches; m_i = match probability; TCL = ten-cell leader



Match probabilities (m_i), n_M=1,764 matches

		Overall	
		(n_M=1,764)	
Parameter i	Agreement condition	% collected	m_i
First name	Jaro-Winkler ≥ 0.8	100.0%	0.95
Second name	Jaro-Winkler ≥ 0.8	99.9%	0.88
Third name	Jaro-Winkler ≥ 0.8	76.2%	0.06
TCL first name	Jaro-Winkler ≥ 0.8	77.5%	0.47
TCL second name	Jaro-Winkler ≥ 0.8	68.3%	0.49
TCL third name	Jaro-Winkler ≥ 0.8	0.3%	0.00
Sex	exact match	99.9%	0.99
Year of birth	within 2 years	98.6%	0.85
Month of birth	exact match	4.5%	0.43
Day of birth	exact match	4.4%	0.32
Village	exact match	95.6%	0.95
Subvillage	exact match	95.6%	0.79

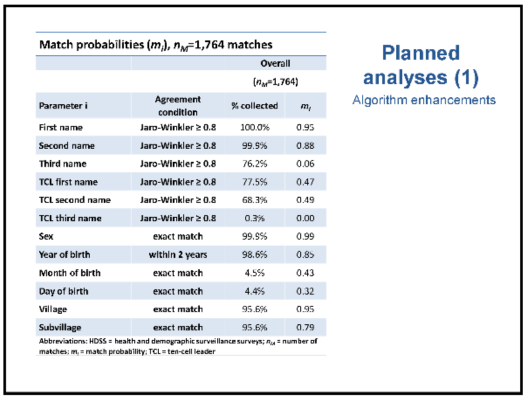
Abbreviations: HDSS = health and demographic surveillance surveys; n_M = number of matches; m_i = match probability; TCL = ten-cell leader



Match probabilities (m_i), n_M=1,764 matches

		Overall	
		(n_M=1,764)	
Parameter i	Agreement condition	% collected	m_i
First name	Jaro-Winkler ≥ 0.8	100.0%	0.95
Second name	Jaro-Winkler ≥ 0.8	99.9%	0.88
Third name	Jaro-Winkler ≥ 0.8	76.2%	0.06
TCL first name	Jaro-Winkler ≥ 0.8	77.5%	0.47
TCL second name	Jaro-Winkler ≥ 0.8	68.3%	0.49
TCL third name	Jaro-Winkler ≥ 0.8	0.3%	0.00
Sex	exact match	99.9%	0.99
Year of birth	within 2 years	98.6%	0.85
Month of birth	exact match	4.5%	0.43
Day of birth	exact match	4.4%	0.32
Village	exact match	95.6%	0.95
Subvillage	exact match	95.6%	0.79

Abbreviations: HDSS = health and demographic surveillance surveys; n_M = number of matches; m_i = match probability; TCL = ten-cell leader



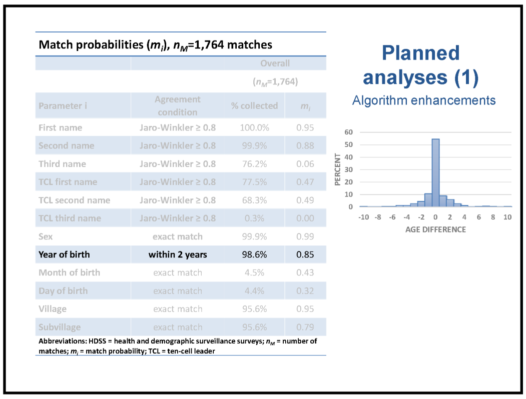
Match probabilities (m_i), n_M=1,764 matches

		Overall	
		(n_M=1,764)	
Parameter i	Agreement condition	% collected	m_i
First name	Jaro-Winkler ≥ 0.8	100.0%	0.95
Second name	Jaro-Winkler ≥ 0.8	99.9%	0.88
Third name	Jaro-Winkler ≥ 0.8	76.2%	0.06
TCL first name	Jaro-Winkler ≥ 0.8	77.5%	0.47
TCL second name	Jaro-Winkler ≥ 0.8	68.3%	0.49
TCL third name	Jaro-Winkler ≥ 0.8	0.3%	0.00
Sex	exact match	99.9%	0.99
Year of birth	within 2 years	98.6%	0.85
Month of birth	exact match	4.5%	0.43
Day of birth	exact match	4.4%	0.32
Village	exact match	95.6%	0.95
Subvillage	exact match	95.6%	0.79

Abbreviations: HDSS = health and demographic surveillance surveys; n_M = number of matches; m_i = match probability; TCL = ten-cell leader

Planned analyses (1)

Algorithm enhancements



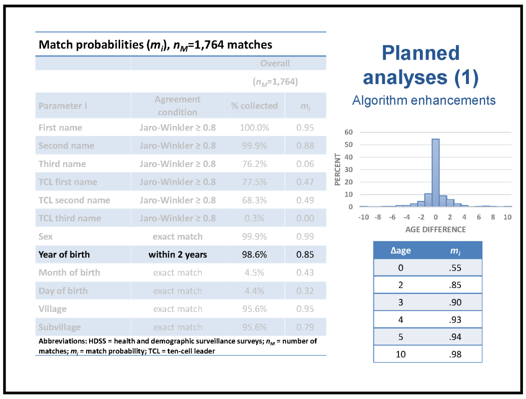
Match probabilities (m_i), n_M=1,764 matches

		Overall	
		(n_M=1,764)	
Parameter i	Agreement condition	% collected	m_i
First name	Jaro-Winkler ≥ 0.8	100.0%	0.95
Second name	Jaro-Winkler ≥ 0.8	99.9%	0.88
Third name	Jaro-Winkler ≥ 0.8	76.2%	0.06
TCL first name	Jaro-Winkler ≥ 0.8	77.5%	0.47
TCL second name	Jaro-Winkler ≥ 0.8	68.3%	0.49
TCL third name	Jaro-Winkler ≥ 0.8	0.3%	0.00
Sex	exact match	99.9%	0.99
Year of birth	**within 2 years**	98.6%	0.85
Month of birth	exact match	4.5%	0.43
Day of birth	exact match	4.4%	0.32
Village	exact match	95.6%	0.95
Subvillage	exact match	95.6%	0.79

Abbreviations: HDSS = health and demographic surveillance surveys; n_M = number of matches; m_i = match probability; TCL = ten-cell leader

Planned analyses (1)

Algorithm enhancements



Match probabilities (m_i), n_M=1,764 matches

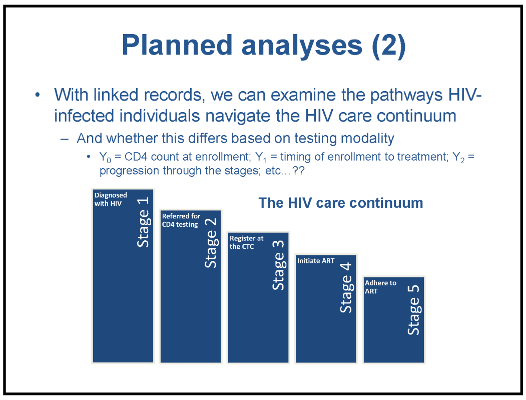
		Overall	
		(n_M=1,764)	
Parameter i	Agreement condition	% collected	m_i
First name	Jaro-Winkler ≥ 0.8	100.0%	0.95
Second name	Jaro-Winkler ≥ 0.8	99.9%	0.88
Third name	Jaro-Winkler ≥ 0.8	76.2%	0.06
TCL first name	Jaro-Winkler ≥ 0.8	77.5%	0.47
TCL second name	Jaro-Winkler ≥ 0.8	68.3%	0.49
TCL third name	Jaro-Winkler ≥ 0.8	0.3%	0.00
Sex	exact match	99.9%	0.99
Year of birth	**within 2 years**	98.6%	0.85
Month of birth	exact match	4.5%	0.43
Day of birth	exact match	4.4%	0.32
Village	exact match	95.6%	0.95
Subvillage	exact match	95.6%	0.79

Abbreviations: HDSS = health and demographic surveillance surveys; n_M = number of matches; m_i = match probability; TCL = ten-cell leader

Planned analyses (1)

Algorithm enhancements

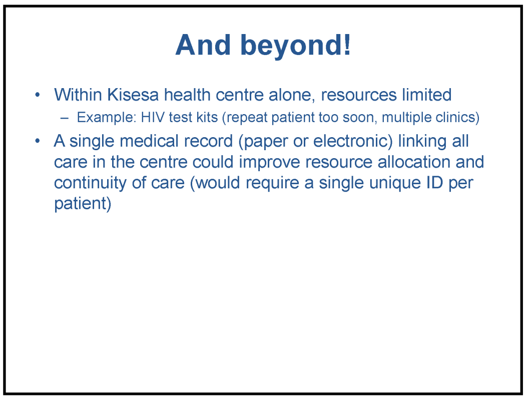
Δage	m_i
0	.55
2	.85
3	.90
4	.93
5	.94
10	.98



Planned analyses (2)

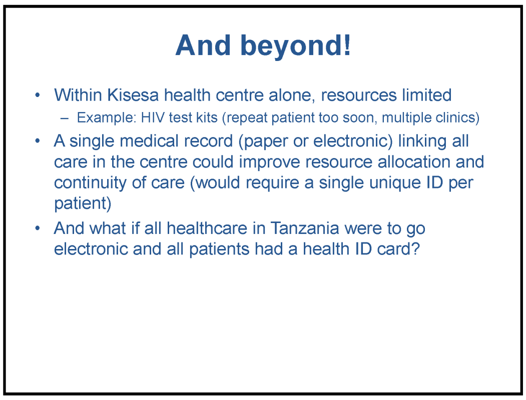
- With linked records, we can examine the pathways HIV-infected individuals navigate the HIV care continuum
  - And whether this differs based on testing modality
    - Y_0 = CD4 count at enrollment; Y_1 = timing of enrollment to treatment; Y_2 = progression through the stages; etc…??

The HIV care continuum



And beyond!

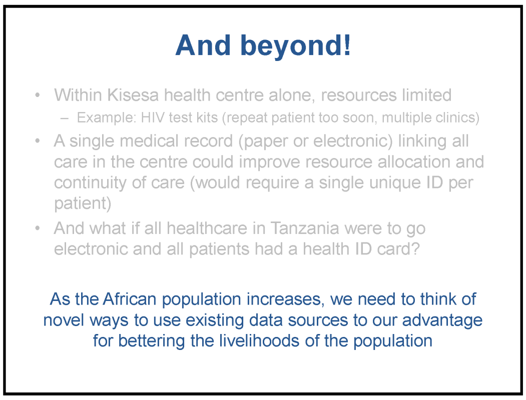
- Within Kisesa health centre alone, resources limited
  - Example: HIV test kits (repeat patient too soon, multiple clinics)
- A single medical record (paper or electronic) linking all care in the centre could improve resource allocation and continuity of care (would require a single unique ID per patient)



And beyond!

- Within Kisesa health centre alone, resources limited
  - Example: HIV test kits (repeat patient too soon, multiple clinics)
- A single medical record (paper or electronic) linking all care in the centre could improve resource allocation and continuity of care (would require a single unique ID per patient)

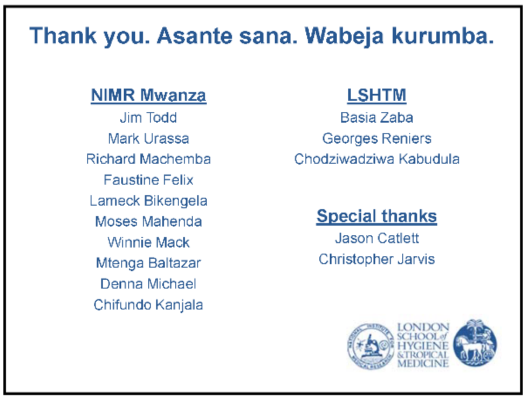
- And what if all healthcare in Tanzania were to go electronic and all patients had a health ID card?



And beyond!

- Within Kisesa health centre alone, resources limited
  - Example: HIV test kits (repeat patient too soon, multiple clinics)
- A single medical record (paper or electronic) linking all care in the centre could improve resource allocation and continuity of care (would require a single unique ID per patient)
- And what if all healthcare in Tanzania were to go electronic and all patients had a health ID card?

As the African population increases, we need to think of novel ways to use existing data sources to our advantage for bettering the livelihoods of the population



Thank you. Asante sana. Wabeja kurumba.

NIMR Mwanza

Jim Todd
Mark Urassa
Richard Machemba
Faustine Felix
Lameck Bikengela
Moses Mahenda
Winnie Mack
Mtenga Baltazar
Denna Michael
Chifundo Kanjala

LSHTM

Basia Zaba
Georges Reniers
Chodziwadziwa Kabudula

Special thanks

Jason Catlett
Christopher Jarvis

LONDON SCHOOL of HYGIENE &TROPICAL MEDICINE

Thank you. Asante sana. Wabeja kurumba.

The Kisesa team

Questions or comments?

Christopher.Rentsch@lshtm.ac.uk

LONDON SCHOOL of HYGIENE &TROPICAL MEDICINE

10.1.2 LSHTM Research Degree Poster Day, 2016

Using record linkage to improve engagement in HIV care

Rentsch C[1], Reniers G[1], Urassa M[2], Todd J[1], Marston M[1], Zaba B[1]

[1] Department of Population Health, London School of Hygiene & Tropical Medicine
[2] The Tazama Project, National Institute for Medical Research, Mwanza, Tanzania

Introduction

- Uptake of HIV care services remains low in east and southern Africa.[1]
- Whether and how communities engage with HIV services, from diagnosis to treatment, is a key consideration in meeting the 90-90-90 target.[2]
- Most analyses of these services are limited to patients in care and lack a population perspective. In contrast, health and demographic surveillance systems (HDSS) rely on self reports of health service use.
- Linking clinic and community data would allow for better monitoring and improving engagement in HIV care.

Methods

- Real-time record linkage was initiated in June 2015 in three clinics serving the Kisesa HDSS, in NW Tanzania:
  - the HIV care and treatment centre (CTC),
  - the HIV testing and counselling clinic (HTC), and
  - the antenatal clinic (ANC), in which prevention of mother-to-child transmission (PMTCT) services are offered
- As patients arrive at the clinics, fieldworkers use specially designed computer software to search the HDSS database for potential matches (Fig 1), consulting the patient about each potential match to locate the true match(es).

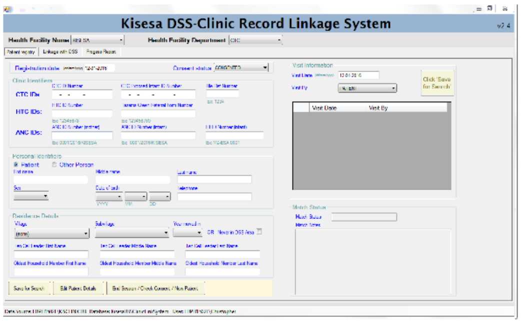


Fig. 1 Screenshot of the real-time record linkage software

- The software invokes a probabilistic algorithm based on an original model by Newcombe et al[3] and formalized by Felligi and Sunter.[4] The algorithm includes up to twelve parameters, each with their own agreement condition (Table 1). Names are compared using the Jaro-Winkler string comparator.[5]

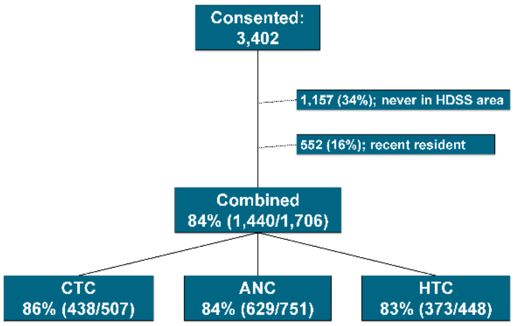
Parameter	Agreement condition
Three names for patient	Jaro-Winkler ≥ 0.8
Three names for ten-cell leader	Jaro-Winkler ≥ 0.8
Sex	exact match
Year of birth	within 2 years
Month and day of birth	exact match
Village and subvillage	exact match

Table 1 Parameters and agreement conditions in algorithm

Preliminary results

Nearly 40 patients are interviewed each day, half of whom are newly consented patients with the other half being repeat visitors.

As of 31 December 2015, 84% of individuals with residence history in the HDSS area were matched to at least one record (Fig 2).


Consented: 3,402

1,157 (34%); never in HDSS area

552 (16%); recent resident

Combined 84% (1,440/1,706)

CTC 86% (438/507)

ANC 84% (629/751)

HTC 83% (373/448)

Fig. 2 Sample size and match percentage through 31 December 2015

The algorithm performs well (with some room to improve) in ranking potential matches (Fig 3), and in its ability to locate multiple residency episodes (Fig 4).

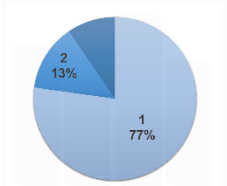


Fig. 3 Ranking of records selected as a true match

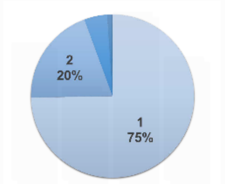


Fig. 4 Number of HDSS records matched to each individual

Planned work

Using the linked data, I plan to:

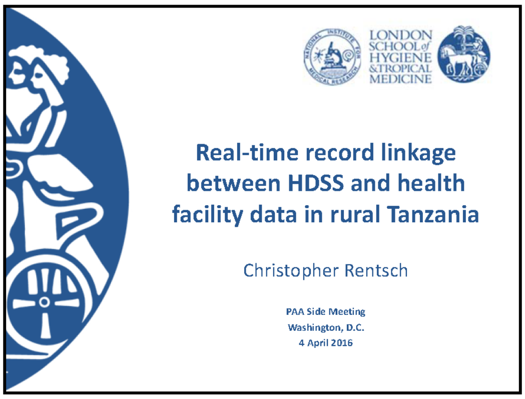
- evaluate the real-time record linkage system in this setting,
- analyse the characteristics of users and non-users of HIV services, examine the pathways they used to navigate the HIV care continuum, and quantify delays and sub-optimal service use, and
- model the errors associated with probabilistic record linkage and use this to estimate error bounds on analyses that cannot benefit from real-time linkage methods.

References

1. Church, K., et al. (2015). "A comparative analysis of national HIV policies in six African countries with generalized epidemics." Bulletin of the World Health Organization 93(7): 457-467.
2. UNAIDS (2014). 90-90-90: an ambitious treatment target to help end the AIDS epidemic.
3. Newcombe H, Kennedy J, Axford S, James A. Automatic Linkage of Vital Records. Science. 1959; 130(3381): 954-959
4. Fellegi IP, Sunter AB. A Theory for Record Linkage. J Am Stat Assoc. 1969; 64(328): 1183-1210
5. Winkler WE. String Comparator Metrics and Enhanced Decision Rules in the Fellegi-Sunter Model of Record Linkage. 1990.

Improving health worldwide   www.lshtm.ac.uk

10.1.3 Invited talk by IUSSP/PAA at The World Bank, 2016

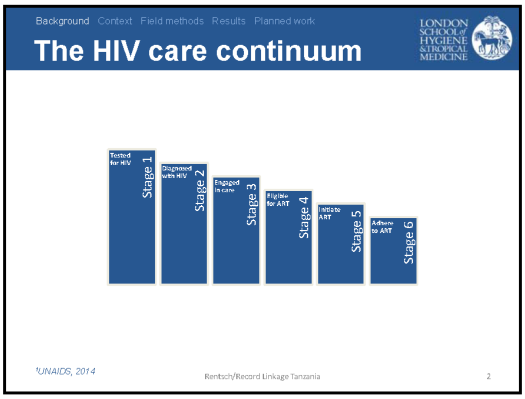


LONDON SCHOOL of HYGIENE &TROPICAL MEDICINE

Real-time record linkage between HDSS and health facility data in rural Tanzania

Christopher Rentsch

PAA Side Meeting
Washington, D.C.
4 April 2016

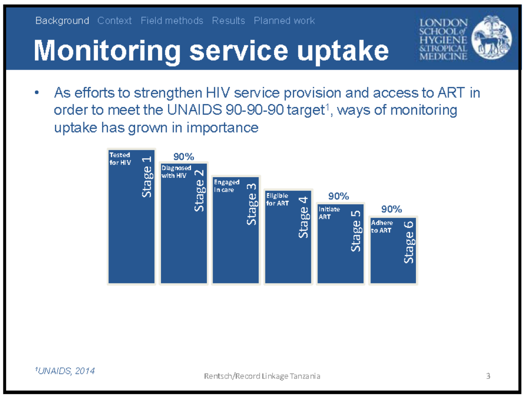


Background Context Field methods Results Planned work

The HIV care continuum

[1] UNAIDS, 2014

Rentsch/Record Linkage Tanzania 2



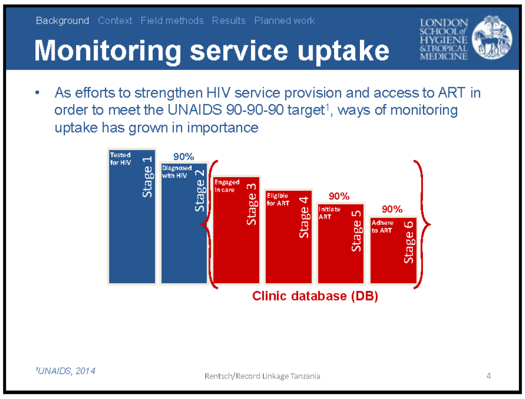
Background Context Field methods Results Planned work

Monitoring service uptake

- As efforts to strengthen HIV service provision and access to ART in order to meet the UNAIDS 90-90-90 target[1], ways of monitoring uptake has grown in importance

[1] UNAIDS, 2014

Rentsch/Record Linkage Tanzania 3



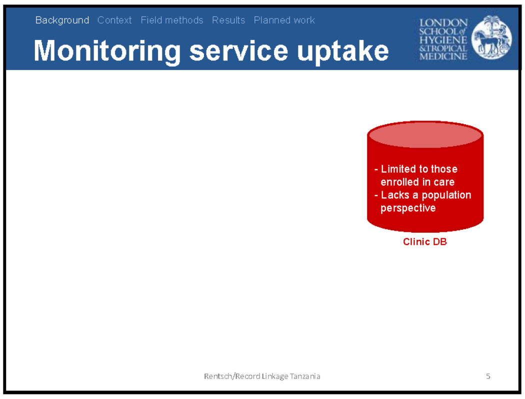
Background Context Field methods Results Planned work

Monitoring service uptake

- As efforts to strengthen HIV service provision and access to ART in order to meet the UNAIDS 90-90-90 target[1], ways of monitoring uptake has grown in importance

[1] UNAIDS, 2014

Rentsch/Record Linkage Tanzania 4



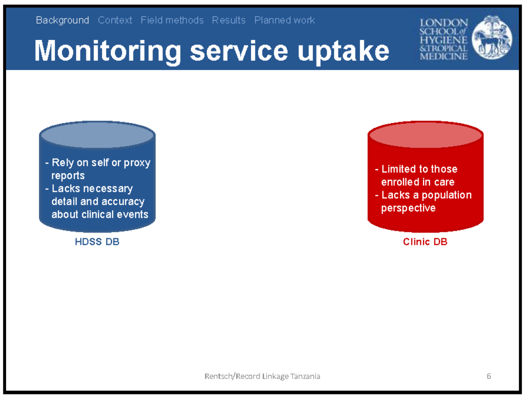
Background Context Field methods Results Planned work

Monitoring service uptake

- Limited to those enrolled in care
- Lacks a population perspective

Clinic DB

Rentsch/Record Linkage Tanzania 5



Background Context Field methods Results Planned work

Monitoring service uptake

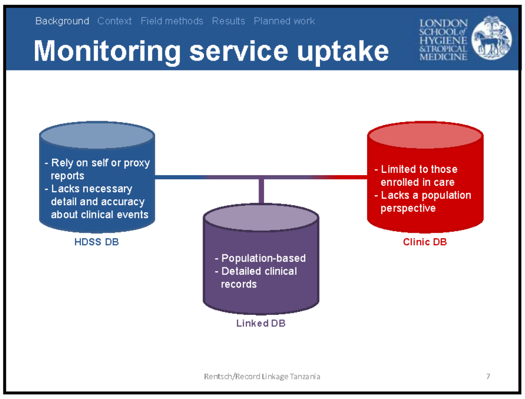
- Rely on self or proxy reports
- Lacks necessary detail and accuracy about clinical events

HDSS DB

- Limited to those enrolled in care
- Lacks a population perspective

Clinic DB

Rentsch/Record Linkage Tanzania 6



Background Context Field methods Results Planned work

Monitoring service uptake

- Rely on self or proxy reports
- Lacks necessary detail and accuracy about clinical events

HDSS DB

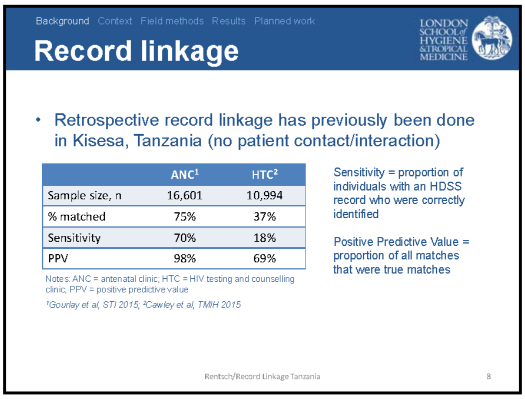
- Limited to those enrolled in care
- Lacks a population perspective

Clinic DB

- Population-based
- Detailed clinical records

Linked DB

Rentsch/Record Linkage Tanzania 7



Background Context Field methods Results Planned work

Record linkage

- Retrospective record linkage has previously been done in Kisesa, Tanzania (no patient contact/interaction)

	ANC[1]	HTC[2]
Sample size, n	16,601	10,994
% matched	75%	37%
Sensitivity	70%	18%
PPV	98%	69%

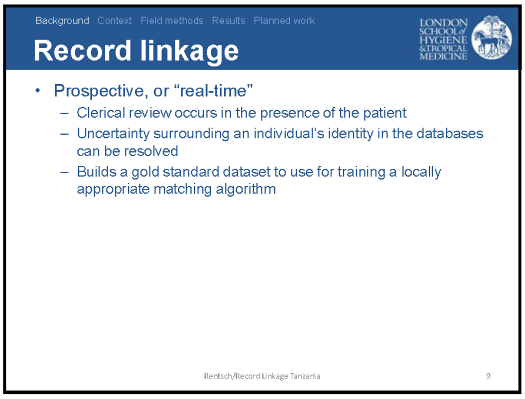
Notes: ANC = antenatal clinic; HTC = HIV testing and counselling clinic; PPV = positive predictive value

[1]Gourlay et al, STI 2015; [2]Cawley et al, TMIH 2015

Sensitivity = proportion of individuals with an HDSS record who were correctly identified

Positive Predictive Value = proportion of all matches that were true matches

Rentsch/Record Linkage Tanzania 8

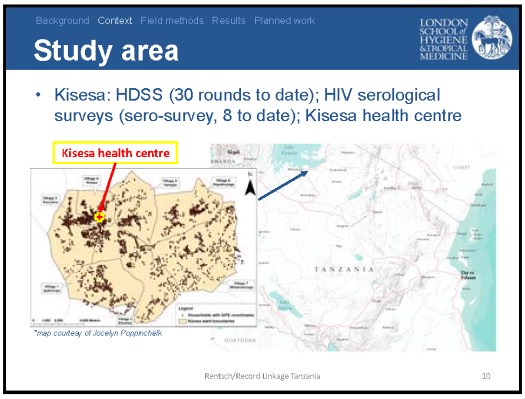


Background Context Field methods Results Planned work

Record linkage

- Prospective, or “real-time”
  - Clerical review occurs in the presence of the patient
  - Uncertainty surrounding an individual’s identity in the databases can be resolved
  - Builds a gold standard dataset to use for training a locally appropriate matching algorithm

Rentsch/Record Linkage Tanzania 9



Background Context Field methods Results Planned work

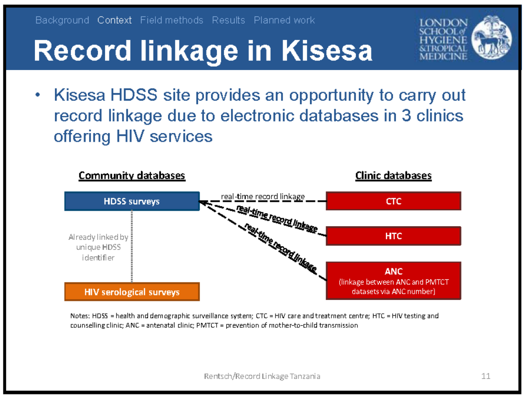
Study area

- Kisesa: HDSS (30 rounds to date); HIV serological surveys (sero-survey, 8 to date); Kisesa health centre

Kisesa health centre

*map courtesy of Jocelyn Poppinchalk

Rentsch/Record Linkage Tanzania 10



Background Context Field methods Results Planned work

Record linkage in Kisesa

- Kisesa HDSS site provides an opportunity to carry out record linkage due to electronic databases in 3 clinics offering HIV services

Community databases

HDSS surveys

Already linked by unique HDSS identifier

HIV serological surveys

Clinic databases

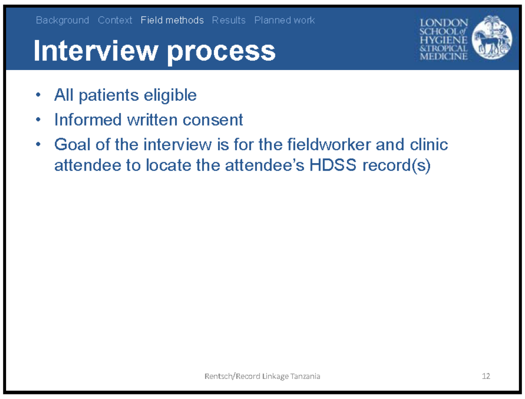
CTC

HTC

ANC (linkage between ANC and PMTCT datasets via ANC number)

Notes: HDSS = health and demographic surveillance system; CTC = HIV care and treatment centre; HTC = HIV testing and counselling clinic; ANC = antenatal clinic; PMTCT = prevention of mother-to-child transmission

Rentsch/Record Linkage Tanzania 11



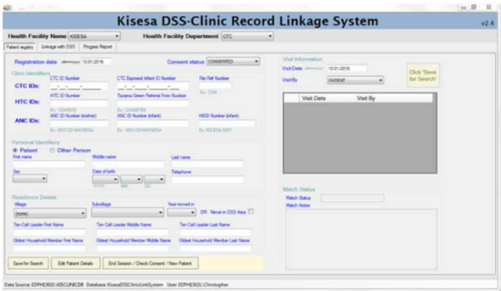
Background Context Field methods Results Planned work

Interview process

- All patients eligible
- Informed written consent
- Goal of the interview is for the fieldworker and clinic attendee to locate the attendee’s HDSS record(s)

Rentsch/Record Linkage Tanzania 12

Software


Kisesa DSS-Clinic Record Linkage System v2.4

Health Facility Name: KISESA

Health Facility Department: CTC

Registration date: 12-01-2016

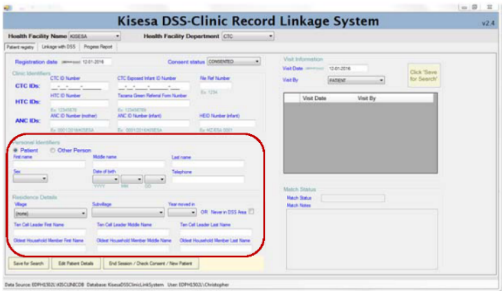
Consent status: CONSENTED

Visit Date: 12-01-2016

Visit By: PATIENT

Data Source: EDPHI302L\KISCLINBCDB Database: KisesaDSSClinicLinkSystem User: EDPHI302L\Christopher

Personal identifiers


Kisesa DSS-Clinic Record Linkage System v2.4

Health Facility Name: KISESA

Health Facility Department: CTC

Registration date: 12-01-2016

Consent status: CONSENTED

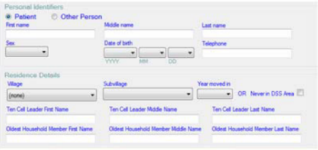
Visit Date: 12-01-2016

Visit By: PATIENT

Data Source: EDPHI302L\KISCLINBCDB Database: KisesaDSSClinicLinkSystem User: EDPHI302L\Christopher

Personal Identifiers
Patient
Other Person
First name
Middle name
Last name
Sex
Date of birth
Telephone
Residence Details
Village
Subvillage
Year moved in
OR Never in DSS Area
Ten Cell Leader First Name
Ten Cell Leader Middle Name
Ten Cell Leader Last Name
Oldest Household Member First Name
Oldest Household Member Middle Name
Oldest Household Member Last Name
Save for Search
Edit Patient Details
End Session / Check Consent / New Patient
Match Status
Match Notes

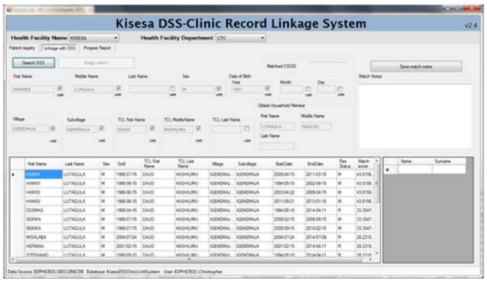
Personal identifiers


Village: (none)

Personal Identifiers
Patient
Other Person
First name
Middle name
Last name
Sex
Date of birth
Telephone
YYYY
MM
DD
Residence Details
Village
Subvillage
Year moved in
OR Never in DSS Area
Ten Cell Leader First Name
Ten Cell Leader Middle Name
Ten Cell Leader Last Name
Oldest Household Member First Name
Oldest Household Member Middle Name
Oldest Household Member Last Name

Parameter	Agreement condition
Three names for patient	Jaro-Winkler ≥0.8
Three names for ten-cell leader	Jaro-Winkler ≥0.8
Sex	exact
Year of birth	±2 years
Month and day of birth	exact
Village and subvillage	exact

Software demonstration

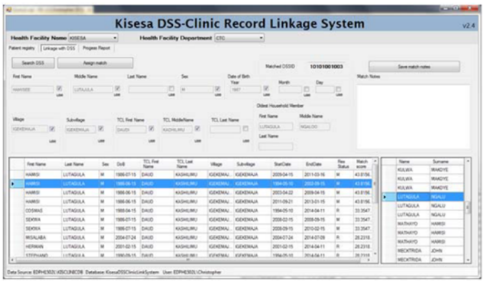

Kisesa DSS-Clinic Record Linkage System v2.4

Health Facility Name: KISESA

Health Facility Department: CTC

Data Source: EDPHI302L\KISCLINBCDB Database: KisesaDSSClinicLinkSystem User: EDPHI302L\Christopher

Software demonstration


Kisesa DSS-Clinic Record Linkage System v2.4

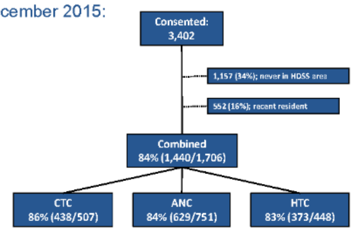
Health Facility Name: KISESA

Health Facility Department: CTC

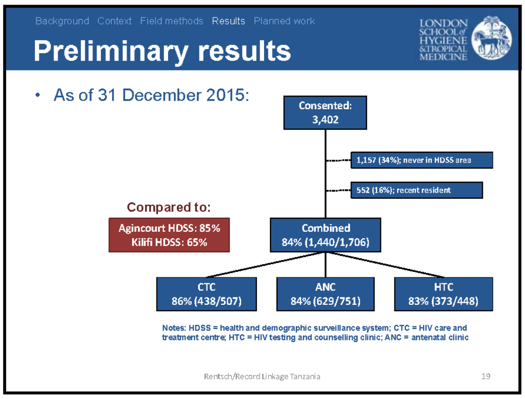
Data Source: EDPHI302L\KISCLINBCDB Database: KisesaDSSClinicLinkSystem User: EDPHI302L\Christopher

Preliminary results

- As of 31 December 2015:


Consented: 3,402
1,157 (34%); never in HDSS area
552 (16%); recent resident
Combined 84% (1,440/1,706)
CTC 86% (438/507)
ANC 84% (629/751)
HTC 83% (373/448)

Notes: HDSS = health and demographic surveillance system; CTC = HIV care and treatment centre; HTC = HIV testing and counselling clinic; ANC = antenatal clinic



Background Context Field methods Results Planned work

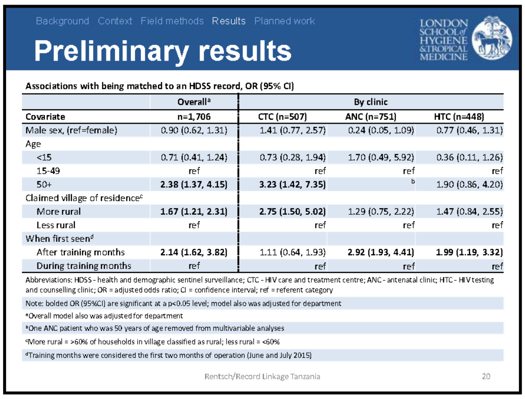
LONDON SCHOOL of HYGIENE &TROPICAL MEDICINE

Preliminary results

- As of 31 December 2015:

Notes: HDSS = health and demographic surveillance system; CTC = HIV care and treatment centre; HTC = HIV testing and counselling clinic; ANC = antenatal clinic

Rentsch/Record Linkage Tanzania 19



Background Context Field methods Results Planned work

LONDON SCHOOL of HYGIENE &TROPICAL MEDICINE

Preliminary results

Associations with being matched to an HDSS record, OR (95% CI)

	Overall[a]	By clinic		
Covariate	**n=1,706**	**CTC (n=507)**	**ANC (n=751)**	**HTC (n=448)**
Male sex, (ref=female)	0.90 (0.62, 1.31)	1.41 (0.77, 2.57)	0.24 (0.05, 1.09)	0.77 (0.46, 1.31)
Age				
<15	0.71 (0.41, 1.24)	0.73 (0.28, 1.94)	1.70 (0.49, 5.92)	0.36 (0.11, 1.26)
15-49	ref	ref	ref	ref
50+	**2.38 (1.37, 4.15)**	**3.23 (1.42, 7.35)**	[b]	1.90 (0.86, 4.20)
Claimed village of residence[c]				
More rural	**1.67 (1.21, 2.31)**	**2.75 (1.50, 5.02)**	1.29 (0.75, 2.22)	1.47 (0.84, 2.55)
Less rural	ref	ref	ref	ref
When first seen[d]				
After training months	**2.14 (1.62, 3.82)**	1.11 (0.64, 1.93)	**2.92 (1.93, 4.41)**	**1.99 (1.19, 3.32)**
During training months	ref	ref	ref	ref

Abbreviations: HDSS - health and demographic sentinel surveillance; CTC - HIV care and treatment centre; ANC - antenatal clinic; HTC - HIV testing and counselling clinic; OR = adjusted odds ratio; CI = confidence interval; ref = referent category

Note: bolded OR (95%CI) are significant at a $p<0.05$ level; model also was adjusted for department

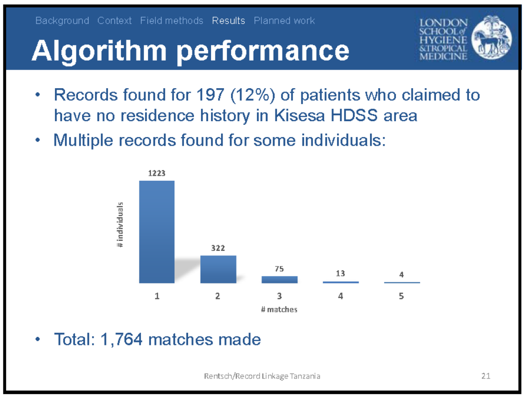
[a] Overall model also was adjusted for department

[b] One ANC patient who was 50 years of age removed from multivariable analyses

[c] More rural = >60% of households in village classified as rural; less rural = <60%

[d] Training months were considered the first two months of operation (June and July 2015)

Rentsch/Record Linkage Tanzania 20



Background Context Field methods Results Planned work

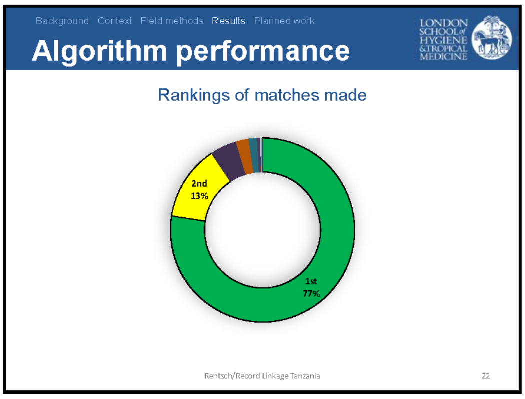
LONDON SCHOOL of HYGIENE &TROPICAL MEDICINE

Algorithm performance

- Records found for 197 (12%) of patients who claimed to have no residence history in Kisesa HDSS area
- Multiple records found for some individuals:

- Total: 1,764 matches made

Rentsch/Record Linkage Tanzania 21



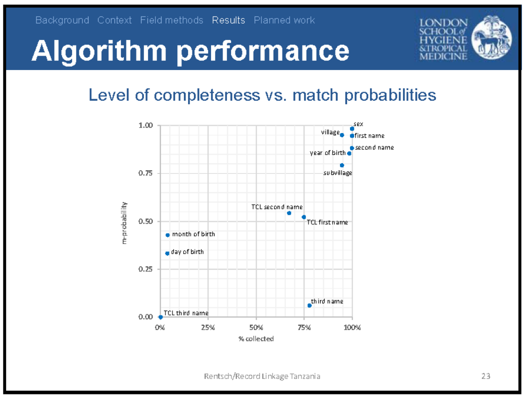
Background Context Field methods Results Planned work

LONDON SCHOOL of HYGIENE &TROPICAL MEDICINE

Algorithm performance

Rankings of matches made

Rentsch/Record Linkage Tanzania 22



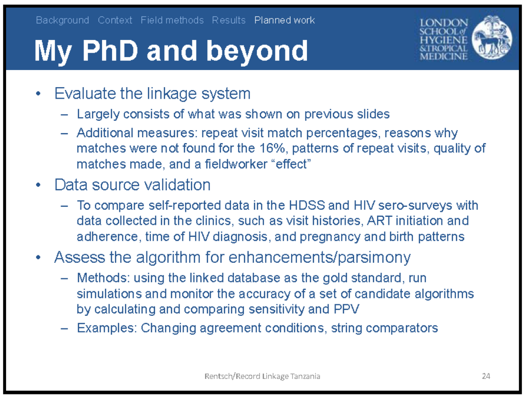
Background Context Field methods Results Planned work

LONDON SCHOOL of HYGIENE &TROPICAL MEDICINE

Algorithm performance

Level of completeness vs. match probabilities

Rentsch/Record Linkage Tanzania 23



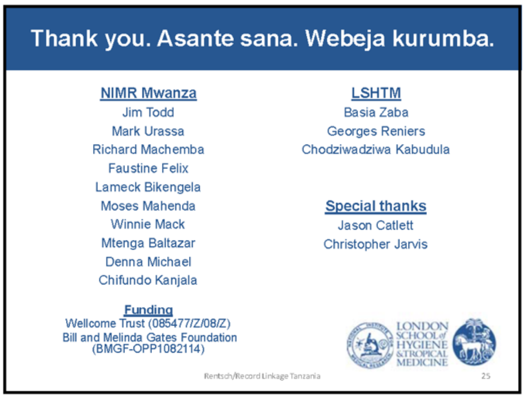
Background Context Field methods Results Planned work

LONDON SCHOOL of HYGIENE &TROPICAL MEDICINE

My PhD and beyond

- Evaluate the linkage system
  - Largely consists of what was shown on previous slides
  - Additional measures: repeat visit match percentages, reasons why matches were not found for the 16%, patterns of repeat visits, quality of matches made, and a fieldworker "effect"
- Data source validation
  - To compare self-reported data in the HDSS and HIV sero-surveys with data collected in the clinics, such as visit histories, ART initiation and adherence, time of HIV diagnosis, and pregnancy and birth patterns
- Assess the algorithm for enhancements/parsimony
  - Methods: using the linked database as the gold standard, run simulations and monitor the accuracy of a set of candidate algorithms by calculating and comparing sensitivity and PPV
  - Examples: Changing agreement conditions, string comparators

Rentsch/Record Linkage Tanzania 24



Thank you. Asante sana. Webeja kurumba.

NIMR Mwanza
Jim Todd
Mark Urassa
Richard Machemba
Faustine Felix
Lameck Bikengela
Moses Mahenda
Winnie Mack
Mtenga Baltazar
Denna Michael
Chifundo Kanjala

LSHTM
Basia Zaba
Georges Reniers
Chodziwadziwa Kabudula

Special thanks
Jason Catlett
Christopher Jarvis

Funding
Wellcome Trust (085477/Z/08/Z)
Bill and Melinda Gates Foundation
(BMGF-OPP1082114)

LONDON SCHOOL of HYGIENE & TROPICAL MEDICINE

Rentsch/Record Linkage Tanzania 25

Thank you. Asante sana. Webeja kurumba.

The Kisesa team

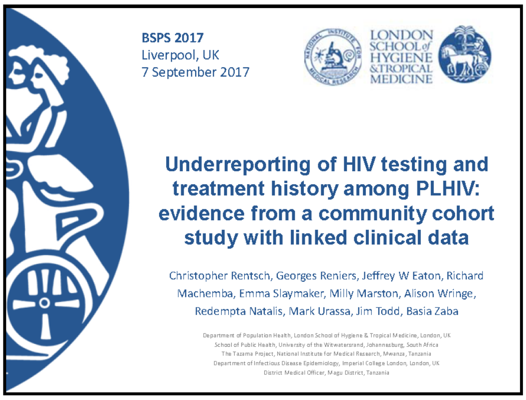
Questions or comments?

Christopher.Rentsch@lshtm.ac.uk

LONDON SCHOOL of HYGIENE & TROPICAL MEDICINE

Rentsch/Record Linkage Tanzania 26

10.1.4 British Society for Population Studies (BSPS), 2017

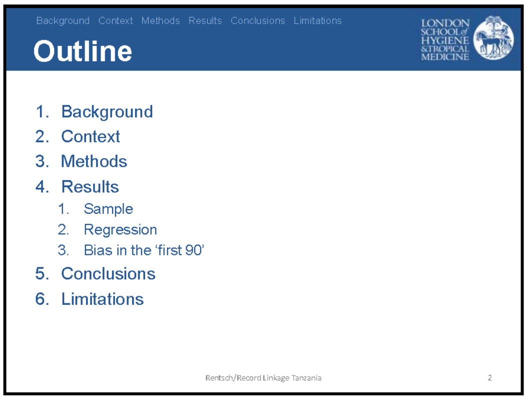


BSPS 2017
Liverpool, UK
7 September 2017

Underreporting of HIV testing and treatment history among PLHIV: evidence from a community cohort study with linked clinical data

Christopher Rentsch, Georges Reniers, Jeffrey W Eaton, Richard Machemba, Emma Slaymaker, Milly Marston, Alison Wringe, Redempta Natalis, Mark Urassa, Jim Todd, Basia Zaba

Department of Population Health, London School of Hygiene & Tropical Medicine, London, UK
School of Public Health, University of the Witwatersrand, Johannesburg, South Africa
The Tazama Project, National Institute for Medical Research, Mwanza, Tanzania
Department of Infectious Disease Epidemiology, Imperial College London, London, UK
District Medical Officer, Magu District, Tanzania

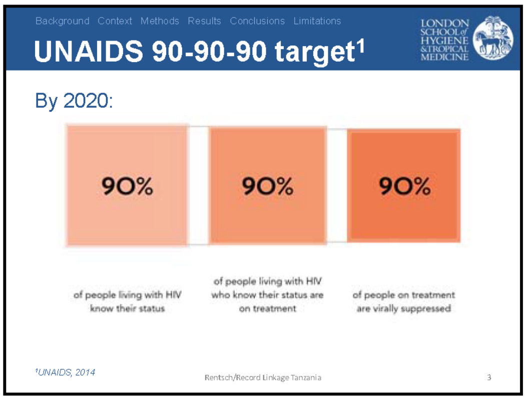


Background Context Methods Results Conclusions Limitations

Outline

1. Background
2. Context
3. Methods
4. Results
   1. Sample
   2. Regression
   3. Bias in the 'first 90'
5. Conclusions
6. Limitations

Rentsch/Record Linkage Tanzania 2



Background Context Methods Results Conclusions Limitations

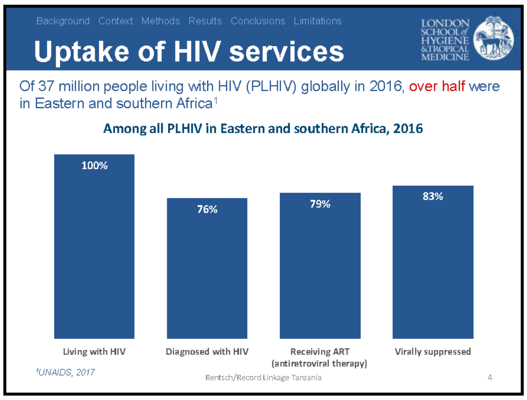
UNAIDS 90-90-90 target[1]

By 2020:

90%
90%
90%
of people living with HIV know their status
of people living with HIV who know their status are on treatment
of people on treatment are virally suppressed

[1]UNAIDS, 2014

Rentsch/Record Linkage Tanzania 3



Background Context Methods Results Conclusions Limitations

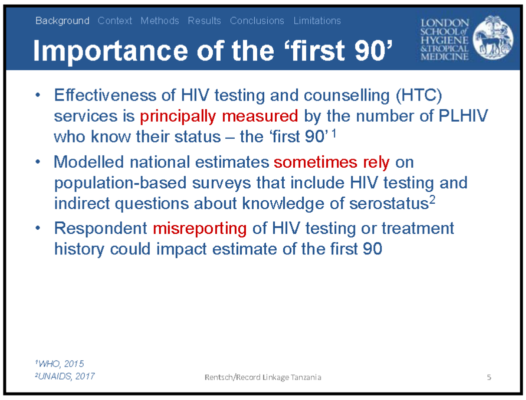
Uptake of HIV services

Of 37 million people living with HIV (PLHIV) globally in 2016, over half were in Eastern and southern Africa[1]

Among all PLHIV in Eastern and southern Africa, 2016

[1]UNAIDS, 2017

Rentsch/Record Linkage Tanzania 4



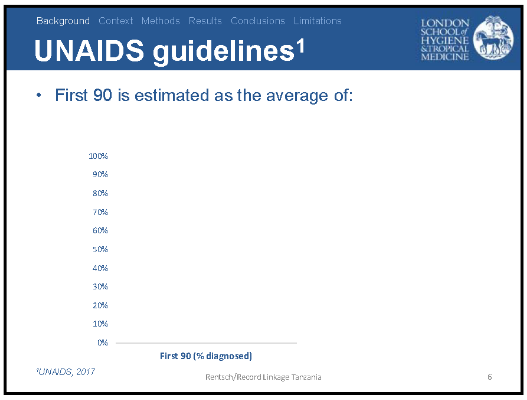
Background Context Methods Results Conclusions Limitations

Importance of the 'first 90'

- Effectiveness of HIV testing and counselling (HTC) services is principally measured by the number of PLHIV who know their status – the 'first 90' [1]
- Modelled national estimates sometimes rely on population-based surveys that include HIV testing and indirect questions about knowledge of serostatus[2]
- Respondent misreporting of HIV testing or treatment history could impact estimate of the first 90

[1]WHO, 2015
[2]UNAIDS, 2017

Rentsch/Record Linkage Tanzania 5



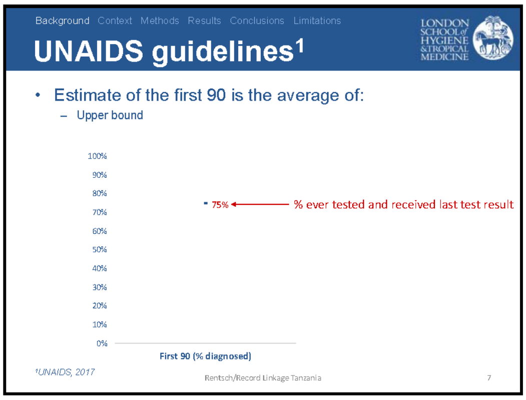
Background Context Methods Results Conclusions Limitations

UNAIDS guidelines[1]

- First 90 is estimated as the average of:

[1]UNAIDS, 2017

Rentsch/Record Linkage Tanzania 6



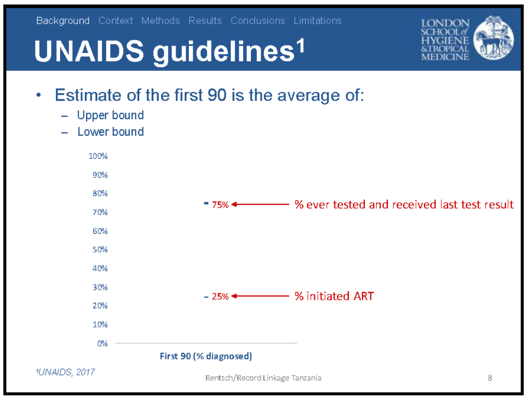
UNAIDS guidelines[1]

- Estimate of the first 90 is the average of:
  - Upper bound

% ever tested and received last test result

[1] UNAIDS, 2017

Rentsch/Record Linkage Tanzania 7



UNAIDS guidelines[1]

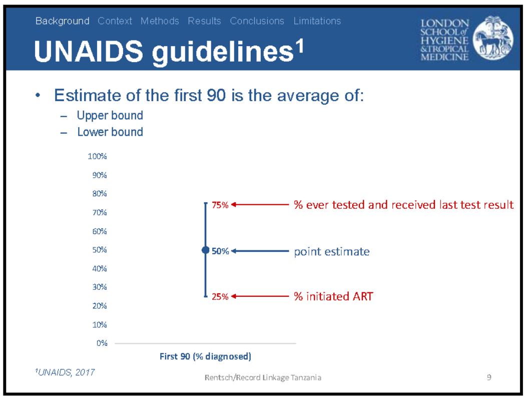
- Estimate of the first 90 is the average of:
  - Upper bound
  - Lower bound

% ever tested and received last test result

% initiated ART

[1] UNAIDS, 2017

Rentsch/Record Linkage Tanzania 8



UNAIDS guidelines[1]

- Estimate of the first 90 is the average of:
  - Upper bound
  - Lower bound

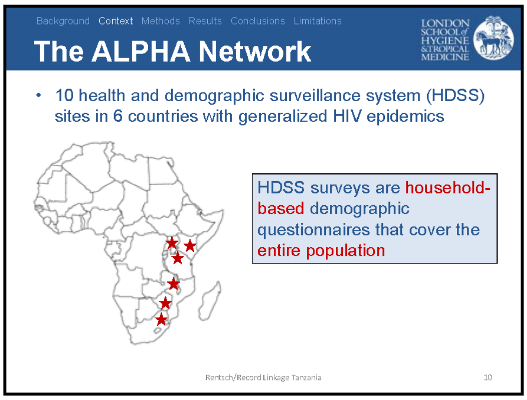
% ever tested and received last test result

point estimate

% initiated ART

[1] UNAIDS, 2017

Rentsch/Record Linkage Tanzania 9



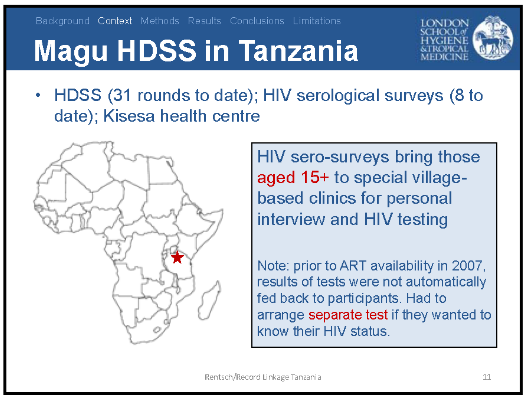
Background Context Methods Results Conclusions Limitations
LONDON SCHOOL of HYGIENE &TROPICAL MEDICINE

The ALPHA Network

- 10 health and demographic surveillance system (HDSS) sites in 6 countries with generalized HIV epidemics

HDSS surveys are household-based demographic questionnaires that cover the entire population

Rentsch/Record Linkage Tanzania 10



Background Context Methods Results Conclusions Limitations
LONDON SCHOOL of HYGIENE &TROPICAL MEDICINE

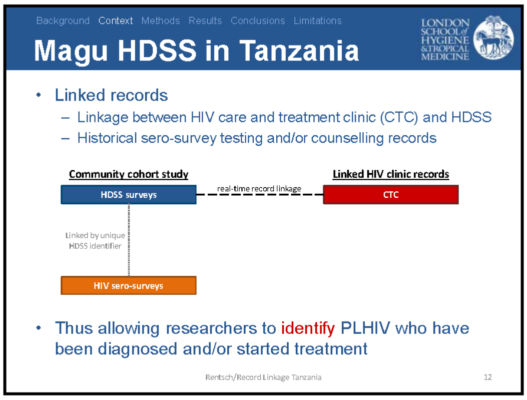
Magu HDSS in Tanzania

- HDSS (31 rounds to date); HIV serological surveys (8 to date); Kisesa health centre

HIV sero-surveys bring those aged 15+ to special village-based clinics for personal interview and HIV testing

Note: prior to ART availability in 2007, results of tests were not automatically fed back to participants. Had to arrange separate test if they wanted to know their HIV status.

Rentsch/Record Linkage Tanzania 11

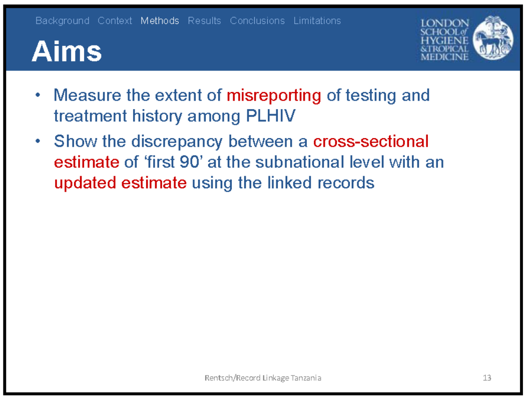


Background Context Methods Results Conclusions Limitations
LONDON SCHOOL of HYGIENE &TROPICAL MEDICINE
Community cohort study
Linked HIV clinic records
HDSS surveys
real-time record linkage
CTC
Linked by unique HDSS identifier
HIV sero-surveys

Magu HDSS in Tanzania

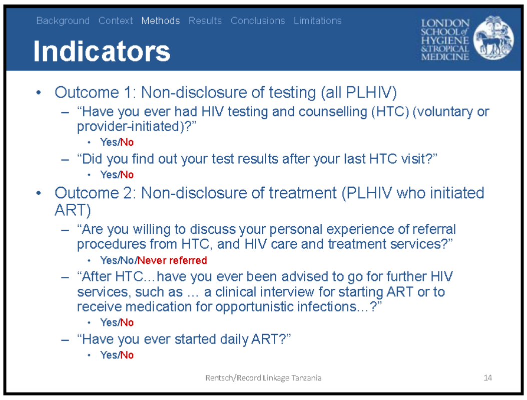
- Linked records
  - Linkage between HIV care and treatment clinic (CTC) and HDSS
  - Historical sero-survey testing and/or counselling records
- Thus allowing researchers to identify PLHIV who have been diagnosed and/or started treatment

Rentsch/Record Linkage Tanzania 12



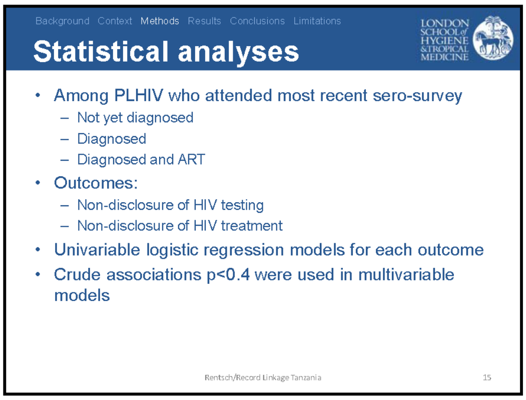
Aims

- Measure the extent of misreporting of testing and treatment history among PLHIV
- Show the discrepancy between a cross-sectional estimate of 'first 90' at the subnational level with an updated estimate using the linked records



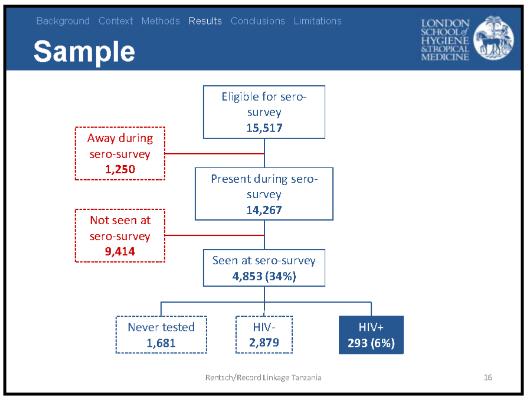
Indicators

- Outcome 1: Non-disclosure of testing (all PLHIV)
  - "Have you ever had HIV testing and counselling (HTC) (voluntary or provider-initiated)?"
    - Yes/No
  - "Did you find out your test results after your last HTC visit?"
    - Yes/No
- Outcome 2: Non-disclosure of treatment (PLHIV who initiated ART)
  - "Are you willing to discuss your personal experience of referral procedures from HTC, and HIV care and treatment services?"
    - Yes/No/Never referred
  - "After HTC…have you ever been advised to go for further HIV services, such as … a clinical interview for starting ART or to receive medication for opportunistic infections…?"
    - Yes/No
  - "Have you ever started daily ART?"
    - Yes/No



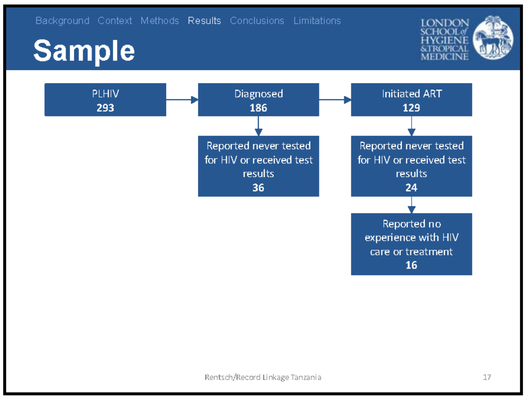
Statistical analyses

- Among PLHIV who attended most recent sero-survey
  - Not yet diagnosed
  - Diagnosed
  - Diagnosed and ART
- Outcomes:
  - Non-disclosure of HIV testing
  - Non-disclosure of HIV treatment
- Univariable logistic regression models for each outcome
- Crude associations p<0.4 were used in multivariable models



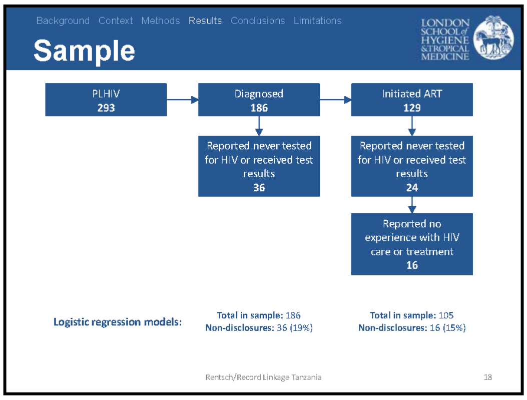
Sample

Eligible for sero-survey 15,517
Away during sero-survey 1,250
Present during sero-survey 14,267
Not seen at sero-survey 9,414
Seen at sero-survey 4,853 (34%)
Never tested 1,681
HIV- 2,879
HIV+ 293 (6%)



Sample

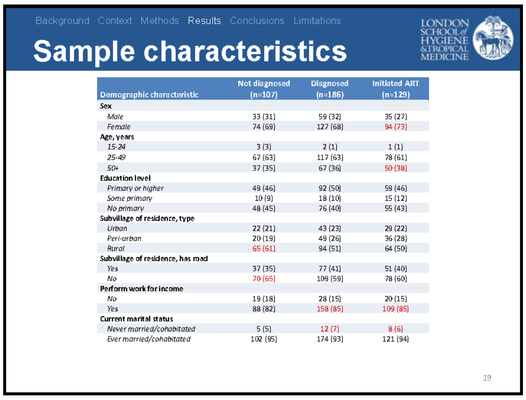
PLHIV 293
Diagnosed 186
Initiated ART 129
Reported never tested for HIV or received test results 36
Reported never tested for HIV or received test results 24
Reported no experience with HIV care or treatment 16



Sample

PLHIV 293
Diagnosed 186
Initiated ART 129
Reported never tested for HIV or received test results 36
Reported never tested for HIV or received test results 24
Reported no experience with HIV care or treatment 16

Logistic regression models: **Total in sample:** 186 **Non-disclosures:** 36 (19%) **Total in sample:** 105 **Non-disclosures:** 16 (15%)

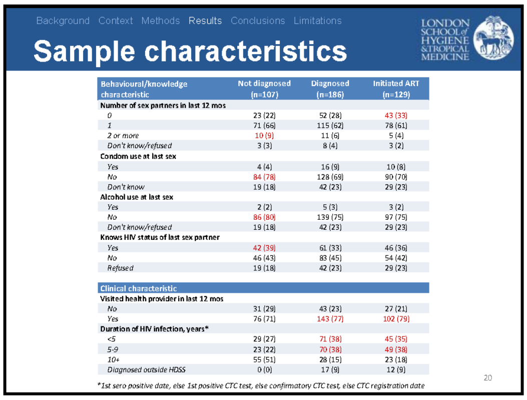


Background Context Methods Results Conclusions Limitations

Sample characteristics

Demographic characteristic	Not diagnosed (n=107)	Diagnosed (n=186)	Initiated ART (n=129)
Sex			
Male	33 (31)	59 (32)	35 (27)
Female	74 (69)	127 (68)	94 (73)
Age, years			
15-24	3 (3)	2 (1)	1 (1)
25-49	67 (63)	117 (63)	78 (61)
50+	37 (35)	67 (36)	50 (38)
Education level			
Primary or higher	49 (46)	92 (50)	59 (46)
Some primary	10 (9)	18 (10)	15 (12)
No primary	48 (45)	76 (40)	55 (43)
Subvillage of residence, type			
Urban	22 (21)	43 (23)	29 (22)
Peri-urban	20 (19)	49 (26)	36 (28)
Rural	65 (61)	94 (51)	64 (50)
Subvillage of residence, has road			
Yes	37 (35)	77 (41)	51 (40)
No	70 (65)	109 (59)	78 (60)
Perform work for income			
No	19 (18)	28 (15)	20 (15)
Yes	88 (82)	158 (85)	109 (85)
Current marital status			
Never married/cohabitated	5 (5)	12 (7)	8 (6)
Ever married/cohabitated	102 (95)	174 (93)	121 (94)

19



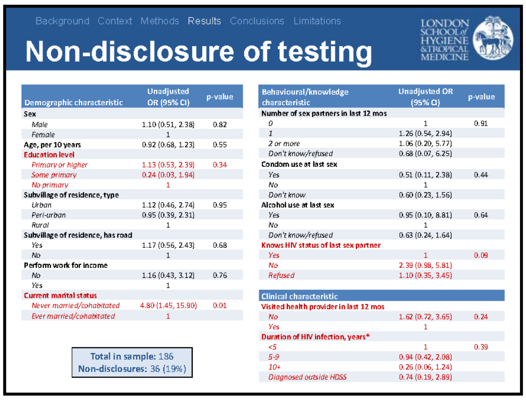
Background Context Methods Results Conclusions Limitations

Sample characteristics

Behavioural/knowledge characteristic	Not diagnosed (n=107)	Diagnosed (n=186)	Initiated ART (n=129)
Number of sex partners in last 12 mos			
0	23 (22)	52 (28)	43 (33)
1	71 (66)	115 (62)	78 (61)
2 or more	10 (9)	11 (6)	5 (4)
Don't know/refused	3 (3)	8 (4)	3 (2)
Condom use at last sex			
Yes	4 (4)	16 (9)	10 (8)
No	84 (78)	128 (69)	90 (70)
Don't know	19 (18)	42 (23)	29 (23)
Alcohol use at last sex			
Yes	2 (2)	5 (3)	3 (2)
No	86 (80)	139 (75)	97 (75)
Don't know/refused	19 (18)	42 (23)	29 (23)
Knows HIV status of last sex partner			
Yes	42 (39)	61 (33)	46 (36)
No	46 (43)	83 (45)	54 (42)
Refused	19 (18)	42 (23)	29 (23)
Clinical characteristic			
Visited health provider in last 12 mos			
No	31 (29)	43 (23)	27 (21)
Yes	76 (71)	143 (77)	102 (79)
Duration of HIV infection, years*			
<5	29 (27)	71 (38)	45 (35)
5-9	23 (22)	70 (38)	49 (38)
10+	55 (51)	28 (15)	23 (18)
Diagnosed outside HDSS	0 (0)	17 (9)	12 (9)

20

*1st sero positive date, else 1st positive CTC test, else confirmatory CTC test, else CTC registration date



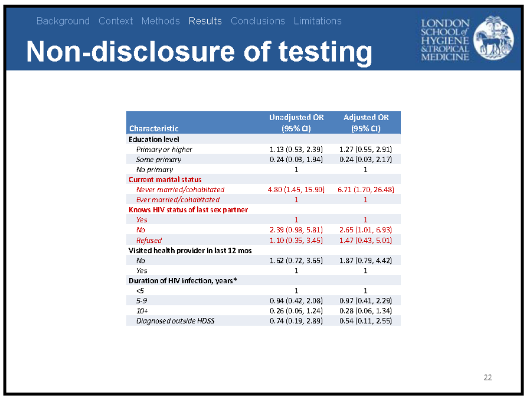
Background Context Methods Results Conclusions Limitations

Non-disclosure of testing

Demographic characteristic	Unadjusted OR (95% CI)	p-value
Sex		
Male	1.10 (0.51, 2.38)	0.82
Female	1	
Age, per 10 years	0.92 (0.68, 1.23)	0.55
Education level		
Primary or higher	1.13 (0.53, 2.39)	0.34
Some primary	0.24 (0.03, 1.94)	
No primary	1	
Subvillage of residence, type		
Urban	1.12 (0.46, 2.74)	0.95
Peri-urban	0.95 (0.39, 2.31)	
Rural	1	
Subvillage of residence, has road		
Yes	1.17 (0.56, 2.43)	0.68
No	1	
Perform work for income		
No	1.16 (0.43, 3.12)	0.76
Yes	1	
Current marital status		
Never married/cohabitated	4.80 (1.45, 15.90)	0.01
Ever married/cohabitated	1	

Total in sample: 186
Non-disclosures: 36 (19%)

Behavioural/knowledge characteristic	Unadjusted OR (95% CI)	p-value
Number of sex partners in last 12 mos		
0	1	0.91
1	1.26 (0.54, 2.94)	
2 or more	1.06 (0.20, 5.77)	
Don't know/refused	0.68 (0.07, 6.25)	
Condom use at last sex		
Yes	0.51 (0.11, 2.38)	0.44
No	1	
Don't know	0.60 (0.23, 1.56)	
Alcohol use at last sex		
Yes	0.95 (0.10, 8.81)	0.64
No	1	
Don't know/refused	0.63 (0.24, 1.64)	
Knows HIV status of last sex partner		
Yes	1	0.09
No	2.39 (0.98, 5.81)	
Refused	1.10 (0.35, 3.45)	
Clinical characteristic		
Visited health provider in last 12 mos		
No	1.62 (0.72, 3.65)	0.24
Yes	1	
Duration of HIV infection, years*		
<5	1	0.39
5-9	0.94 (0.42, 2.08)	
10+	0.26 (0.06, 1.24)	
Diagnosed outside HDSS	0.74 (0.19, 2.89)	

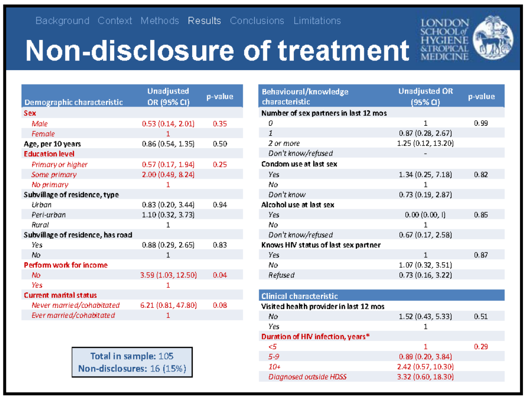


Background Context Methods Results Conclusions Limitations

Non-disclosure of testing

Characteristic	Unadjusted OR (95% CI)	Adjusted OR (95% CI)
Education level		
Primary or higher	1.13 (0.53, 2.39)	1.27 (0.55, 2.91)
Some primary	0.24 (0.03, 1.94)	0.24 (0.03, 2.17)
No primary	1	1
Current marital status		
Never married/cohabitated	4.80 (1.45, 15.90)	6.71 (1.70, 26.48)
Ever married/cohabitated	1	1
Knows HIV status of last sex partner		
Yes	1	1
No	2.39 (0.98, 5.81)	2.65 (1.01, 6.93)
Refused	1.10 (0.35, 3.45)	1.47 (0.43, 5.01)
Visited health provider in last 12 mos		
No	1.62 (0.72, 3.65)	1.87 (0.79, 4.42)
Yes	1	1
Duration of HIV infection, years*		
<5	1	1
5-9	0.94 (0.42, 2.08)	0.97 (0.41, 2.29)
10+	0.26 (0.06, 1.24)	0.28 (0.06, 1.34)
Diagnosed outside HDSS	0.74 (0.19, 2.89)	0.54 (0.11, 2.55)

22



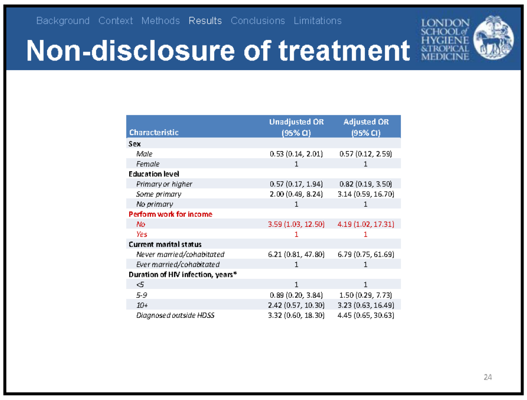
Background Context Methods Results Conclusions Limitations

Non-disclosure of treatment

Demographic characteristic	Unadjusted OR (95% CI)	p-value
Sex		
Male	0.53 (0.14, 2.01)	0.35
Female	1	
Age, per 10 years	0.86 (0.54, 1.35)	0.50
Education level		
Primary or higher	0.57 (0.17, 1.94)	0.25
Some primary	2.00 (0.49, 8.24)	
No primary	1	
Subvillage of residence, type		
Urban	0.83 (0.20, 3.44)	0.94
Peri-urban	1.10 (0.32, 3.73)	
Rural	1	
Subvillage of residence, has road		
Yes	0.88 (0.29, 2.65)	0.83
No	1	
Perform work for income		
No	3.59 (1.03, 12.50)	0.04
Yes	1	
Current marital status		
Never married/cohabitated	6.21 (0.81, 47.80)	0.08
Ever married/cohabitated	1	

Total in sample: 105
Non-disclosures: 16 (15%)

Behavioural/knowledge characteristic	Unadjusted OR (95% CI)	p-value
Number of sex partners in last 12 mos		
0	1	0.99
1	0.87 (0.28, 2.67)	
2 or more	1.25 (0.12, 13.20)	
Don't know/refused	-	
Condom use at last sex		
Yes	1.34 (0.25, 7.18)	0.82
No	1	
Don't know	0.73 (0.19, 2.87)	
Alcohol use at last sex		
Yes	0.00 (0.00, I)	0.85
No	1	
Don't know/refused	0.67 (0.17, 2.58)	
Knows HIV status of last sex partner		
Yes	1	0.87
No	1.07 (0.32, 3.51)	
Refused	0.73 (0.16, 3.22)	
Clinical characteristic		
Visited health provider in last 12 mos		
No	1.52 (0.43, 5.33)	0.51
Yes	1	
Duration of HIV infection, years*		
<5	1	0.29
5-9	0.89 (0.20, 3.84)	
10+	2.42 (0.57, 10.30)	
Diagnosed outside HDSS	3.32 (0.60, 18.30)	

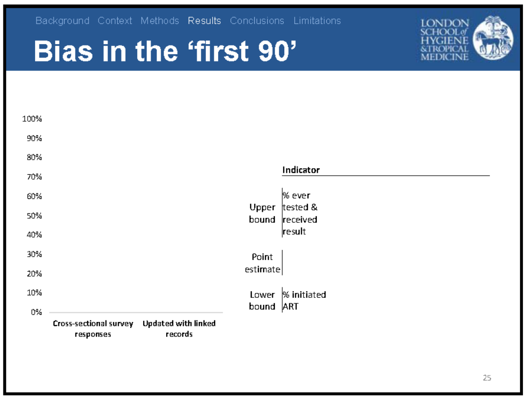


Background Context Methods Results Conclusions Limitations

Non-disclosure of treatment

Characteristic	Unadjusted OR (95% CI)	Adjusted OR (95% CI)
Sex		
Male	0.53 (0.14, 2.01)	0.57 (0.12, 2.59)
Female	1	1
Education level		
Primary or higher	0.57 (0.17, 1.94)	0.82 (0.19, 3.50)
Some primary	2.00 (0.49, 8.24)	3.14 (0.59, 16.70)
No primary	1	1
Perform work for income		
No	3.59 (1.03, 12.50)	4.19 (1.02, 17.31)
Yes	1	1
Current marital status		
Never married/cohabitated	6.21 (0.81, 47.80)	6.79 (0.75, 61.69)
Ever married/cohabitated	1	1
Duration of HIV infection, years*		
<5	1	1
5-9	0.89 (0.20, 3.84)	1.50 (0.29, 7.73)
10+	2.42 (0.57, 10.30)	3.23 (0.63, 16.49)
Diagnosed outside HDSS	3.32 (0.60, 18.30)	4.45 (0.65, 30.63)

24

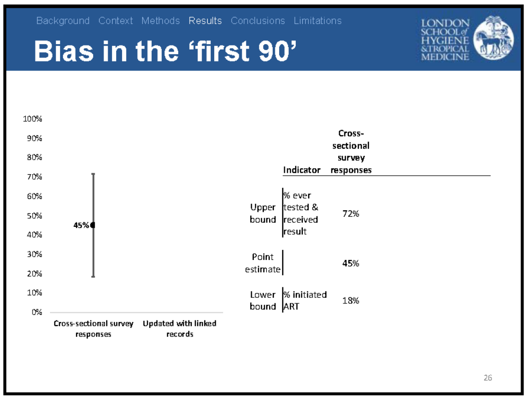


Background Context Methods Results Conclusions Limitations

Bias in the ‘first 90’

	Indicator
Upper bound	% ever tested & received result
Point estimate	
Lower bound	% initiated ART

25

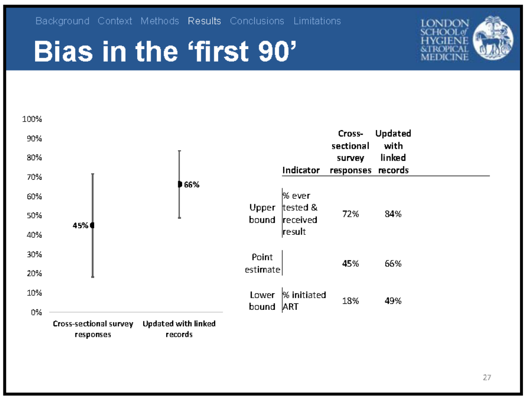


Background Context Methods Results Conclusions Limitations

Bias in the ‘first 90’

	Indicator	Cross-sectional survey responses
Upper bound	% ever tested & received result	72%
Point estimate		45%
Lower bound	% initiated ART	18%

26

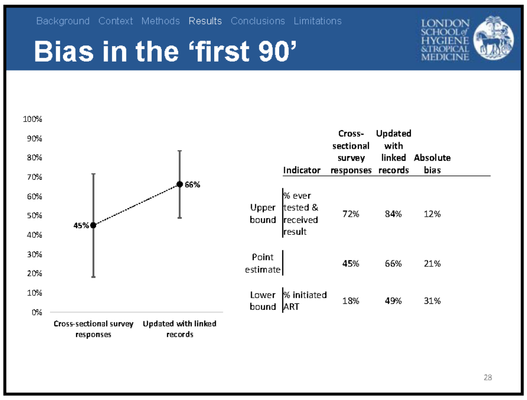


Background Context Methods Results Conclusions Limitations

Bias in the ‘first 90’

	Indicator	Cross-sectional survey responses	Updated with linked records
Upper bound	% ever tested & received result	72%	84%
Point estimate		45%	66%
Lower bound	% initiated ART	18%	49%

27

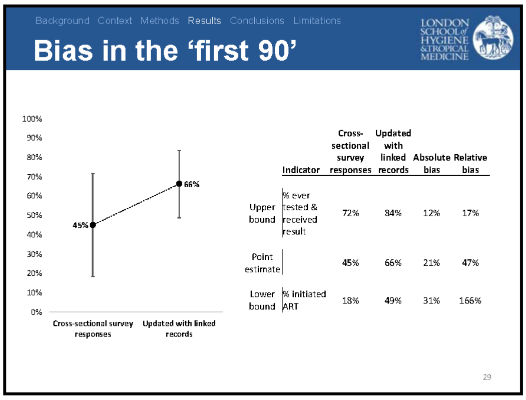


Background Context Methods Results Conclusions Limitations

Bias in the ‘first 90’

	Indicator	Cross-sectional survey responses	Updated with linked records	Absolute bias
Upper bound	% ever tested & received result	72%	84%	12%
Point estimate		45%	66%	21%
Lower bound	% initiated ART	18%	49%	31%

28

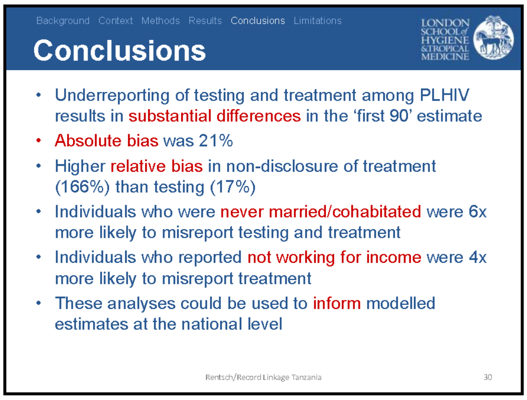


Background Context Methods Results Conclusions Limitations

Bias in the ‘first 90’

	Indicator	Cross-sectional survey responses	Updated with linked records	Absolute bias	Relative bias
Upper bound	% ever tested & received result	72%	84%	12%	17%
Point estimate		45%	66%	21%	47%
Lower bound	% initiated ART	18%	49%	31%	166%

29



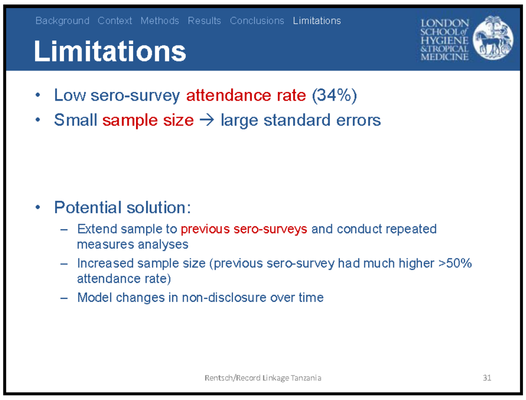
Background Context Methods Results Conclusions Limitations

Conclusions

- Underreporting of testing and treatment among PLHIV results in substantial differences in the ‘first 90’ estimate
- Absolute bias was 21%
- Higher relative bias in non-disclosure of treatment (166%) than testing (17%)
- Individuals who were never married/cohabitated were 6x more likely to misreport testing and treatment
- Individuals who reported not working for income were 4x more likely to misreport treatment
- These analyses could be used to inform modelled estimates at the national level

Rentsch/Record Linkage Tanzania

30



Background Context Methods Results Conclusions Limitations

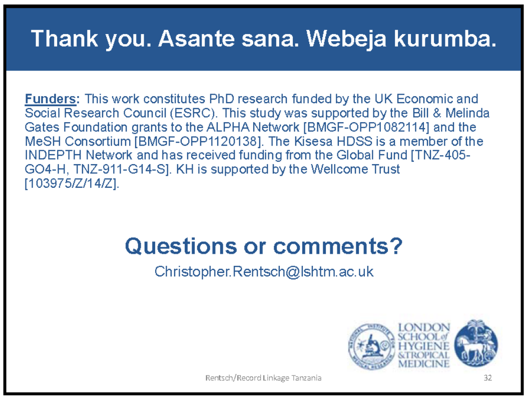
LONDON SCHOOL of HYGIENE &TROPICAL MEDICINE

Limitations

- Low sero-survey attendance rate (34%)
- Small sample size → large standard errors

- Potential solution:
  - Extend sample to previous sero-surveys and conduct repeated measures analyses
  - Increased sample size (previous sero-survey had much higher >50% attendance rate)
  - Model changes in non-disclosure over time

Rentsch/Record Linkage Tanzania 31



Thank you. Asante sana. Webeja kurumba.

Funders: This work constitutes PhD research funded by the UK Economic and Social Research Council (ESRC). This study was supported by the Bill & Melinda Gates Foundation grants to the ALPHA Network [BMGF-OPP1082114] and the MeSH Consortium [BMGF-OPP1120138]. The Kisesa HDSS is a member of the INDEPTH Network and has received funding from the Global Fund [TNZ-405-GO4-H, TNZ-911-G14-S]. KH is supported by the Wellcome Trust [103975/Z/14/Z].

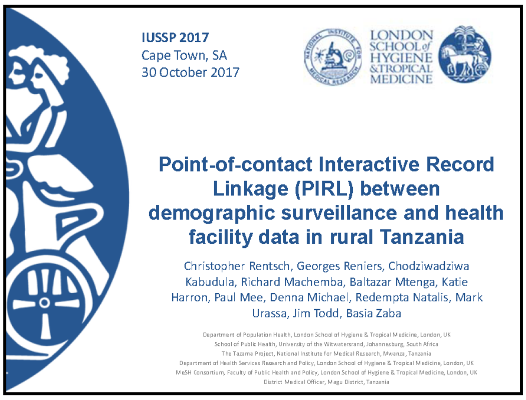
Questions or comments?

Christopher.Rentsch@lshtm.ac.uk

LONDON SCHOOL of HYGIENE &TROPICAL MEDICINE

Rentsch/Record Linkage Tanzania 32

10.1.5 International Union for the Scientific Study of Population (IUSSP), 2017



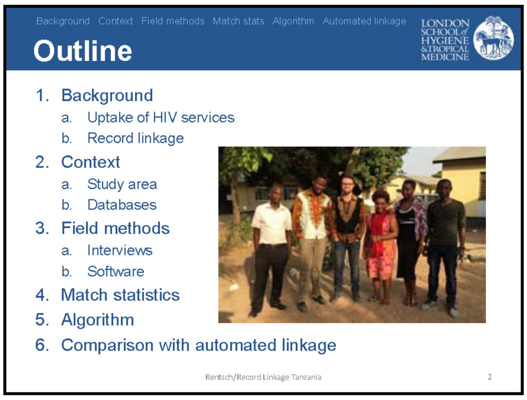
IUSSP 2017
Cape Town, SA
30 October 2017

LONDON SCHOOL of HYGIENE & TROPICAL MEDICINE

Point-of-contact Interactive Record Linkage (PIRL) between demographic surveillance and health facility data in rural Tanzania

Christopher Rentsch, Georges Reniers, Chodziwadziwa Kabudula, Richard Machemba, Baltazar Mtenga, Katie Harron, Paul Mee, Denna Michael, Redempta Natalis, Mark Urassa, Jim Todd, Basia Zaba

Department of Population Health, London School of Hygiene & Tropical Medicine, London, UK
School of Public Health, University of the Witwatersrand, Johannesburg, South Africa
The Tazama Project, National Institute for Medical Research, Mwanza, Tanzania
Department of Health Services Research and Policy, London School of Hygiene & Tropical Medicine, London, UK
MeSH Consortium, Faculty of Public Health and Policy, London School of Hygiene & Tropical Medicine, London, UK
District Medical Officer, Magu District, Tanzania

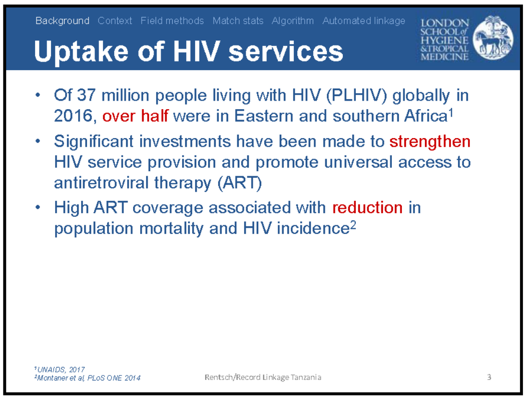


Background Context Field methods Match stats Algorithm Automated linkage

Outline

1. Background
   a. Uptake of HIV services
   b. Record linkage
2. Context
   a. Study area
   b. Databases
3. Field methods
   a. Interviews
   b. Software
4. Match statistics
5. Algorithm
6. Comparison with automated linkage

Rentsch/Record Linkage Tanzania 2



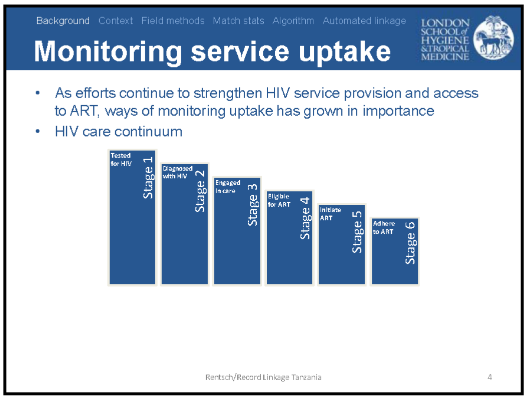
Background Context Field methods Match stats Algorithm Automated linkage

Uptake of HIV services

- Of 37 million people living with HIV (PLHIV) globally in 2016, over half were in Eastern and southern Africa[1]
- Significant investments have been made to strengthen HIV service provision and promote universal access to antiretroviral therapy (ART)
- High ART coverage associated with reduction in population mortality and HIV incidence[2]

[1] *UNAIDS, 2017*
[2] *Montaner et al, PLoS ONE 2014*

Rentsch/Record Linkage Tanzania 3

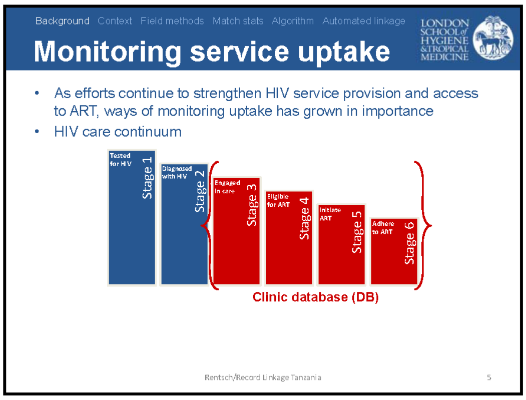


Background Context Field methods Match stats Algorithm Automated linkage

Monitoring service uptake

- As efforts continue to strengthen HIV service provision and access to ART, ways of monitoring uptake has grown in importance
- HIV care continuum

Rentsch/Record Linkage Tanzania 4

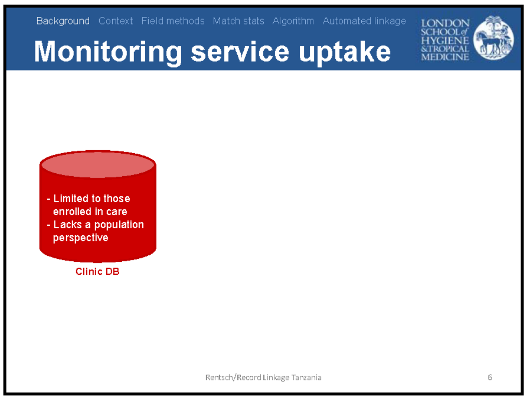


Background Context Field methods Match stats Algorithm Automated linkage

Monitoring service uptake

- As efforts continue to strengthen HIV service provision and access to ART, ways of monitoring uptake has grown in importance
- HIV care continuum

Rentsch/Record Linkage Tanzania 5

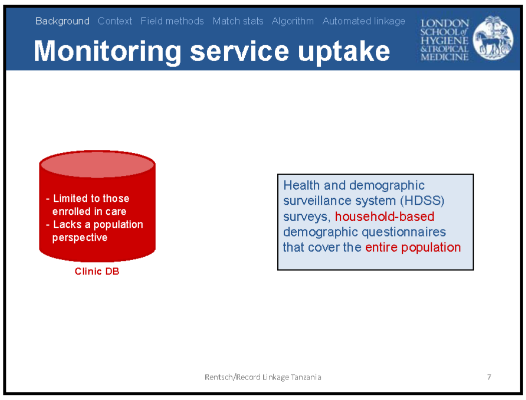


Background Context Field methods Match stats Algorithm Automated linkage

Monitoring service uptake

- Limited to those enrolled in care
- Lacks a population perspective
Clinic DB

Rentsch/Record Linkage Tanzania 6

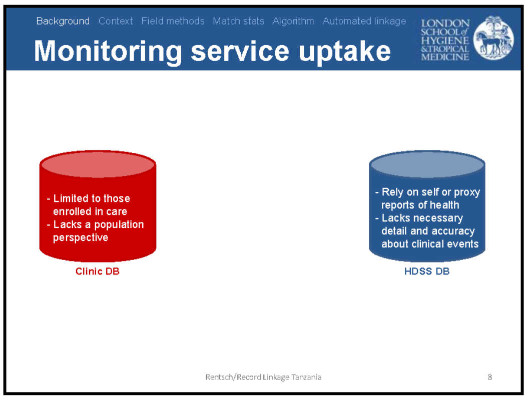


Monitoring service uptake

- Limited to those enrolled in care
- Lacks a population perspective

Clinic DB

Health and demographic surveillance system (HDSS) surveys, household-based demographic questionnaires that cover the entire population



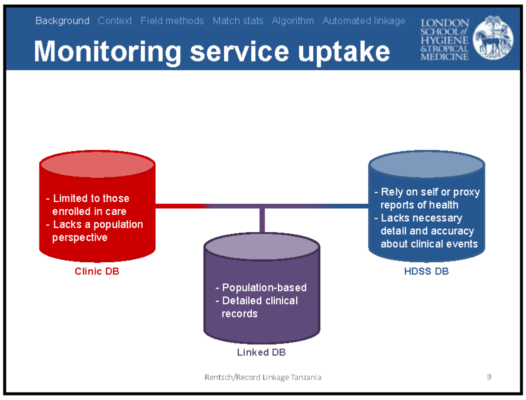
Monitoring service uptake

- Limited to those enrolled in care
- Lacks a population perspective

Clinic DB

- Rely on self or proxy reports of health
- Lacks necessary detail and accuracy about clinical events

HDSS DB



Monitoring service uptake

- Limited to those enrolled in care
- Lacks a population perspective

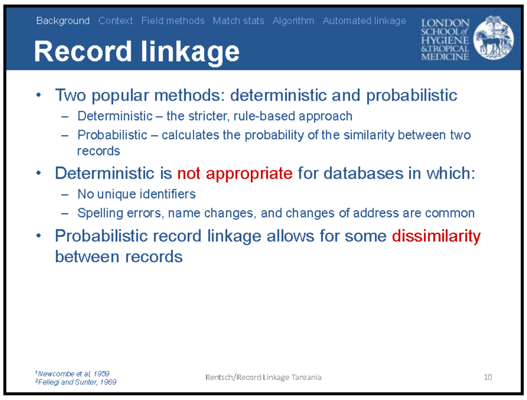
Clinic DB

- Rely on self or proxy reports of health
- Lacks necessary detail and accuracy about clinical events

HDSS DB

- Population-based
- Detailed clinical records

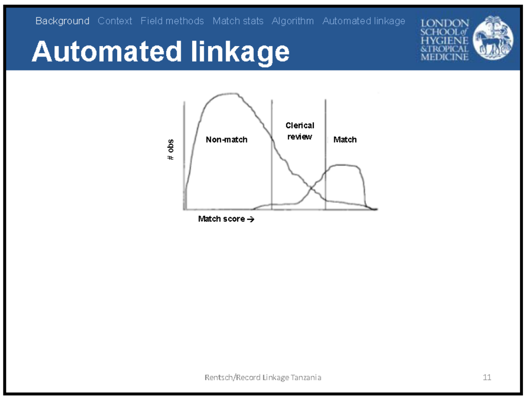
Linked DB



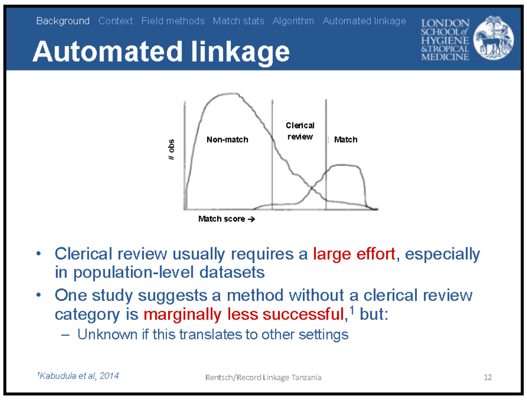
Record linkage

- Two popular methods: deterministic and probabilistic
  - Deterministic – the stricter, rule-based approach
  - Probabilistic – calculates the probability of the similarity between two records
- Deterministic is not appropriate for databases in which:
  - No unique identifiers
  - Spelling errors, name changes, and changes of address are common
- Probabilistic record linkage allows for some dissimilarity between records

[1] *Newcombe et al, 1959*
[2] *Fellegi and Sunter, 1969*



Automated linkage

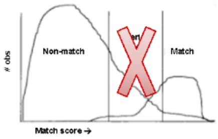


Automated linkage

- Clerical review usually requires a large effort, especially in population-level datasets
- One study suggests a method without a clerical review category is marginally less successful,[1] but:
  - Unknown if this translates to other settings

[1] *Kabudula et al, 2014*

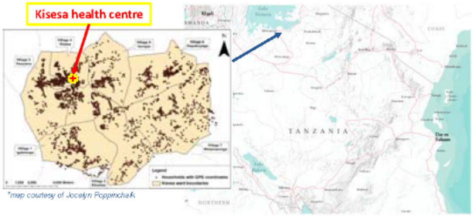
PIRL



- Point-of-contact Interactive Record Linkage (PIRL)
  - Prospective approach
  - Clerical review occurs in the presence of the individual
  - Uncertainty surrounding an individual's identity in the databases can be resolved
  - Offers chance to obtain informed consent

Study area

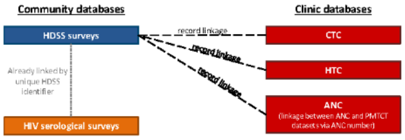
- Kisesa: HDSS (31 rounds to date); HIV serological surveys (sero-survey, 8 to date); Kisesa health centre



*map courtesy of Jocelyn Poppinchalk

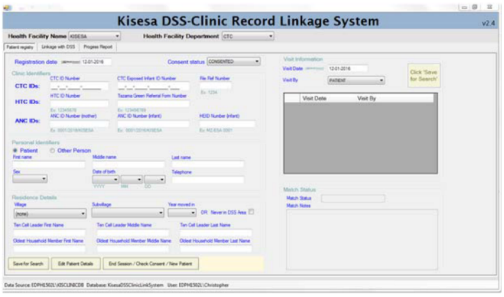
Record linkage in Kisesa

- Kisesa HDSS site provides an opportunity to carry out record linkage due to electronic databases in 3 clinics offering HIV services



Notes: HDSS = health and demographic surveillance system; CTC = HIV care and treatment centre; HTC = HIV testing and counselling clinic; ANC = antenatal clinic; PMTCT = prevention of mother-to-child transmission

Software



Kisesa DSS-Clinic Record Linkage System v2.4

Health Facility Name: KISESA
Health Facility Department: CTC

Patient registry | Linkage with DSS | Progress Report

Registration date: 12-01-2016
Consent status: CONSENTED

Clinic Identifiers
CTC IDs: CTC ID Number; CTC Exposed Infant ID Number; File Ref Number (Ex: 1234)
HTC IDs: HTC ID Number (Ex: 12345678); Tazama Green Referral Form Number (Ex: 123456789)
ANC IDs: ANC ID Number (mother) (Ex: 0001/2016/KISESA); ANC ID Number (infant) (Ex: 0001/2016/KISESA); HEID Number (infant) (Ex: KISESA 0001)

Personal Identifiers
● Patient ○ Other Person
First name; Middle name; Last name
Sex; Date of birth (YYYY MM DD); Telephone

Residence Details
Village: (none); Subvillage; Year moved in; OR Never in DSS Area □
Ten Cell Leader First Name; Ten Cell Leader Middle Name; Ten Cell Leader Last Name
Oldest Household Member First Name; Oldest Household Member Middle Name; Oldest Household Member Last Name

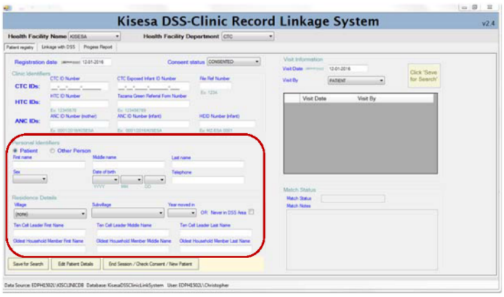
Visit Information
Visit Date: 12-01-2016
Visit By: PATIENT
Click 'Save for Search'
Visit Date | Visit By

Match Status
Match Status
Match Notes

Save for Search | Edit Patient Details | End Session / Check Consent / New Patient

Data Source: EDPHI5302\KISCLINICDB Database: KisesaDSSClinicLinkSystem User: EDPHI5302\Christopher

Personal identifiers



Kisesa DSS-Clinic Record Linkage System v2.4

Health Facility Name: KISESA
Health Facility Department: CTC

Patient registry | Linkage with DSS | Progress Report

Registration date: 12-01-2016
Consent status: CONSENTED

Clinic Identifiers
CTC IDs: CTC ID Number; CTC Exposed Infant ID Number; File Ref Number (Ex: 1234)
HTC IDs: HTC ID Number (Ex: 12345678); Tazama Green Referral Form Number (Ex: 123456789)
ANC IDs: ANC ID Number (mother) (Ex: 0001/2016/KISESA); ANC ID Number (infant) (Ex: 0001/2016/KISESA); HEID Number (infant) (Ex: KISESA 0001)

Personal Identifiers
● Patient ○ Other Person
First name; Middle name; Last name
Sex; Date of birth (YYYY MM DD); Telephone

Residence Details
Village: (none); Subvillage; Year moved in; OR Never in DSS Area □
Ten Cell Leader First Name; Ten Cell Leader Middle Name; Ten Cell Leader Last Name
Oldest Household Member First Name; Oldest Household Member Middle Name; Oldest Household Member Last Name

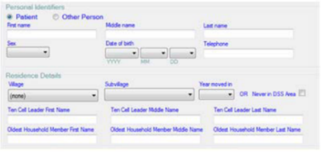
Visit Information
Visit Date: 12-01-2016
Visit By: PATIENT
Click 'Save for Search'
Visit Date | Visit By

Match Status
Match Status
Match Notes

Save for Search | Edit Patient Details | End Session / Check Consent / New Patient

Data Source: EDPHI5302\KISCLINICDB Database: KisesaDSSClinicLinkSystem User: EDPHI5302\Christopher

Personal identifiers

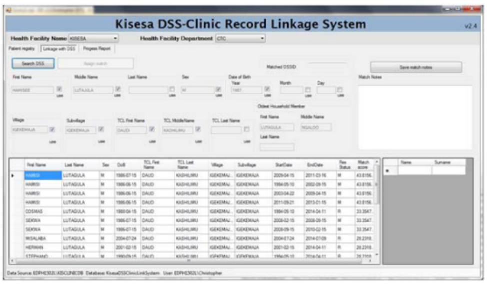


Personal Identifiers
● Patient ○ Other Person
First name; Middle name; Last name
Sex; Date of birth (YYYY MM DD); Telephone

Residence Details
Village: (none); Subvillage; Year moved in; OR Never in DSS Area □
Ten Cell Leader First Name; Ten Cell Leader Middle Name; Ten Cell Leader Last Name
Oldest Household Member First Name; Oldest Household Member Middle Name; Oldest Household Member Last Name

Parameter	Agreement condition
Three names for patient	Jaro-Winkler ≥0.8
Three names for ten-cell leader	Jaro-Winkler ≥0.8
Three names for HH member	manual review
Sex	exact
Year of birth	±2 years
Month and day of birth	exact
Village and subvillage	exact

Software demonstration



Kisesa DSS-Clinic Record Linkage System v2.4

Health Facility Name: KISESA Health Facility Department: CTC

Patient registry | Linkage with DSS | Progress Report

Search DSS | Assign match | Matched DSSID | Save match notes

First Name: HAMISEE ☑ | Middle Name: LUTAJULA ☑ | Last Name: ☐ | Sex: M ☑ | Date of Birth Year: 1987 ☑ Month: ☐ Day: ☐ | Match Notes

Village: IGEKEMAJA ☑ | Subvillage: IGEKEMAJA ☑ | TCL First Name: DAUDI ☑ | TCL MiddleName: KASHILIMU ☑ | TCL Last Name: ☐

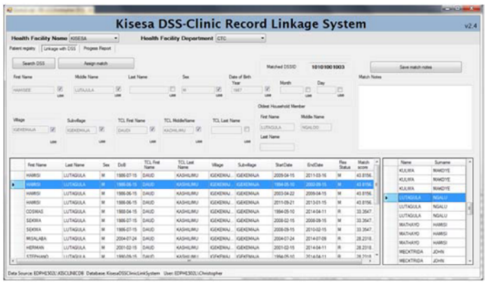
Oldest Household Member — First Name: LUTAGULA | Middle Name: NGALOO | Last Name:

First Name	Last Name	Sex	DoB	TCL First Name	TCL Last Name	Village	Subvillage	StartDate	EndDate	Res Status	Match score
HAMISI	LUTAGULA	M	1986-07-15	DAUD	KASHILIMU	IGEKEMAJ...	IGEKEMAJA	2009-04-15	2011-03-16	M	43.8156...
HAMISI	LUTAGULA	M	1986-06-15	DAUD	KASHILIMU	IGEKEMAJ...	IGEKEMAJA	1994-05-10	2002-09-15	M	43.8156...
HAMISI	LUTAGULA	M	1986-06-15	DAUD	KASHILIMU	IGEKEMAJ...	IGEKEMAJA	2003-04-22	2009-04-15	M	43.8156...
HAMISI	LUTAGULA	M	1986-06-15	DAUD	KASHILIMU	IGEKEMAJ...	IGEKEMAJA	2011-09-21	2013-01-15	M	43.8156...
COSMAS	LUTAGULA	M	1988-04-15	DAUD	KASHILIMU	IGEKEMAJ...	IGEKEMAJA	1994-05-10	2014-04-11	R	33.3547...
SEKWA	LUTAGULA	M	1986-07-15	DAUD	KASHILIMU	IGEKEMAJ...	IGEKEMAJA	2008-02-15	2008-09-15	M	33.3547...
SEKWA	LUTAGULA	M	1986-07-15	DAUD	KASHILIMU	IGEKEMAJ...	IGEKEMAJA	2008-09-15	2010-02-15	M	33.3547...
MSALABA	LUTAGULA	M	2004-07-24	DAUD	KASHILIMU	IGEKEMAJ...	IGEKEMAJA	2004-07-24	2014-07-09	R	28.2318...
HERMAN	LUTAGULA	M	2001-02-15	DAUD	KASHILIMU	IGEKEMAJ...	IGEKEMAJA	2001-02-15	2014-04-11	R	28.2318...
STEPHANO	LUTAGULA	M	1990-09-15	DAUD	KASHILIMU	IGEKEMAJ...	IGEKEMAJA	1994-05-10	2014-04-11	R	28.2318...

Name	Surname

Data Source: EDPHE502L\KISCLINICDB Database: KisesaDSSClinicLinkSystem User: EDPHE502L\Christopher

Software demonstration



Kisesa DSS-Clinic Record Linkage System v2.4

Health Facility Name: KISESA Health Facility Department: CTC

Patient registry | Linkage with DSS | Progress Report

Search DSS | Assign match | Matched DSSID 10101001003 | Save match notes

First Name: HAMISEE ☑ | Middle Name: LUTAJULA ☑ | Last Name: ☐ | Sex: M ☑ | Date of Birth Year: 1987 ☑ Month: ☐ Day: ☐ | Match Notes

Village: IGEKEMAJA ☑ | Subvillage: IGEKEMAJA ☑ | TCL First Name: DAUDI ☑ | TCL MiddleName: KASHILIMU ☑ | TCL Last Name: ☐

Oldest Household Member — First Name: LUTAGULA | Middle Name: NGALOO | Last Name:

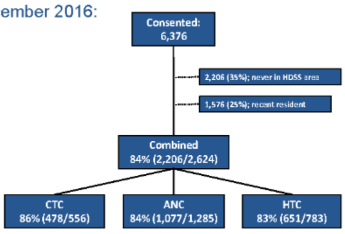
First Name	Last Name	Sex	DoB	TCL First Name	TCL Last Name	Village	Subvillage	StartDate	EndDate	Res Status	Match score
HAMISI	LUTAGULA	M	1986-07-15	DAUD	KASHILIMU	IGEKEMAJ...	IGEKEMAJA	2009-04-15	2011-03-16	M	43.8156...
HAMISI	LUTAGULA	M	1986-06-15	DAUD	KASHILIMU	IGEKEMAJ...	IGEKEMAJA	1994-05-10	2002-09-15	M	43.8156...
HAMISI	LUTAGULA	M	1986-06-15	DAUD	KASHILIMU	IGEKEMAJ...	IGEKEMAJA	2003-04-22	2009-04-15	M	43.8156...
HAMISI	LUTAGULA	M	1986-06-15	DAUD	KASHILIMU	IGEKEMAJ...	IGEKEMAJA	2011-09-21	2013-01-15	M	43.8156...
COSMAS	LUTAGULA	M	1988-04-15	DAUD	KASHILIMU	IGEKEMAJ...	IGEKEMAJA	1994-05-10	2014-04-11	R	33.3547...
SEKWA	LUTAGULA	M	1986-07-15	DAUD	KASHILIMU	IGEKEMAJ...	IGEKEMAJA	2008-02-15	2008-09-15	M	33.3547...
SEKWA	LUTAGULA	M	1986-07-15	DAUD	KASHILIMU	IGEKEMAJ...	IGEKEMAJA	2008-09-15	2010-02-15	M	33.3547...
MSALABA	LUTAGULA	M	2004-07-24	DAUD	KASHILIMU	IGEKEMAJ...	IGEKEMAJA	2004-07-24	2014-07-09	R	28.2318...
HERMAN	LUTAGULA	M	2001-02-15	DAUD	KASHILIMU	IGEKEMAJ...	IGEKEMAJA	2001-02-15	2014-04-11	R	28.2318...
STEPHANO	LUTAGULA	M	1990-09-15	DAUD	KASHILIMU	IGEKEMAJ...	IGEKEMAJA	1994-05-10	2014-04-11	R	28.2318...

Name	Surname
KULWA	MAKOYE
KULWA	MAKOYE
KULWA	MAKOYE
LUTAGULA	NGALU
LUTAGULA	NGALU
LUTAGULA	NGALU
MATHAYO	HAMISI
MATHAYO	HAMISI
MATHAYO	HAMISI
MECKTRIDA	JOHN
MECKTRIDA	JOHN

Data Source: EDPHE502L\KISCLINICDB Database: KisesaDSSClinicLinkSystem User: EDPHE502L\Christopher

Sample population

- As of 31 December 2016:



Consented: 6,376

2,206 (35%); never in HDSS area

1,576 (25%); recent resident

Combined 84% (2,206/2,624)

CTC 86% (478/556)

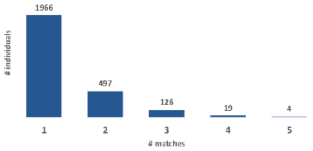
ANC 84% (1,077/1,285)

HTC 83% (651/783)

Notes: HDSS = health and demographic surveillance system; CTC = HIV care and treatment centre; HTC = HIV testing and counselling clinic; ANC = antenatal clinic

Multiple HDSS records

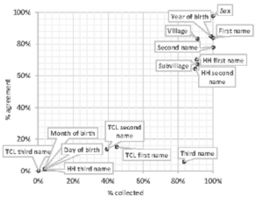
- Records found for 406 (11%) of patients who reported to have no residence history in Kisesa HDSS area
- Multiple records found for some individuals:



- Total: 3,434 matches made on 2,612 individuals

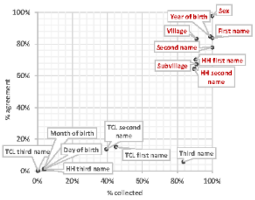
Parameter performance

% collected vs. % agreement

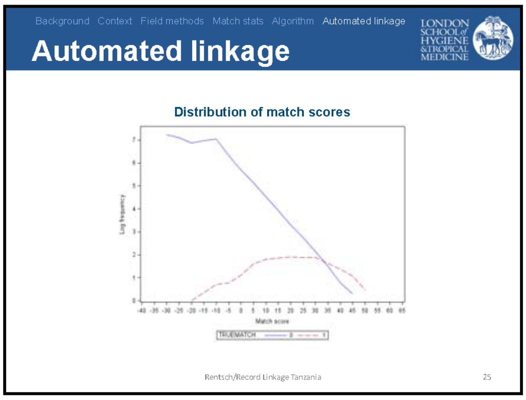


Parameter performance

% collected vs. % agreement

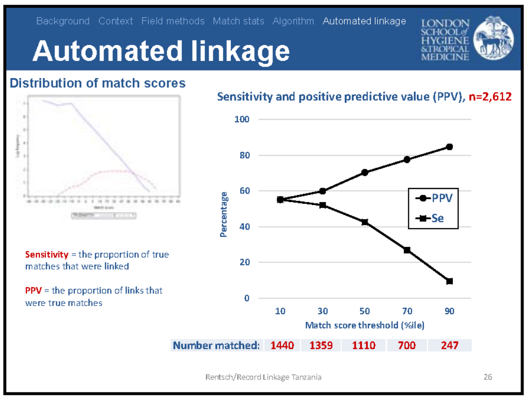


'high-performing' parameters



Automated linkage

Distribution of match scores



Automated linkage

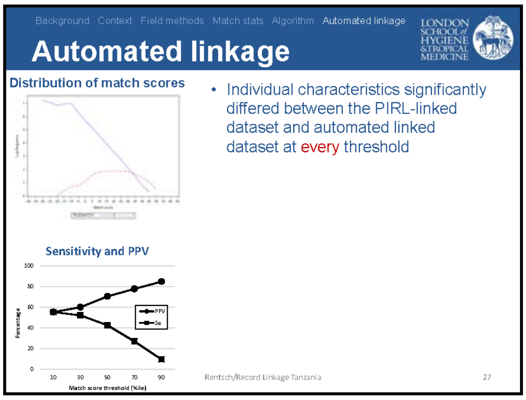
Distribution of match scores

Sensitivity and positive predictive value (PPV), n=2,612

Number matched: 1440 1359 1110 700 247

Sensitivity = the proportion of true matches that were linked

PPV = the proportion of links that were true matches

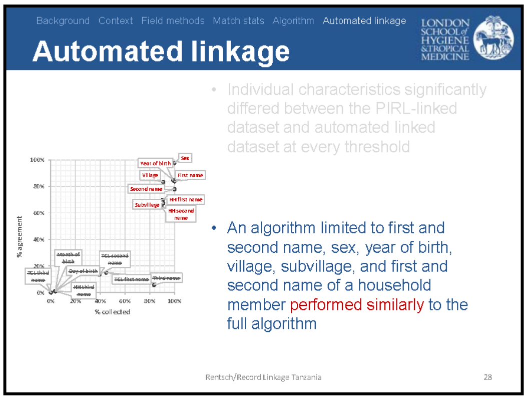


Automated linkage

Distribution of match scores

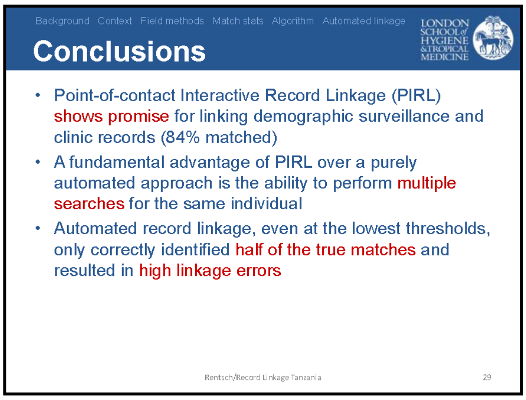
Sensitivity and PPV

- Individual characteristics significantly differed between the PIRL-linked dataset and automated linked dataset at every threshold



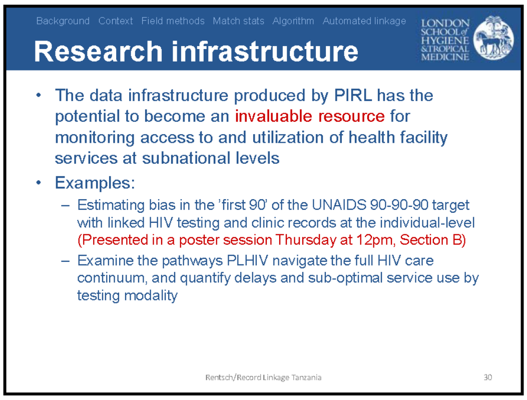
Automated linkage

- Individual characteristics significantly differed between the PIRL-linked dataset and automated linked dataset at every threshold
- An algorithm limited to first and second name, sex, year of birth, village, subvillage, and first and second name of a household member performed similarly to the full algorithm



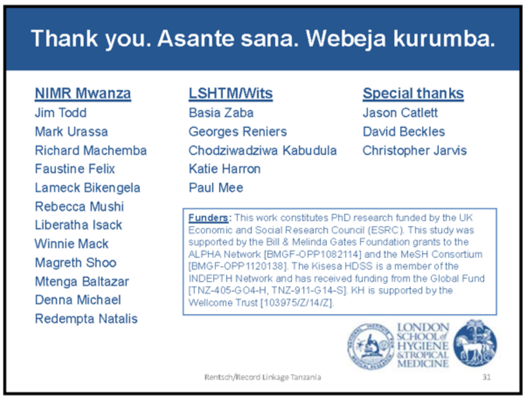
Conclusions

- Point-of-contact Interactive Record Linkage (PIRL) shows promise for linking demographic surveillance and clinic records (84% matched)
- A fundamental advantage of PIRL over a purely automated approach is the ability to perform multiple searches for the same individual
- Automated record linkage, even at the lowest thresholds, only correctly identified half of the true matches and resulted in high linkage errors



Research infrastructure

- The data infrastructure produced by PIRL has the potential to become an invaluable resource for monitoring access to and utilization of health facility services at subnational levels
- Examples:
  - Estimating bias in the 'first 90' of the UNAIDS 90-90-90 target with linked HIV testing and clinic records at the individual-level (Presented in a poster session Thursday at 12pm, Section B)
  - Examine the pathways PLHIV navigate the full HIV care continuum, and quantify delays and sub-optimal service use by testing modality



Thank you. Asante sana. Webeja kurumba.

NIMR Mwanza
Jim Todd
Mark Urassa
Richard Machemba
Faustine Felix
Lameck Bikengela
Rebecca Mushi
Liberatha Isack
Winnie Mack
Magreth Shoo
Mtenga Baltazar
Denna Michael
Redempta Natalis

LSHTM/Wits
Basia Zaba
Georges Reniers
Chodziwadziwa Kabudula
Katie Harron
Paul Mee

Special thanks
Jason Catlett
David Beckles
Christopher Jarvis

Funders: This work constitutes PhD research funded by the UK Economic and Social Research Council (ESRC). This study was supported by the Bill & Melinda Gates Foundation grants to the ALPHA Network [BMGF-OPP1082114] and the MeSH Consortium [BMGF-OPP1120138]. The Kisesa HDSS is a member of the INDEPTH Network and has received funding from the Global Fund [TNZ-405-GO4-H, TNZ-911-G14-S]. KH is supported by the Wellcome Trust [103975/Z/14/Z].

Rentsch/Record Linkage Tanzania 31

Thank you. Asante sana. Webeja kurumba.

The Kisesa team

Questions or comments?

Christopher.Rentsch@lshtm.ac.uk

Rentsch/Record Linkage Tanzania 32

10.1.6 International Workshop on HIV Observational Databases (IWHOD), 2018

Time from HIV diagnosis to care by testing modality in a rural Tanzanian community

Rentsch CT[1], Reniers G[1,2], Machemba R[3], Mtenga B[3], Michael D[3], Kabudula C[2], Natalis R[4], Urassa M[3], Todd J[1,3], Zaba B[1]

LONDON SCHOOL of HYGIENE &TROPICAL MEDICINE

[1]Department of Population Health, London School of Hygiene & Tropical Medicine, London, UK
[2]School of Public Health, University of the Witwatersrand. Johannesburg, South Africa
[3]The Tazama Project, National Institute for Medical Research, Mwanza, Tanzania
[4]District Medical Office, Magu District, Tanzania

Introduction

- HIV testing and counselling (HTC) is the first critical step for subsequent linkage to care and antiretroviral treatment (ART)
- Community-based HTC, such as that conducted during population-based HIV serological surveys, has increased testing coverage in high HIV burden areas like sub-Saharan Africa[1]
- However, it is not yet clear whether community-based HTC results in better linkage to care than facility-based testing

Specific Aims

AIM 1 Compare time from first positive HIV diagnostic test to registration at a local HIV care and treatment clinic (CTC) by testing modality (community- vs. facility-based HTC)

Hypothesis: Individuals who tested HIV-positive in a facility will have higher rates of linkage to care than individuals who tested HIV-positive during a sero-survey

AIM 2 Among those in care, compare initial CD4 count and time from CTC registration to ART initiation by testing modality

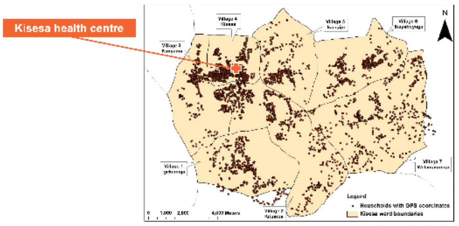
Hypothesis: Individuals who tested HIV-positive in a facility will have lower initial CD4 counts and more rapidly initiate ART than individuals who tested HIV-positive during a sero-survey

Methods

Data sources

The TAZAMA Project, which includes multiple rounds of health and demographic surveillance surveys (HDSS) and population-based HIV sero-surveys, is located in a rural ward in the Magu district of Mwanza region in northwest Tanzania

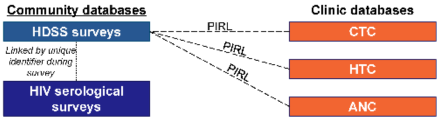
Fig. 1 Map of Kisesa, Tanzania



*map courtesy of Jocelyn Poppinchelli

Since 2015, medical records from three clinics within the HDSS area have been prospectively linked to HDSS and sero-survey records using point-of-contact interactive record linkage (PIRL) (Fig. 2)[2,3]

Fig. 2 Point-of-contact interactive record linkage (PIRL) in The TAZAMA Project



CTC = HIV care and treatment centre; HTC = HIV testing and counselling clinic; ANC = antenatal clinic

Sample

- All adults (≥15 years) who received a positive HIV diagnostic test in a sero-survey, HTC, or ANC between 2015-2017
- Excluded: individuals who had a diagnostic HIV+ test in a previous sero-survey or initiated care prior to their positive HIV test (repeat testers), and those who reported residence outside HDSS area or not seen in 2016/17 HDSS survey

Statistical analyses

- Chi-square (χ2) and Wilcoxon Rank-Sum tests were used to assess differences by testing modality
- Crude and adjusted Cox regression models were used to compare time to CTC registration by testing modality

Results

Table 1 Demographic characteristics by location of first HIV+ test

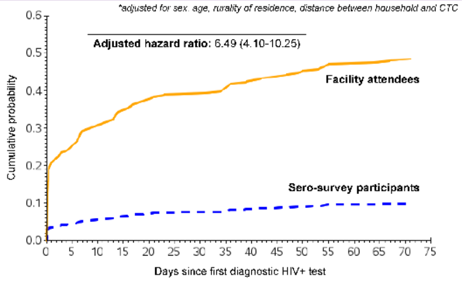
Characteristic	Facility attendees (n=146)	Sero-survey participants (n=265)	p-value
Clinic			
ANC	17 (11.6)	-	-
HTC	129 (88.4)	-	
Sex			
Female	100 (68.5)	172 (64.9)	0.4619
Male	46 (31.5)	93 (35.1)	
Age, years			
15-29	58 (39.7)	62 (23.4)	0.0039
30-39	49 (33.6)	98 (37.0)	
40-49	22 (15.1)	58 (21.9)	
50+	17 (11.6)	47 (17.7)	
Village			
Igekemaja	16 (11.0)	30 (11.3)	0.0455
Ihayabuyaga	7 (4.8)	32 (12.1)	
Isangijo	14 (9.6)	27 (10.2)	
Kanyama	24 (16.4)	36 (13.6)	
Kisesa	46 (31.5)	70 (26.4)	
Kitumba	29 (19.9)	35 (13.2)	
Welamasonga	10 (6.9)	35 (13.2)	
Subvillage of residence, type			
Rural	61 (41.8)	145 (54.7)	0.0348
Peri-urban	42 (28.8)	54 (20.4)	
Urban	43 (29.5)	66 (24.9)	
Subvillage of residence, has road			
Yes	67 (45.9)	104 (39.3)	0.1909
No	79 (54.1)	161 (60.8)	
Distance from HH to clinic, km			
<1	28 (19.2)	44 (16.6)	0.0957
1-1.9	43 (29.5)	54 (20.4)	
2-4.9	29 (19.9)	49 (18.5)	
5-11	35 (24.0)	86 (32.5)	
Unknown	11 (7.5)	32 (12.1)	
Registered at CTC	71 (48.6)	27 (10.2)	<0.0001

146 facility attendees and 265 sero-survey participants received their first HIV+ diagnostic test between 2015-2017

- Facility attendees were proportionately younger and from more urban areas

Fig. 3 Adjusted* cumulative probability to register for care by testing modality

*adjusted for sex, age, rurality of residence, distance between household and CTC



Half of facility attendees and 10% of sero-survey participants registered for care in the Kisesa CTC ≤90 days after HIV+ test

- Men 56% more likely than women to register at CTC (95% CI: 1.00-2.41)

38/71 (54%) of facility attendees who registered at CTC did so on day of HIV+ test (CTC <20 metres from HTC and ANC)

- After removing these 38 patients in sensitivity analysis: HR: 3.42 (95% CI: 2.06-5.67); sex differential increases to 92% (1.10-3.35)

Among those who registered for care:

- **Median initial CD4 count** was 302 cells/mm^3 (IQR 156-398) among facility attendees vs. 436 cells/mm^3 (IQR 126-557) among sero-survey participants (p=0.25)
- No difference in **ART initiation rates**: 94% vs. 93% (p=0.67)

Conclusions

Community-based HTC is important to expand testing coverage; however, these efforts should include interventions to link individuals who test positive into care

References

1. Isingo R et al. Trends in the uptake of voluntary counselling and testing for HIV in rural Tanzania in the context of the scale up of antiretroviral therapy. Trop Med Int Health. 2012;17(8):e15-25
2. Rentsch CT et al. Point of contact Interactive Record Linkage (PIRL): A software tool to prospectively link demographic surveillance and health facility data [version 2; referees: 2 approved]. Gates Open Res. 2018;1(8)
3. Rentsch CT et al. Point-of-contact interactive record linkage (PIRL) between demographic surveillance and health facility data in rural Tanzania. International Journal for Population Data Science. 2017;2(1)

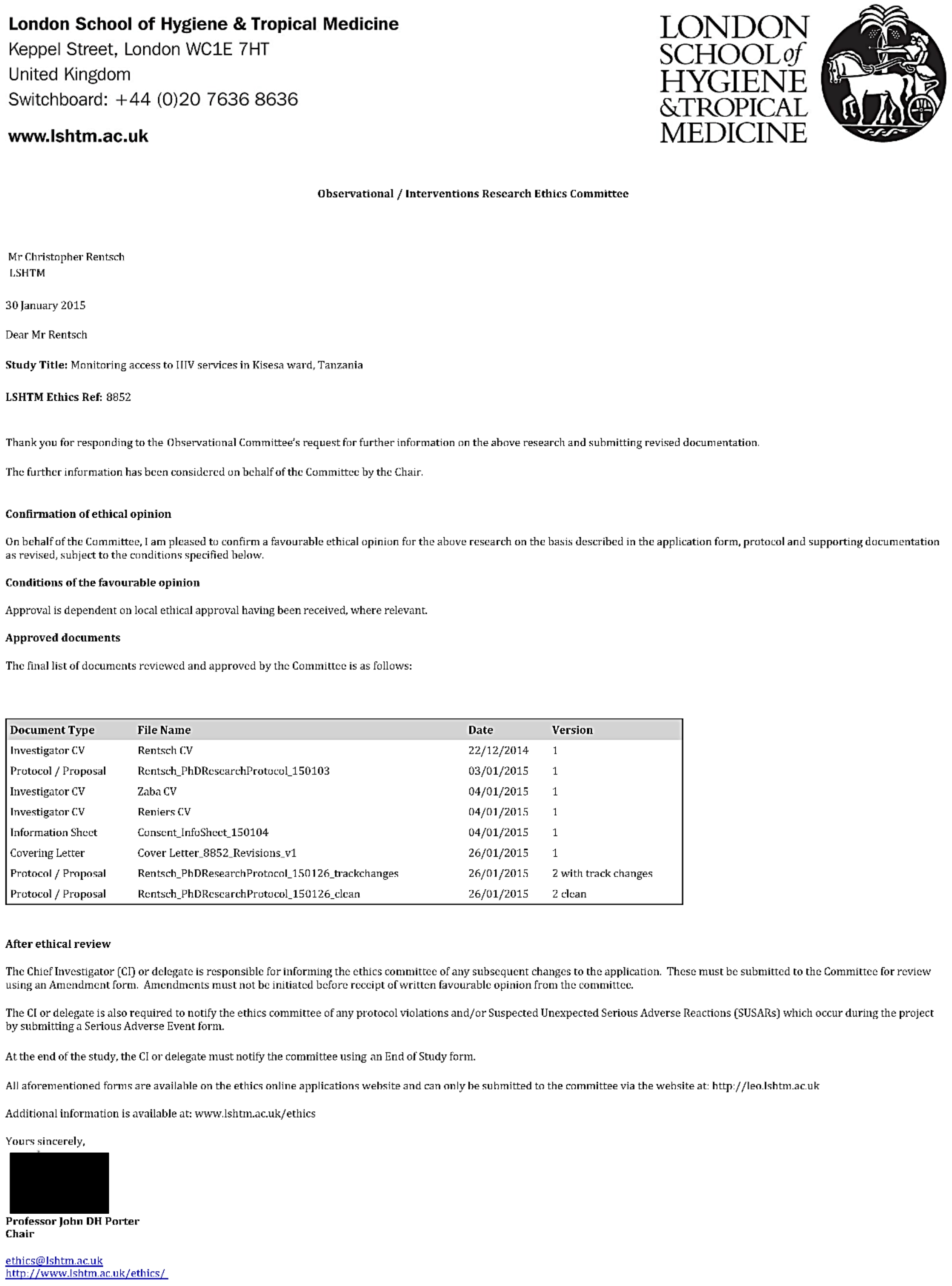
Funding

This work constitutes PhD research funded by the UK Economic and Social Research Council (ESRC). This study was supported by the Bill & Melinda Gates Foundation grants to the ALPHA Network [OPP1082114], MeSH Consortium [OPP1120138], and HIV Modelling Consortium [OPP1084364]. The Kisesa HDSS is a member of the INDEPTH Network and has received funding from the Global Fund [TNZ-405-G04-H, TNZ-911-G14-S].

Improving health worldwide www.lshtm.ac.uk

10.2 Ethical clearances

10.2.1 LSHTM (original)



London School of Hygiene & Tropical Medicine
Keppel Street, London WC1E 7HT
United Kingdom
Switchboard: +44 (0)20 7636 8636

www.lshtm.ac.uk

LONDON SCHOOL of HYGIENE &TROPICAL MEDICINE

Observational / Interventions Research Ethics Committee

Mr Christopher Rentsch
LSHTM

30 January 2015

Dear Mr Rentsch

Study Title: Monitoring access to HIV services in Kisesa ward, Tanzania

LSHTM Ethics Ref: 8852

Thank you for responding to the Observational Committee's request for further information on the above research and submitting revised documentation.

The further information has been considered on behalf of the Committee by the Chair.

Confirmation of ethical opinion

On behalf of the Committee, I am pleased to confirm a favourable ethical opinion for the above research on the basis described in the application form, protocol and supporting documentation as revised, subject to the conditions specified below.

Conditions of the favourable opinion

Approval is dependent on local ethical approval having been received, where relevant.

Approved documents

The final list of documents reviewed and approved by the Committee is as follows:

Document Type	File Name	Date	Version
Investigator CV	Rentsch CV	22/12/2014	1
Protocol / Proposal	Rentsch_PhDResearchProtocol_150103	03/01/2015	1
Investigator CV	Zaba CV	04/01/2015	1
Investigator CV	Reniers CV	04/01/2015	1
Information Sheet	Consent_InfoSheet_150104	04/01/2015	1
Covering Letter	Cover Letter_8852_Revisions_v1	26/01/2015	1
Protocol / Proposal	Rentsch_PhDResearchProtocol_150126_trackchanges	26/01/2015	2 with track changes
Protocol / Proposal	Rentsch_PhDResearchProtocol_150126_clean	26/01/2015	2 clean

After ethical review

The Chief Investigator (CI) or delegate is responsible for informing the ethics committee of any subsequent changes to the application. These must be submitted to the Committee for review using an Amendment form. Amendments must not be initiated before receipt of written favourable opinion from the committee.

The CI or delegate is also required to notify the ethics committee of any protocol violations and/or Suspected Unexpected Serious Adverse Reactions (SUSARs) which occur during the project by submitting a Serious Adverse Event form.

At the end of the study, the CI or delegate must notify the committee using an End of Study form.

All aforementioned forms are available on the ethics online applications website and can only be submitted to the committee via the website at: http://leo.lshtm.ac.uk

Additional information is available at: www.lshtm.ac.uk/ethics

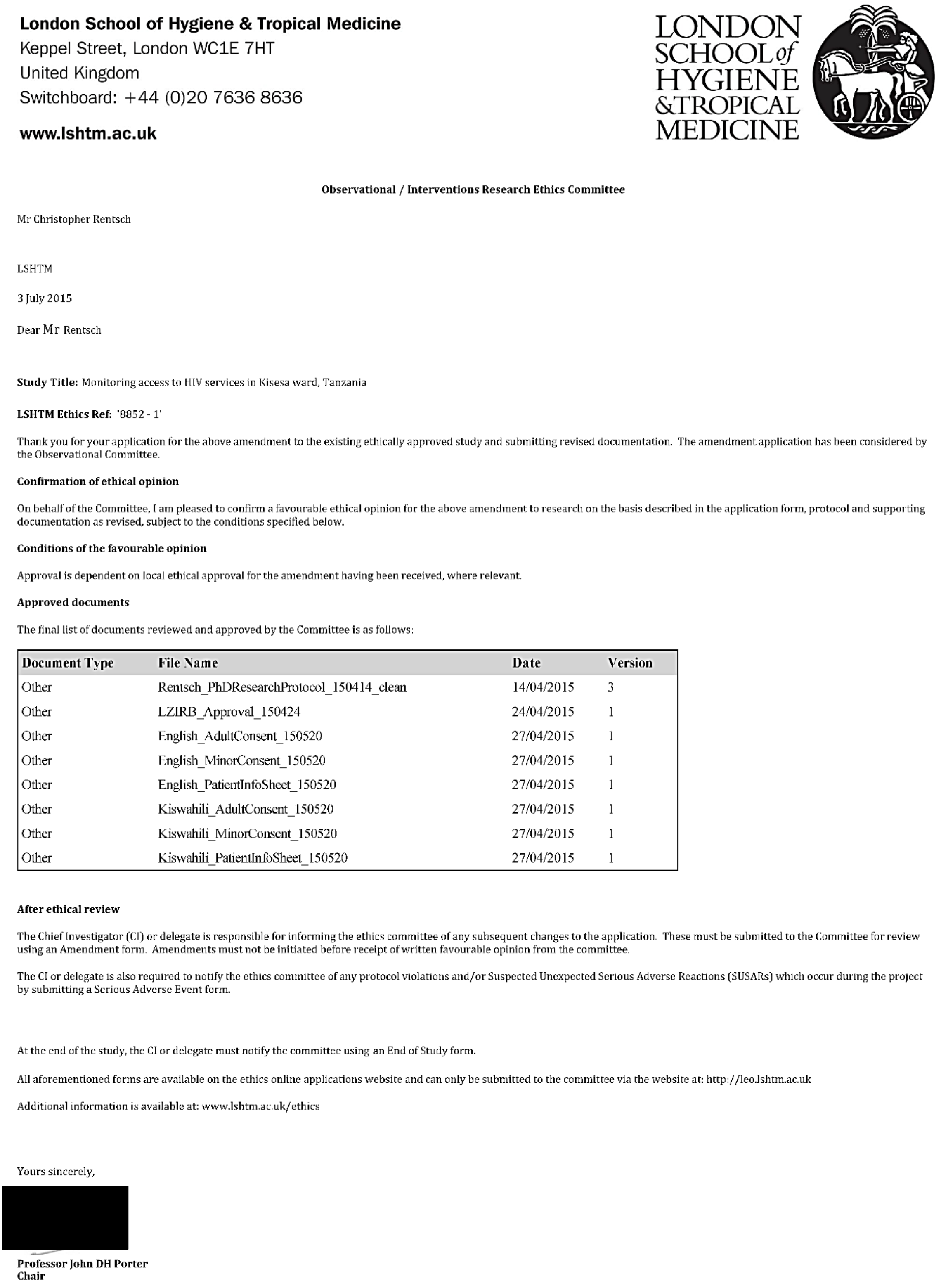
Yours sincerely,

[redacted]

Professor John DH Porter
Chair

ethics@lshtm.ac.uk
http://www.lshtm.ac.uk/ethics/

10.2.2 LSHTM (amendment)



London School of Hygiene & Tropical Medicine
Keppel Street, London WC1E 7HT
United Kingdom
Switchboard: +44 (0)20 7636 8636

www.lshtm.ac.uk

Observational / Interventions Research Ethics Committee

Mr Christopher Rentsch

LSHTM

3 July 2015

Dear Mr Rentsch

Study Title: Monitoring access to HIV services in Kisesa ward, Tanzania

LSHTM Ethics Ref: '8852 - 1'

Thank you for your application for the above amendment to the existing ethically approved study and submitting revised documentation. The amendment application has been considered by the Observational Committee.

Confirmation of ethical opinion

On behalf of the Committee, I am pleased to confirm a favourable ethical opinion for the above amendment to research on the basis described in the application form, protocol and supporting documentation as revised, subject to the conditions specified below.

Conditions of the favourable opinion

Approval is dependent on local ethical approval for the amendment having been received, where relevant.

Approved documents

The final list of documents reviewed and approved by the Committee is as follows:

Document Type	File Name	Date	Version
Other	Rentsch_PhDResearchProtocol_150414_clean	14/04/2015	3
Other	LZIRB_Approval_150424	24/04/2015	1
Other	English_AdultConsent_150520	27/04/2015	1
Other	English_MinorConsent_150520	27/04/2015	1
Other	English_PatientInfoSheet_150520	27/04/2015	1
Other	Kiswahili_AdultConsent_150520	27/04/2015	1
Other	Kiswahili_MinorConsent_150520	27/04/2015	1
Other	Kiswahili_PatientInfoSheet_150520	27/04/2015	1

After ethical review

The Chief Investigator (CI) or delegate is responsible for informing the ethics committee of any subsequent changes to the application. These must be submitted to the Committee for review using an Amendment form. Amendments must not be initiated before receipt of written favourable opinion from the committee.

The CI or delegate is also required to notify the ethics committee of any protocol violations and/or Suspected Unexpected Serious Adverse Reactions (SUSARs) which occur during the project by submitting a Serious Adverse Event form.

At the end of the study, the CI or delegate must notify the committee using an End of Study form.

All aforementioned forms are available on the ethics online applications website and can only be submitted to the committee via the website at: http://leo.lshtm.ac.uk

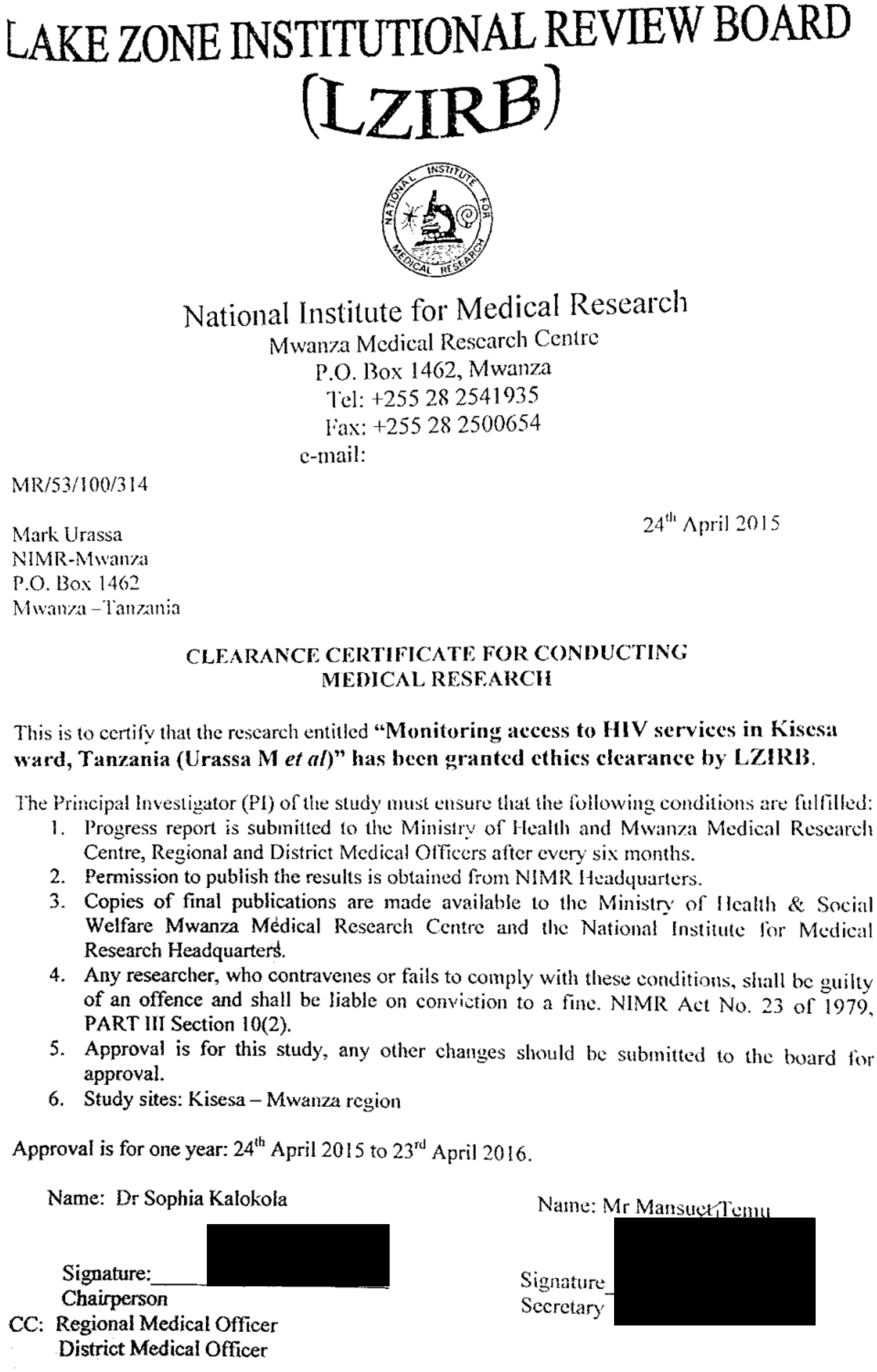
Additional information is available at: www.lshtm.ac.uk/ethics

Yours sincerely,

[redacted]

Professor John DH Porter
Chair

10.2.3 Local approval



LAKE ZONE INSTITUTIONAL REVIEW BOARD (LZIRB)

National Institute for Medical Research
Mwanza Medical Research Centre
P.O. Box 1462, Mwanza
Tel: +255 28 2541935
Fax: +255 28 2500654
e-mail:

MR/53/100/314

24th April 2015

Mark Urassa
NIMR-Mwanza
P.O. Box 1462
Mwanza –Tanzania

CLEARANCE CERTIFICATE FOR CONDUCTING MEDICAL RESEARCH

This is to certify that the research entitled **"Monitoring access to HIV services in Kisesa ward, Tanzania (Urassa M *et al*)" has been granted ethics clearance by LZIRB.**

The Principal Investigator (PI) of the study must ensure that the following conditions are fulfilled:

1. Progress report is submitted to the Ministry of Health and Mwanza Medical Research Centre, Regional and District Medical Officers after every six months.
2. Permission to publish the results is obtained from NIMR Headquarters.
3. Copies of final publications are made available to the Ministry of Health & Social Welfare Mwanza Médical Research Centre and the National Institute for Medical Research Headquarters.
4. Any researcher, who contravenes or fails to comply with these conditions, shall be guilty of an offence and shall be liable on conviction to a fine. NIMR Act No. 23 of 1979, PART III Section 10(2).
5. Approval is for this study, any other changes should be submitted to the board for approval.
6. Study sites: Kisesa – Mwanza region

Approval is for one year: 24th April 2015 to 23rd April 2016.

Name: Dr Sophia Kalokola

Signature:[redacted]
Chairperson

Name: Mr Mansuet Temu

Signature[redacted]
Secretary

CC: Regional Medical Officer
District Medical Officer

10.3 CONSENT/ASSENT FORMS

10.3.1 Adult (≥15 years)

English (shown), Kiswahili also made available

TAZAMA PROJECT, NATIONAL INSTITUTE FOR MEDICAL RESEARCH
Study title: Monitoring access to HIV services in Kisesa ward, Tanzania
ADULT INFORMED CONSENT

Greetings,

My name is _____ (name of researcher), I am working for the TAZAMA Project at the National Institute for Medical Research. The TAZAMA Project has been working with the communities in Kisesa and Bukandwe since 1994 to help understand and improve the health problems that they face. We are now doing a study on the use of health care services in this community. In order to do that we want to link the information we are collecting during the TAZAMA Project census you or somebody in your household has likely already completed with that from clinics and health centres.

Our long-term goal is to improve health services in this community, and in order to do that we hope to learn from this study which groups of people use the clinics and health centres for chronic care services, and which people do not. We are not really interested in information about any individual in particular, but we want to learn about the patterns in the community in general.

We will use the demographic information (name, date of birth, village of residence) you give us to search for the number that we have used for you in the TAZAMA Project census. This process will take about 5-10 minutes. After we link the information we will remove all the names and other information that could identify you. We treat all the information that we collect with the highest confidentiality and we do not share any identifying information with anyone outside this project.

Your participation is entirely voluntary. By that I mean that you may refuse or accept to participate and your decision will not benefit or harm you in any way and you will be assisted in this health facility as usual. You may also withdraw your participation in the study at any time without any further consequences.

This study has been approved by the Lake Zone Institutional Review Board (LZIRB), the National Health Research Ethics Sub-Committee (NatREC), and the London School of Hygiene & Tropical Medicine (LSHTM) Research Ethics Committee. If you have questions about this study, you may contact Mark Urassa, the TAZAMA Project Leader at the National Institute for Medical Research, PO Box 1462, Isimilo Rd, Mwanza or by telephone, 0784 74 13 60. You may also contact the LSHTM Research Ethics Committee at ethics@lshtm.ac.uk.

Please also accept this information sheet that contains a general explanation of this study and the contact details of the people that are responsible for this study in case you have more questions or complaints.

ADULT CONSENT:

- I confirm that I have been informed about the nature, conduct, benefits, and risks of this study entitled “Monitoring access to HIV services in Kisesa ward, Tanzania.”
- I have also read/listened and understood the above information regarding this study, and I have received a copy of the patient information sheet.
- I am aware that the results of the study, including personal identification details and medical details will be anonymously processed into a study report.
- I may, at any stage, without prejudice, withdraw my consent and participation in the study.
- I have had sufficient opportunity to ask questions and declare myself prepared to participate in the study.

ADULT PARTICIPANT:

Printed Name Signature / Mark or Thumbprint Date

WITNESS:

Printed Name Signature / Mark or Thumbprint Date

I, ____________________________________ [*Study clerk]*, confirm that the above participant has been fully informed about the nature, conduct and risks of the above study.

STUDY STAFF/INVESTIGATOR:

Printed Name Signature Date

10.3.2 Child (<15 years)

English (shown), Kiswahili also made available

TAZAMA PROJECT, NATIONAL INSTITUTE FOR MEDICAL RESEARCH

Study title: Monitoring access to HIV services in Kisesa ward, Tanzania

MINOR INFORMED CONSENT

Greetings,

My name is _____ (name of researcher), I am working for the TAZAMA Project at the National Institute for Medical Research. The TAZAMA Project has been working with the communities in Kisesa and Bukandwe since 1994 to help understand and improve the health problems that they face. We are now doing a study on the use of health care services in this community. In order to do that we want to link the information we are collecting during the TAZAMA Project census you or somebody in your household has likely already completed with that from clinics and health centres.

Our long-term goal is to improve health services in this community, and in order to do that we hope to learn from this study which groups of people use the clinics and health centres for chronic care services, and which people do not. We are not really interested in information about any individual in particular, but we want to learn about the patterns in the community in general.

We will use the demographic information (name, date of birth, village of residence) you give us to search for the number that we have used for you in the TAZAMA Project census. This process will take about 5-10 minutes. After we link the information we will remove all the names and other information that could identify you. We treat all the information that we collect with the highest confidentiality and we do not share any identifying information with anyone outside this project.

Your participation is entirely voluntary. By that I mean that you may refuse or accept to participate and your decision will not benefit or harm you in any way and you will be assisted in this health facility as usual. You may also withdraw your participation in the study at any time without any further consequences.

This study has been approved by the Lake Zone Institutional Review Board (LZIRB), the National Health Research Ethics Sub-Committee (NatREC), and the London School of Hygiene & Tropical Medicine (LSHTM) Research Ethics Committee. If you have questions about this study, you may contact Mark Urassa, the TAZAMA Project Leader at the National Institute for Medical Research, PO Box 1462, Isimilo Rd, Mwanza or by telephone, 0784 74 13 60. You may also contact the LSHTM Research Ethics Committee at ethics@lshtm.ac.uk.

Please also accept this information sheet that contains a general explanation of this study and the contact details of the people that are responsible for this study in case you have more questions or complaints.

PARENTAL CONSENT:

- I confirm that I have been informed about the nature, conduct, benefits, and risks of this study entitled "Monitoring access to HIV services in Kisesa ward, Tanzania."
- I have also read/listened and understood the above information regarding this study, and I have received a copy of the patient information sheet.
- I am aware that the results of the study, including personal identification details and medical details will be anonymously processed into a study report.
- My child may, at any stage, without prejudice, withdraw my consent and participation in the study.
- My child and I have had sufficient opportunity to ask questions and I freely consent for my child to participate in this study.

PARENT (OR LEGAL REPRESENTATIVE):

Printed Name Signature / Mark or Thumbprint Date

MINOR ASSENT:

- I confirm that I have been informed about the nature, conduct, benefits, and risks of this study entitled "Monitoring access to HIV services in Kisesa ward, Tanzania."
- I have also read/listened and understood the above information regarding this study, and I have received a copy of the patient information sheet.
- I am aware that the results of the study, including personal identification details and medical details will be anonymously processed into a study report.
- I may, at any stage, without prejudice, withdraw my consent and participation in the study.
- I have had sufficient opportunity to ask questions and declare myself prepared to participate in the study.

MINOR PARTICIPANT:

Printed Name Signature / Mark or Thumbprint Date

WITNESS:

Printed Name Signature / Mark or Thumbprint Date

I, ________________________________ [*Study clerk]*, confirm that the above participant has been fully informed about the nature, conduct and risks of the above study.

STUDY STAFF/INVESTIGATOR:

Printed Name Signature Date

10.3.3 Patient information sheet

English (shown), Kiswahili also made available

TAZAMA PROJECT, NATIONAL INSTITUTE FOR MEDICAL RESEARCH

Study Title: Monitoring access to HIV services in Kisesa ward, Tanzania

PATIENT INFORMATION SHEET

Greetings,

The TAZAMA Project at the National Institute for Medical Research has been working with the communities in Kisesa and Bukandwe for over 20 years, and reports to the Government and to the local district council so that they can better plan for the health needs of the people living in this area. We are currently doing a study that will help us understand the use of health services in this community, and want to link the information we collect during the TAZAMA Project census with that from clinics and health centres.

Our long-term goal is to improve health services in this community, and in order to do that we hope to learn from this study which groups of people use the clinics and health centres for chronic care services, and which people do not. We are not really interested in information about any individual in particular, but we want to learn about the patterns in the community in general.

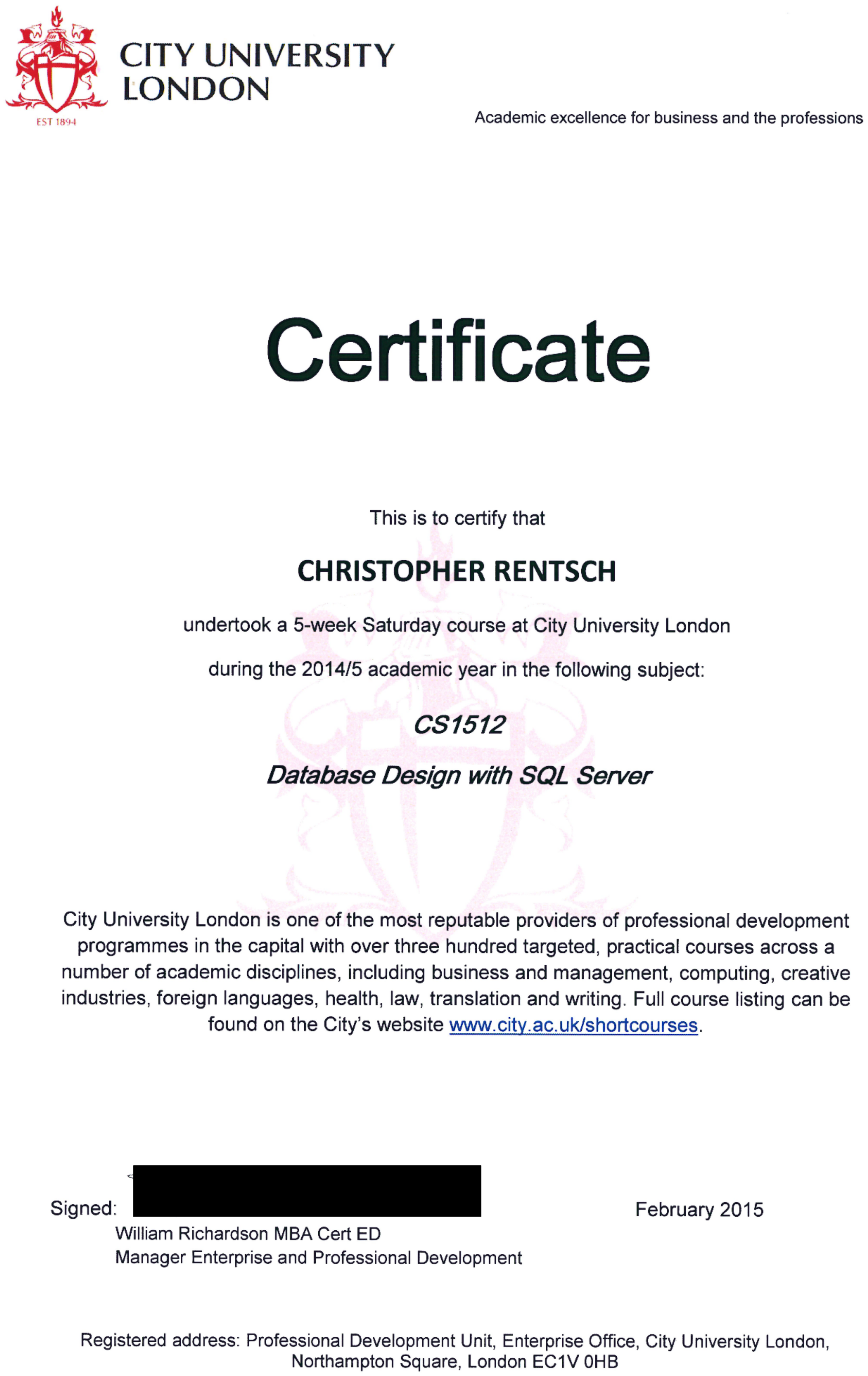
We would like to use the demographic information (name, date of birth, village of residence) you give us to search for the number that we have used for you in the TAZAMA Project census. After we link the information we will remove all the names and other information that could identify you. We treat all the information that we collect with the highest confidentiality and we do not share any identifying information with anyone outside this project.

Your participation is entirely voluntary. By that I mean that you may refuse or accept to participate and your decision will not benefit or harm you in any way and you will be assisted in this health facility as usual. If you allow us to link your information from the health centre to your information in the TAZAMA Project census, you may also withdraw your participation in the study at any time without any further consequences.

This study has been approved by the Lake Zone Institutional Review Board (LZIRB), the National Health Research Ethics Sub-Committee (NatREC), and the London School of Hygiene & Tropical Medicine (LSHTM) Research Ethics Committee. If you have questions about this study, you may contact Mark Urassa, the TAZAMA Project Leader at the National Institute for Medical Research, PO Box 1462, Isimilo Rd, Mwanza or by telephone, 0784 74 13 60. You may also contact the LSHTM Research Ethics Committee at ethics@lshtm.ac.uk.

10.4 TRAINING CERTIFICATES

10.4.1 City University London, Database Design with SQL Server



CITY UNIVERSITY LONDON

EST 1894

Academic excellence for business and the professions

Certificate

This is to certify that

CHRISTOPHER RENTSCH

undertook a 5-week Saturday course at City University London

during the 2014/5 academic year in the following subject:

CS1512

Database Design with SQL Server

City University London is one of the most reputable providers of professional development programmes in the capital with over three hundred targeted, practical courses across a number of academic disciplines, including business and management, computing, creative industries, foreign languages, health, law, translation and writing. Full course listing can be found on the City's website www.city.ac.uk/shortcourses.

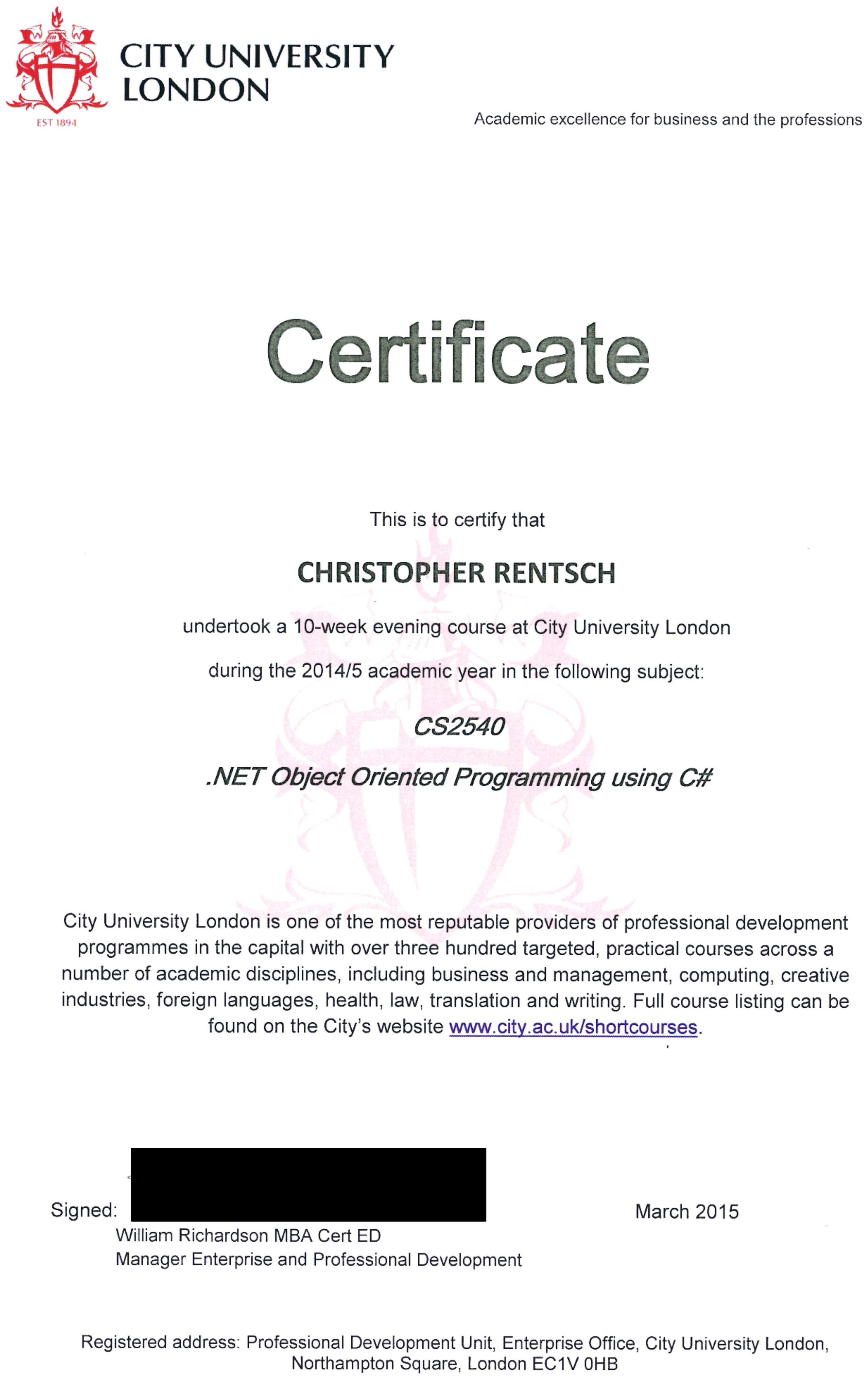
Signed: [redacted]

William Richardson MBA Cert ED
Manager Enterprise and Professional Development

February 2015

Registered address: Professional Development Unit, Enterprise Office, City University London,
Northampton Square, London EC1V 0HB

10.4.2 City University London, .NET Object Oriented Programming using C#



CITY UNIVERSITY
LONDON

EST 1894

Academic excellence for business and the professions

Certificate

This is to certify that

CHRISTOPHER RENTSCH

undertook a 10-week evening course at City University London

during the 2014/5 academic year in the following subject:

CS2540

.NET Object Oriented Programming using C#

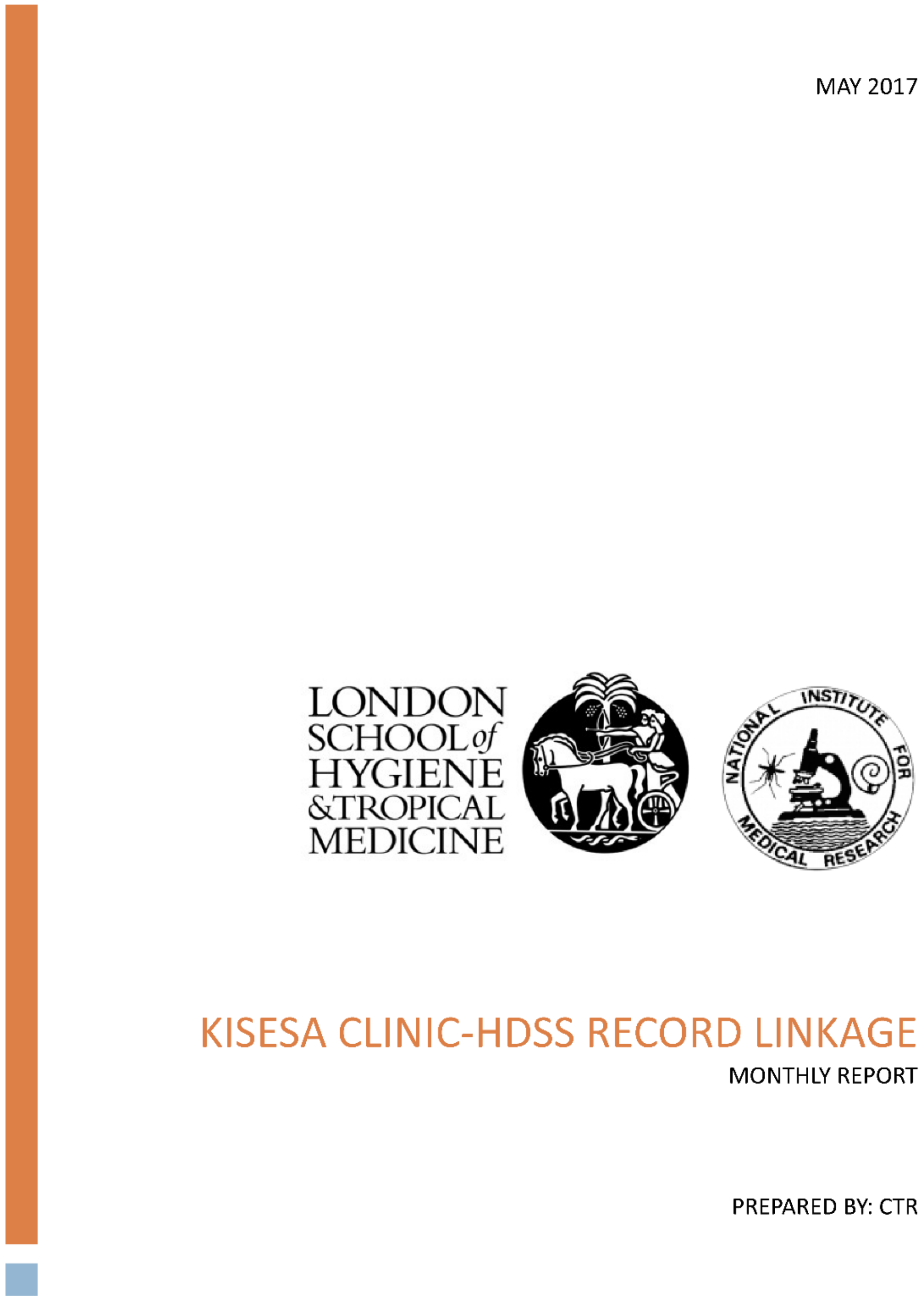
City University London is one of the most reputable providers of professional development programmes in the capital with over three hundred targeted, practical courses across a number of academic disciplines, including business and management, computing, creative industries, foreign languages, health, law, translation and writing. Full course listing can be found on the City's website www.city.ac.uk/shortcourses.

Signed: [redacted]
William Richardson MBA Cert ED
Manager Enterprise and Professional Development

March 2015

Registered address: Professional Development Unit, Enterprise Office, City University London,
Northampton Square, London EC1V 0HB

10.5 Example monthly report to NIMR/TAZAMA (May 2017)



MAY 2017

LONDON SCHOOL of HYGIENE & TROPICAL MEDICINE

NATIONAL INSTITUTE FOR MEDICAL RESEARCH

KISESA CLINIC-HDSS RECORD LINKAGE

MONTHLY REPORT

PREPARED BY: CTR

OVERVIEW

Executive summary

- This document provides updates and feedback to all those involved in record linkage in Kisesa Health Centre. I am most happy to hear your comments and any feedback.
- All of the statistics in this document are updated through 31 May 2017.

Upgrades

- I have added a field validation check to the ANC mother and infant ID fields in the software. With this check, the issue of mistyping an ANC ID should be mitigated.
- I performed a manual, back-end inspection of the data to verify the matches made in the field. These data integrity checks flagged individuals who were matched to multiple HDSS records with large age differences (>10 years), of conflicting sex, within the same household, or with overlapping residency episodes in which one record's start date occurred before another record's end date. Over the last two years, only 8 matches were deemed unlikely and were deleted from the system.

Areas to improve

- Be very cautious when we are working to do a comprehensive search. Also, please remember to have daily meetings to discuss any questions anyone may have, especially with patients where no match is found.
- Clinic data capture in the ANC is a major priority for all fieldworkers and data clerks. We have made significant headway to collect these data, but we still do not have a single complete book cleaned for analysis!

MATCH PERCENTAGE

Cumulative match percent* among all patients

	CTC	ANC	HTC	TOTAL
CURRENT %	86	84	82	**84**

*(total number matched to at least one HDSS record on first or repeat visits)/(total number claiming to have residence history in HDSS area)

Match percent* by month

	CTC	ANC	HTC	TOTAL
JUNE 2015	85	70	69	77
JULY 2015	90	81	84	84
AUGUST 2015	84	85	86	85
SEPTEMBER 2015	93	95	90	93
OCTOBER 2015	88	92	87	90
NOVEMBER 2015	91	84	90	88
DECEMBER 2015	75	86	80	81
JANUARY 2016	67	91	80	84
FEBRUARY 2016	67	80	85	81
MARCH 2016	100	85	72	83
APRIL 2016	75	91	-	90
MAY 2016	88	82	84	83
JUNE 2016	100	85	86	86
JULY 2016	40	88	89	86
AUGUST 2016	88	81	94	87
SEPTEMBER 2016	86	83	83	83
OCTOBER 2016	100	85	100	92
NOVEMBER 2016	50	85	80	82
DECEMBER 2016	-	85	74	80
JANUARY 2017	80	93	62	73
FEBRUARY 2017	67	86	83	82
MARCH 2017	63	84	71	76
APRIL 2017	89	81	87	85
MAY 2017	100	83	83	84

*(total number matched to at least one HDSS record on first visit only)/(total number claiming to have residence history in HDSS area)

Match percent among individuals claiming no residence history in HDSS area

CUMULATIVE %	10.2

PATIENT FREQUENCY

SAMPLE SIZE

Total number of patients consented

	CTC	ANC	HTC	TOTAL
JUNE 2015	312	211	165	688
JULY 2015	189	248	244	681
AUGUST 2015	132	213	183	528
SEPTEMBER 2015	103	129	168	400
OCTOBER 2015	101	125	142	368
NOVEMBER 2015	93	89	117	299
DECEMBER 2015	66	116	155	337
JANUARY 2016	26	98	146	270
FEBRUARY 2016	41	126	205	372
MARCH 2016	18	130	75	223
APRIL 2016	25	133	0	158
MAY 2016	22	117	146	285
JUNE 2016	16	136	84	236
JULY 2016	25	108	129	262
AUGUST 2016	28	115	165	308
SEPTEMBER 2016	36	124	142	302
OCTOBER 2016	25	94	73	192
NOVEMBER 2016	49	151	12	212
DECEMBER 2016	11	120	124	255
JANUARY 2017	25	99	161	285
FEBRUARY 2017	43	110	104	257
MARCH 2017	58	116	158	332
APRIL 2017	28	63	72	163
MAY 2017	18	76	54	148
	1490	**3047**	**3024**	**7561**

Number of repeat visits (true repeats, even if software thinks it's a new patient)

	CTC	ANC	HTC	TOTAL
JUNE 2015	6	3	1	10
JULY 2015	72	32	15	119
AUGUST 2015	140	49	18	207
SEPTEMBER 2015	148	96	17	261
OCTOBER 2015	276	79	25	380
NOVEMBER 2015	249	114	12	375
DECEMBER 2015	322	97	24	443
JANUARY 2016	181	110	27	318
FEBRUARY 2016	347	143	32	522
MARCH 2016	331	130	18	479
APRIL 2016	254	98	0	352
MAY 2016	383	118	27	528
JUNE 2016	316	162	4	482
JULY 2016	332	121	16	469
AUGUST 2016	500	186	29	715
SEPTEMBER 2016	395	255	26	676
OCTOBER 2016	477	293	6	776
NOVEMBER 2016	810	156	4	970
DECEMBER 2016	535	177	23	735
JANUARY 2017	400	171	34	605
FEBRUARY 2017	597	204	19	820
MARCH 2017	619	193	41	853
APRIL 2017	412	132	34	578
MAY 2017	272	94	8	374
	8374	**3213**	**460**	**12047**

EXCLUSION CRITERIA

Total number of patients claiming never lived in HDSS area

	CTC	ANC	HTC	TOTAL
JUNE 2015	87	4	49	140
JULY 2015	104	40	98	242
AUGUST 2015	78	38	90	206
SEPTEMBER 2015	56	22	82	160
OCTOBER 2015	62	14	78	154
NOVEMBER 2015	50	14	56	120
DECEMBER 2015	32	35	79	146
JANUARY 2016	14	30	70	114
FEBRUARY 2016	22	17	99	138
MARCH 2016	9	30	43	82
APRIL 2016	17	15	0	32
MAY 2016	11	18	72	101
JUNE 2016	11	21	44	76
JULY 2016	14	11	76	101
AUGUST 2016	11	14	100	125
SEPTEMBER 2016	20	18	55	93
OCTOBER 2016	3	13	31	47
NOVEMBER 2016	33	24	2	59
DECEMBER 2016	8	15	47	70
JANUARY 2017	13	13	42	68
FEBRUARY 2017	23	14	46	83
MARCH 2017	24	15	51	90
APRIL 2017	13	4	24	41
MAY 2017	7	6	24	37
	722	**445**	**1358**	**2525**

Total number of patients recently born/moved into HDSS area

	CTC	ANC	HTC	TOTAL
JUNE 2015	8	75	33	116
JULY 2015	3	26	48	77
AUGUST 2015	4	49	37	90
SEPTEMBER 2015	2	26	36	64
OCTOBER 2015	7	35	27	69
NOVEMBER 2015	11	34	22	67
DECEMBER 2015	11	27	22	60
JANUARY 2016	4	15	29	48
FEBRUARY 2016	7	35	29	71
MARCH 2016	3	53	12	68
APRIL 2016	4	51	0	55
MAY 2016	3	44	36	83
JUNE 2016	2	64	18	84
JULY 2016	9	49	22	80
AUGUST 2016	10	59	30	99
SEPTEMBER 2016	9	65	57	131
OCTOBER 2016	14	55	26	95
NOVEMBER 2016	12	86	5	103
DECEMBER 2016	3	67	46	116
JANUARY 2017	7	58	66	131
FEBRUARY 2017	11	53	35	99
MARCH 2017	26	59	69	154
APRIL 2017	6	38	18	62
MAY 2017	8	47	12	67
	184	**1170**	**735**	**2089**

*the searched HDSS database currently extends through 2014

MATCHES

Total number of patients matched during first visit

	CTC	ANC	HTC	TOTAL
JUNE 2015	198	96	60	354
JULY 2015	88	160	92	340
AUGUST 2015	54	123	60	237
SEPTEMBER 2015	49	87	60	196
OCTOBER 2015	38	78	43	159
NOVEMBER 2015	38	44	48	130
DECEMBER 2015	22	59	47	128
JANUARY 2016	7	54	44	105
FEBRUARY 2016	12	69	77	158
MARCH 2016	8	52	18	78
APRIL 2016	5	73	0	78
MAY 2016	7	55	42	104
JUNE 2016	4	58	27	89
JULY 2016	3	58	35	96
AUGUST 2016	7	41	49	97
SEPTEMBER 2016	8	42	39	89
OCTOBER 2016	10	28	24	62
NOVEMBER 2016	3	44	5	52
DECEMBER 2016	1	34	27	62
JANUARY 2017	6	31	45	82
FEBRUARY 2017	6	47	23	76
MARCH 2017	11	44	34	89
APRIL 2017	11	19	28	58
MAY 2017	4	21	15	40
	600	**1417**	**942**	**2959**

*regardless of exclusion criteria met

Number of matches made to each individual

NUMBER OF MATCHES	NUMBER OF INDIVIDUALS
1	2246
2	553
3	132
4	24
5	4
TOTAL	**3864**

QUALITY OF MATCHES MADE/QUALITY OF ALGORITHM

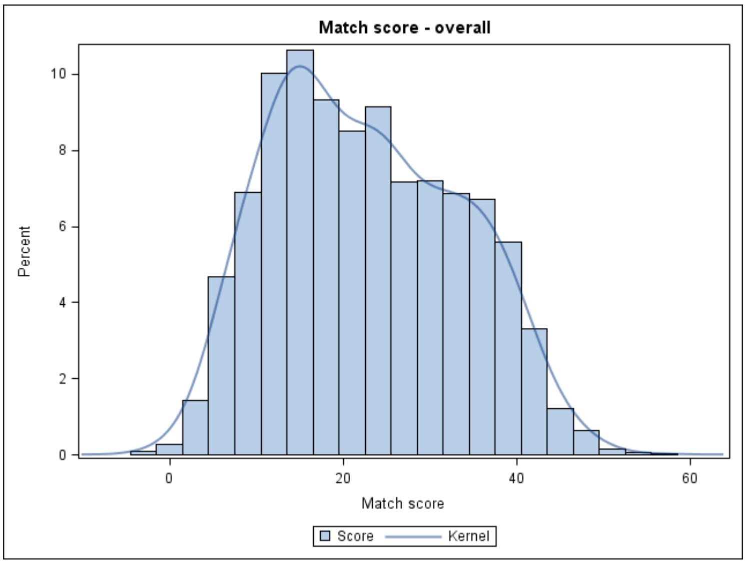
When matches are ranked by score (no gap)

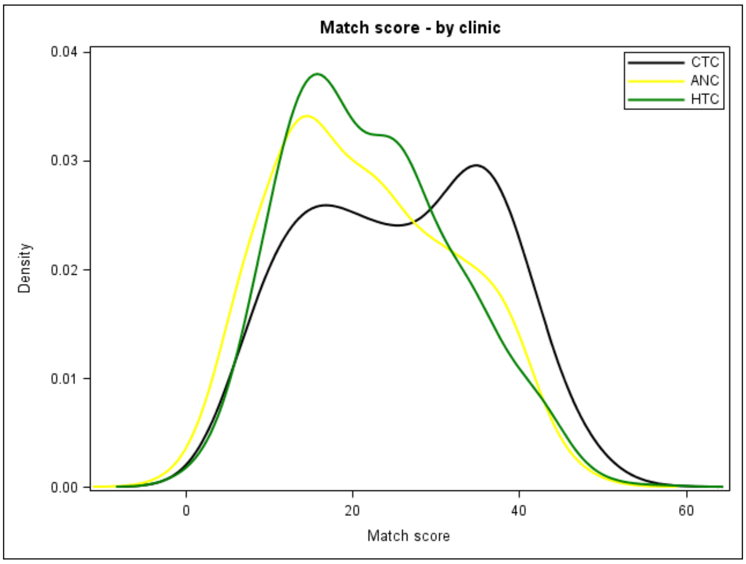
RANK	N (%) OF MATCHES
1	3109 (80%)
2	458 (12%)
3-11	297 (8%)

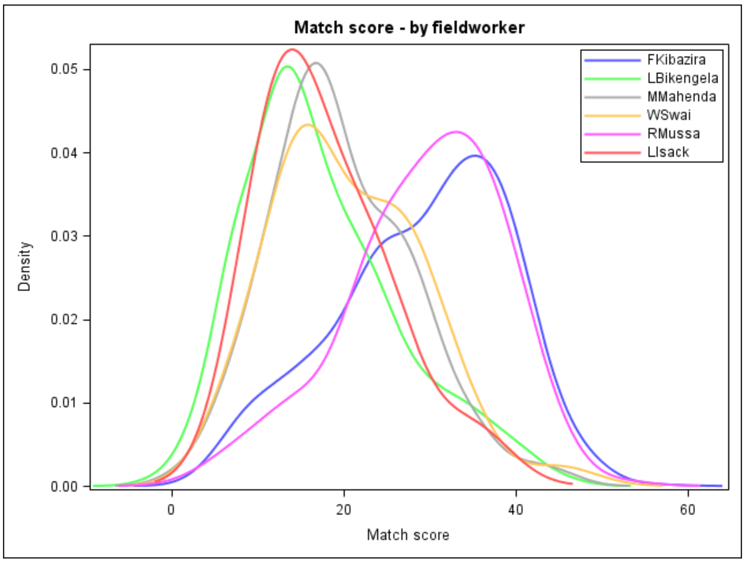
When matches are ranked iteratively

RANK	N (%) OF MATCHES
1	2631 (68%)
2	528 (14%)
3	243 (6%)
4-20	462 (12%)

Match score*







*The higher the match score, the more similar the information you collected matched with the DSS record.

For example, if you typed the name “MASANJA MACHEMBA” into the record linkage software, and you matched to a DSS record with name “MASANJA MACHEMBA”, you will get a perfect match score for names. However, if you incorrectly typed in the name as “MSANGA MACHEBA”, you will get a lower match score. We want to be sure we are asking the patient to spell their name for us before we type it in.

Match score can be made higher by collecting information from the patient exactly how they say it is, including names, birthdate, and residence details.

We know that for many patients, it is impossible to get a perfect match score because they may not know when their birthday is or which sub-village they live in. That is okay! Just try your best to type in everything correctly.

10.6 Fieldworker training agenda

Kisesa-HDSS Record Linkage Training

May 26, 2015

Meeting called by Christopher, Richard, Mtenga

Attendees: Faustine, Lamik, Moses, Winnie, James Beard, Emma, Sero Data Managers

I. Introduction
    a. Reminder of the benefit of record linkage and what we are all working toward

II. Interview
    a. Inquire about residence *history*, not just current living situation
    b. Important questions to ask all patients
        i. Where do you live now?
        ii. Have you ever moved? If so, from and two where around what year? (WRITE THIS DOWN!)
        iii. Have you ever went by any other name? When (what year) did you use that name and where were you living at the time? (WRITE THIS DOWN!)
        iv. Use this information to help guide your search.
        v. This isn't trying to find just one record for each patient. The more successful links we make, the better success we will have showing this project's usefulness!

III. Demo software
    a. Always select Department
    b. MUST CLICK Village/Subvillage. Do not type in ever! Leave blank if claims to be outsider.
    c. Look for household member given
        i. If found, ask if s/he <u>lived with</u> another person on that list. If yes, match!
            1. If not, inquire further into that household
        ii. If not found, ask about 1-2 members from the HH member list
    d. We are trying to find that patient's record. We have to be careful not to select someone else in that same household who may have similar details as the patient sitting in front of you.
    e. If consent, fill in
        i. Clinic ID(s)
        ii. Visit
        iii. Match(es)
    f. Match notes only work if consented and Clinic ID is entered (to store why we CANT find a match).
        i. If a link is made in error, make note to Richard and let him know at end of day!
    g. Show how data tables look on back end
    h. Repeat visits

IV. Consent Process Diagram

a. Greeting (in brief)
   i. A census has been done in the area for over 20 years
   ii. Trying to get a sense of how well people are accessing health services
   iii. So that we can improve health facilities in the area
   iv. We would like to take a few minutes to try and find your records in the census
   v. We will ask questions like name, village, and the name of a household member
   vi. But it is only to help us search. No contact will be made with you or anyone else you mention outside of this clinic

b. Ending the session
   i. Click “End Session/New Patient” on software
   ii. File away consent form
   iii. Prepare forms for next patient

c. Consent forms
   i. Adult vs Child consents
      1. One search session for each of the parent and child
   ii. Patient Information Sheet (PIS)

V. Daily logistics (show Sign Out/In forms, Checklist of materials)
   a. Sign out sheets at NIMR and in field
   b. Return to NIMR
      i. Sign in equipment
      ii. Pass any notes to Richard
      iii. Clear to go when Richard verifies each machine’s collected data

VI. CTC Conversion + Village/subvillage list (show new desktop)

VII. Questions? What can I better explain? What can I explain again?

Notes:

10.7 Evidence of retention of copyright

Gate Open Research (Paper A)

"Gates Open Research articles are published under a CC BY license, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited, and leaves the copyright of the article with the current copyright holder (usually the author or his/her institution)."

Accessed 14 May 2018: https://gatesopenresearch.org/about/policies#licenses

International Journal of Population Data Science (Paper B)

"Authors funded by RCUK / Wellcome Trust can publish their manuscripts in IJPDS using Creative Commons Attribution License (CC-BY). If you require this licence, your request must be indicated clearly in the 'Comments for the Editor' field on the submission form. All other authors will be published using Creative Commons Attribution-Non-Commercial-No-Derivatives 4.0 International Public Licence (CC BY-NC-ND 4.0). Authors retain the copyright to their articles published by IJPDS."

Accessed 14 May 2018: https://ijpds.org/author-guidelines/

NB. Papers C, D, and E are being sought for open-access publication at journals that offer similar licenses obtained for Papers A and B.

10.8 EVALUATION OF PIRL SOFTWARE

Filling out the checklist laid out by Herzog et al.[78]

Question	Responses for PIRL
General	
1. Is the software a generalized system or specific to a given application?	Specific to linking demographic surveillance and medical records in Kisesa, Tanzania
2. What is the form of the software?	A complete system, but can be tailored to other data sources
3. Which types of linkage does the software support?	Simultaneously linking multiple files
4. On which of the following types of computers will the software operate?	Personal computer
5. On which of the following operating systems will the software run?	Windows 7 (tested)
6. For PC-based systems, what is the minimum size of (1) the CPU, (2) RAM, and (3) hard-drive?	(1) Intel Core i5 (4th Gen) 1.9 GHz, (2) 8 GB, (3) 500GB
7. Can the system perform linkages interactively (in real time)? Can it operate in batch mode?	Interactively, not batch mode
8. How fast does the software run on the user's hardware given the expected size of the user's files? If the software is interactive, does it run at an acceptable speed?	10-15 seconds per search on database of 100k records
9. If the software is to be used as part of a statistical analysis system, are the methods used in the software statistically defensible?	Not to be used as part of a statistical analysis system; methods are directly derived from literature
10. Is the software reliable? Can the vendor provide adequate technical support? Is the vendor expected to remain in business through the projected life of the software? If there is doubt about this, is a software escrow available? Is the user prepared to support the software himself/herself?	Published version of software is reliable. The software is available open-source and comes with no technical support.
11. How well is the software documented?	Commented code and a full user guide including install and import/export instructions are published alongside software
12. What additional features docs the vendor plan to make available in then future (i.e., in the next version of the software)?	No current plans to add features
13. Is there a user group? Who else is using the software? What features would they like to see added? Have they developed any custom solutions (e.g., front ends, comparison functions) that they would be willing to share?	This is new software published alongside this thesis. No user groups have been formed at this time.
14. Is other software, such as database package or editor needed to run the system?	SQL Server Management Studio 2012 is required to run this software. Visual Studio is additionally required if any changes are made to the software itself

15. Does the software contain security and data integrity protection features?	Yes, fields are double-entry and have encoded data intregrity checks pertaining to Kisesa data. These can be amended for other purposes
16. How many and what type of staff personnel will be required to develop a system from the software? To run the system? What type of training will they need? Will the vendor provide such training?	Detailed installation instructions have been published. A skilled programmer would be needed to make any changes to the software. No formal training provided by the vendor.
Record Linkage Methodology	
1. Is the record linkage methodology based on (1) Fellegi-Sunter, (2) information-theoretic methods, and/or (3) Bayesian techniques?	Fellegi-Sunter
2. How much control does the user have over the linkage process? Is the system a "black box" or can the user set parameters to control the record linkage process?	Complete control through the source code
3. Does the software require any parameter files? If so, is there a utility provided for generating these files? How effectively does it automate the process? Can the utility be customized?	All parameters can be amended in the source code
4. Does the user specify the linking variables and the types of comparisons?	All linking variables and comparisons can be amended in the source code
5. What kinds of comparison functions are available for different types of variables? Do the methods give proportional weights (i.e., allow degrees of agreement)?	Comparison functions differ based on type of variable. Strings use Jaro-Winkler string comparator, year of birth allows for two-year difference, etc. Comparison functions and agreement conditions can be amended in the source code
6. Can the user specify critical variables that must agree for a link to take place?	This functionality can be added in the source code
7. How does the system handle missing values for linkage variables? Does the system: (1) compute a weight just a it does for any other value? (2) use a median between agreement and disagreement weights? (3) use a weight of zero? (4) allow the user to specify the desired approach?	Uses a weight of zero. This approach can be amended in the source code.
8. Does the system allow array-valued variables (e.g., multiple values for phone number)? How do array-valued comparisons work?	Not currently, but this functionality can be added in the source code
9. What is the maximum number of linking variables?	No built-in limit on number of linking variables, but would need to consider balancing number of linking variables with time the software takes per search
10. Does the software support multiple linkage passes with different blocking and different linkage variables?	Not currently, but this functionality can be added in the source code
11. How does the software block variables? Do users set blocking variables? Can a pass be blocked on more than one variable?	No blocking currently done, but this functionality can be added in the source code

12. Does the software contain or support routines for estimating linkage errors?	Not currently, but this functionality can be added in the source code
Felligi-Sunter Systems	
1. How does the system determine the m- and u-probabilities? Can the user set m- and u-probabilities? Does the software contain utilities that set the m- and u-probabilities?	M-probabilities are hard-coded in the software and can be amended in the source code. U-probabilities are calculated within the software per standard procedures.
2. How does the system determine the weight cut-offs? Can the user set these? Does the software contain any utilities for determining the weight cut-offs?	Weight cut-offs are hard-coded in the software and can be amended in the source code.
3. Does the software permit the user to specify (1) the linkage weights and/or (2) the weights for missing values?	Not currently, but this functionality can be added in the source code.
Data Management	
1. In what file formats can the software use data - (1) flat file, (2) SAS dataset, and/or (3) database? If the answer is "yes" to item (3), what types of database can be employed: Fox, Pro, Informix, Sybase, Oracle, MySQL, DBII?	The software requires a SQL database
2. What is the maximum file size (number of records) that the software can handle?	No built-in limit on file size, but would need to consider balancing file size with time the software takes per search
3. How does the software manage records? Does it use temporary files or sorted files? Does it use pointers? Does it take advantage of database indexing?	Management of records all occurs in SQL Server Management Studio. Within the SQL code, temporary files are created for each search. Permanent files are created with each search attempt, matched record, recorded visit, and saved match notes
4. Can the user specifiy the subsets of the data to be linked?	Not currently, but this functionality can be added in the source code.
5. Does the software provide for "test matches" of a few hundred records to test the specifications?	Not currently, but this functionality can be added in the source code.
6. Does the software contain a utility for viewing and manipulating data records?	Users are permitted to amend the information collected at any time during a linkage session as many times as they wish. Users are not permitted to amend the data in the search database.
Post-linkage Functions	

1. Does the software contain a utility for review of possible links? If so, what kind of functionality is provided for? What kind of interface does the utility use - character-based or GUI? Does the utility allow for review between passes, or only at the end of the process? Can two or more people work on the record review simultaneously? Can records be "put aside" for review later? Is there any provision for adding comments to the reviewed records pairs in the form of hypertext? Can pairs of groups of records be update? Can the user "back up" or restore possible links before committing to decisions? Can a "master" record be created that combines values from two or more records for different fields?	The software outputs the top 20 most likely matches based on the highest match score, from which the user can select which records belong to the individual. After selecting each potential match, the user will be presented with the full list of names associated with living at the same household at that time. The prospective nature of this software only allows for one person to work on record review at a time. Records cannot be "put aside" for review later. There is a match notes field available for users to input any text they wish to save off an be associated with a particular linkage session. Records can be updated in the back-end data. The user cannot restore possible links before committing to a decision. A "master" record can be created that combines values from two or more records for different fields in the back-end data, but not within the software itself.
2. Does the software provide for results of earlier linkages (particularly reviews of possible links) to be applied to the current linkage process?	Yes, a flag denoting an individual has been previously matches is shown to the user if the individiual has been seen before.
3. Does the software contain a utility for generating reports on the linked, unlinked, duplicate, and possible linked records? Can the format of the report be customised? Is the report viewed in character mode or is the report review done in a graphical environment? Can the report be printed? If so, what type of printer is required?	Not currently, but this functionality can be added in the source code.
4. Does the software contain a utility that extracts files of linked and unlinked records? Can the user specify the formats of such extracts?	Not currently, but this functionality can be added in the source code.
5. Does the software generate statistics for evaluating the linkage process? Can the user customise the statistics generated by the system?	Not currently, but this functionality can be added in the source code.
Standardisation	
1. Does the software permit the user to partition variables in order to maximize the use of the information contained in these variables? For example, can a telephone number be partitioned into its (1) area code, (2) exchange, and (3) last four digits?	Not currently, but this functionality can be added in the source code.
2. Can the name and address standardization/parsing components be customized? Can different processes be applied to different files?	They can be customised in the source code.
3. Does the address standardization conform to US Postal Service standards?	N/A

4. Does the standardization modify the original data fields, or does it append standardized fields to the original data record?	N/A
5. How well do the standardization components work on the types of names the user wishes to link? For example, does the standardizer work well with Hispanic name? What about Asian names?	No standardisation is done on names in the software. The software uses a Jaro-Winkler string comparator for names, which has been previously shown to work well in an eastern/southern African context.
6. How well do the standardization components work on the types of addresses (e.g., rural or foreign) that the user expects to encounter?	N/A
Costs	
1. What are (l) the purchase price and (2) the annual maintenance cost of (i) the basic software system, (ii) additional features (e.g., database packages), and (iii) new or upgraded computer hardware?	Costs associated with implementing this software are: personnel (1 data manager + 1 user for each location of linkage), personal computers (1 for data manager + 1 for each user)
2. What is the cost of training staff to use the system?	Staff were trained in an office setting over two days by a PhD student overseeing the implementation of the software, followed by two months of oversight in the field.
3. What are the estimated personnel and (in the case of mainframe systems) computer-time costs associated with running the system?	Costs will be dependent on the setting. In Kisesa, Tanzania, data manager was paid $360 USD/month and users were paid $180 USD/month. The data manager machine costed £920. User machines costed £600/unit
4. Is the cost of developing a new system for the intended purpose using the software within the available budget?	To be determined by end-users
5. What is the upgrade path for the software? What will upgrades cost?	No planned upgrades made by LSHTM staff. All upgrades to be determined by end-users
6. What kind of maintenance/support agreements is available? What do they cost?	No maintenance/support provided by LSHTM staff.

10.9 SUPPLEMENTARY MATERIAL FROM PUBLICATIONS

10.9.1 Paper B

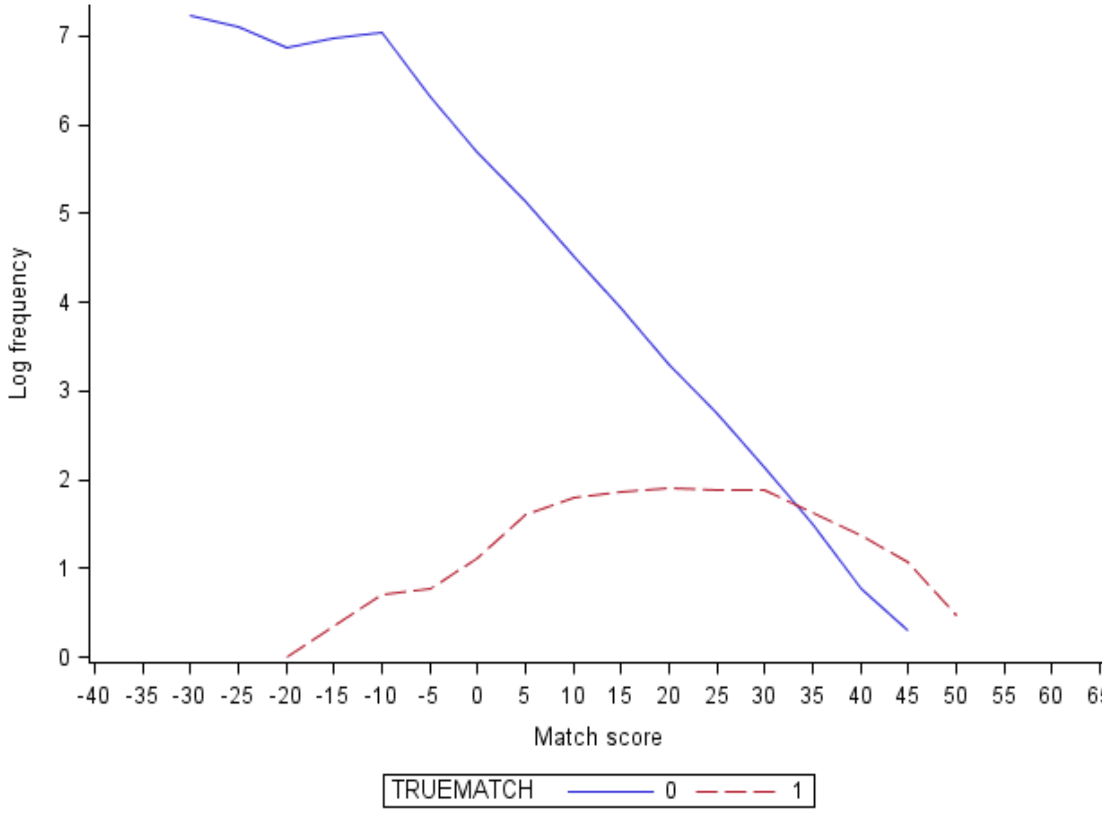
Supplemental Table 1. Agreement conditions, match (*m*) probabilities, proportion collected, and proportion of records with agreement for each field (*i*) in the probabilistic algorithm, by first and matched search attempts, n_M=2,612

			First search		Matched search		Change (Δ)=matched-first	
Field *i*	**Agreement condition**	***m*-prob**	**% collected**	**% agreement**	**% collected**	**% agreement**	**Δ% collected**	**Δ% agreement**
First name	Jaro-Winkler ≥ 0.8	0.87	100.0%	83.8%	100.0%	94.1%	0.0%	10.3%
Second name	Jaro-Winkler ≥ 0.8	0.87	100.0%	77.9%	100.0%	87.9%	0.0%	10.1%
Third name	Jaro-Winkler ≥ 0.8	0.85	83.4%	5.7%	82.0%	5.3%	-1.4%	-0.3%
TCL first name	Jaro-Winkler ≥ 0.8	0.87	44.8%	15.1%	65.8%	42.9%	20.9%	27.8%
TCL second name	Jaro-Winkler ≥ 0.8	0.87	39.4%	13.6%	60.8%	40.9%	21.5%	27.3%
TCL third name	Jaro-Winkler ≥ 0.8	0.85	0.2%	0.0%	0.2%	0.2%	0.0%	0.1%
HH first name	Jaro-Winkler ≥ 0.8	0.52	90.5%	70.1%	93.2%	75.2%	2.7%	5.1%
HH second name	Jaro-Winkler ≥ 0.8	0.52	89.6%	64.3%	92.2%	70.8%	2.6%	6.5%
HH third name	Jaro-Winkler ≥ 0.8	0.52	4.1%	1.1%	4.4%	1.1%	0.3%	0.0%
Sex	exact match	0.99	99.8%	97.6%	99.8%	97.7%	0.0%	0.1%
Year of birth	within 2 years	0.80	98.7%	84.9%	99.1%	87.0%	0.4%	2.1%
Month of birth	exact match	0.63	3.7%	1.4%	4.0%	1.6%	0.3%	0.2%
Day of birth	exact match	0.57	3.6%	1.0%	3.9%	1.2%	0.3%	0.2%
Village	exact match	0.89	90.9%	83.3%	93.0%	89.4%	2.1%	6.1%
Sub-village	exact match	0.89	90.9%	67.2%	93.0%	78.0%	2.1%	10.8%

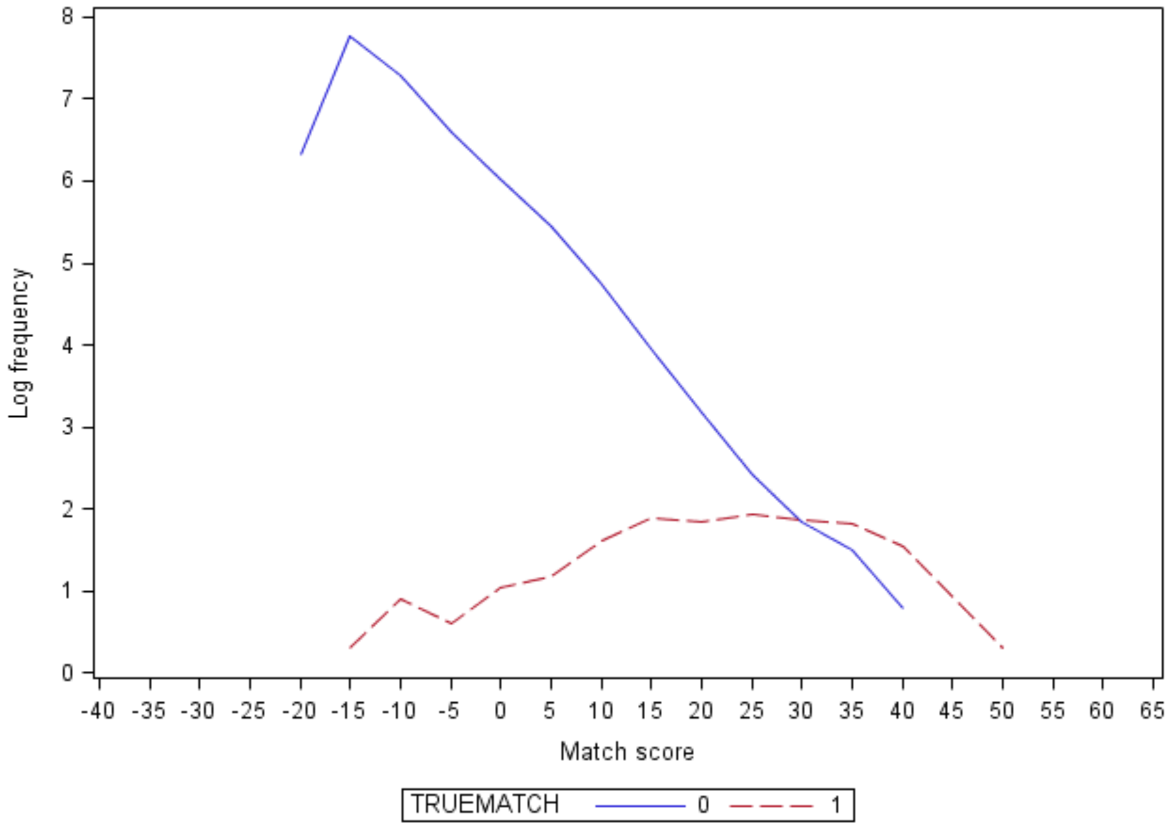
Abbreviations: HDSS = health and demographic surveillance surveys; n_M = number of matches; *m*-prob = match probability; TCL = ten-cell leader; HH = household member

Notes: TCL = an individual for a group of ten households; % collected = proportion of matched records with completed information; % agreement = proportion of matched records with agreeing information

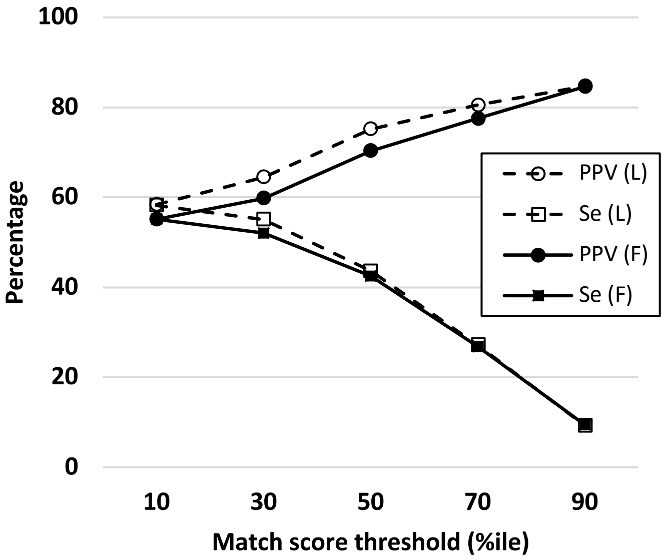
Supplemental Figure 1. Log frequency of match scores calculated for all pairwise comparisons using full algorithm, by true match status



Supplemental Figure 2. Log frequency of match scores calculated for all pairwise comparisons using limited algorithm, by true match status



Supplemental Figure 3. Sensitivity (Se) and positive predictive value (PPV) of automated retrospective record linkage at various match score percentile thresholds, full (F) vs. limited (L) algorithm



Supplemental Table 2. Distribution of individual characteristics in the dataset matched using point-of-contact interactive record linkage (PIRL) compared with those matched using a purely automated probabilistic approach using a full and limited algorithm, by match score threshold

		Automated: full algorithm						**Automated: limited algorithm**					
	PIRL match	**Threshold=10%ile**		**Threshold=50%ile**		**Threshold=90%ile**		**Threshold=10%ile**		**Threshold=50%ile**		**Threshold=90%ile**	
Characteristic	**n (%)**	**n (%)**	***p*-value***	**n (%)**	***p*-value***	**n (%)**	***p*-value***	**n (%)**	***p*-value***	**n (%)**	***p*-value***	**n (%)**	***p*-value***
Total matched (PPV)	2,612	2,612 (55.1)		1,579 (70.3)		292 (84.6)		2602 (58.4)		1,514 (75.2)		288 (84.7)	
Sex													
Female	2,061 (78.9)	2,036 (78.0)	0.4004	1,185 (75.1)	0.0038	213 (73.0)	0.0191	2,059 (79.1)	0.8409	1,158 (76.5)	0.0706	209 (72.6)	0.0133
Male	551 (21.1)	576 (22.1)		394 (25.0)		79 (27.1)		543 (20.9)		356 (23.5)		79 (27.4)	
Age, in years													
<5	125 (4.8)	198 (7.6)	<0.0001	132 (8.4)	<0.0001	46 (15.8)	<0.0001	198 (7.6)	<0.0001	122 (8.1)	0.0013	33 (11.5)	<0.0001
5-17	393 (15.1)	464 (17.8)		239 (15.2)		35 (12.0)		453 (17.4)		211 (14.0)		34 (11.8)	
18-34	1,384 (53.0)	1,301 (49.9)		770 (48.8)		125 (42.8)		1,325 (51.0)		765 (50.6)		121 (42.0)	
35-49	522 (20.0)	433 (16.6)		301 (19.1)		68 (23.3)		437 (16.8)		296 (19.6)		74 (25.7)	
50-64	160 (6.1)	162 (6.2)		105 (6.7)		15 (5.1)		144 (5.5)		99 (6.5)		23 (8.0)	
65+	28 (1.1)	52 (2.0)		30 (1.9)		3 (1.0)		43 (1.7)		20 (1.3)		3 (1.0)	
Village of residence													

Supplemental Table 2. Distribution of individual characteristics in the dataset matched using point-of-contact interactive record linkage (PIRL) compared with those matched using a purely automated probabilistic approach using a full and limited algorithm, by match score threshold

		Automated: full algorithm						**Automated: limited algorithm**					
	PIRL match	**Threshold=10%ile**		**Threshold=50%ile**		**Threshold=90%ile**		**Threshold=10%ile**		**Threshold=50%ile**		**Threshold=90%ile**	
Characteristic	**n (%)**	**n (%)**	***p*-value***	**n (%)**	***p*-value***	**n (%)**	***p*-value***	**n (%)**	***p*-value***	**n (%)**	***p*-value***	**n (%)**	***p*-value***
Kisesa	999 (38.3)	982 (37.6)	0.9340	586 (37.1)	0.8100	111 (38.0)	0.3320	981 (37.7)	0.6773	531 (35.1)	0.3071	73 (25.4)	0.0002
Kanyama	521 (20.0)	529 (20.3)		302 (19.1)		46 (15.8)		527 (20.3)		299 (19.8)		59 (20.5)	
Kitumba	424 (16.2)	444 (17.0)		262 (16.6)		48 (16.4)		436 (16.8)		254 (16.8)		49 (17.0)	
Isangijo	257 (9.8)	258 (9.9)		176 (11.2)		39 (13.4)		254 (9.8)		177 (11.7)		46 (16.0)	
Ihayabuyaga	152 (5.8)	138 (5.3)		89 (5.6)		21 (7.2)		129 (5.0)		87 (5.8)		22 (7.6)	
Igekemaja	141 (5.4)	150 (5.7)		94 (6.0)		13 (4.5)		163 (6.3)		94 (6.2)		24 (8.3)	
Welamasonga	118 (4.5)	111 (4.3)		70 (4.4)		14 (4.8)		112 (4.3)		72 (4.8)		15 (5.2)	
Marital status[a]													
Never married	362 (24.0)	272 (24.1)	0.9997	179 (22.5)	0.4266	33 (22.3)	0.6089	286 (25.3)	0.8668	176 (22.5)	0.7093	26 (16.5)	0.0139
Married once	724 (48.0)	540 (47.8)		403 (50.6)		72 (48.7)		536 (47.4)		391 (49.9)		80 (50.6)	
Remarried	175 (11.6)	132 (11.7)		99 (12.4)		22 (14.9)		124 (11.0)		95 (12.1)		30 (19.0)	

Supplemental Table 2. Distribution of individual characteristics in the dataset matched using point-of-contact interactive record linkage (PIRL) compared with those matched using a purely automated probabilistic approach using a full and limited algorithm, by match score threshold

		Automated: full algorithm						Automated: limited algorithm					
	PIRL match	Threshold=10%ile		Threshold=50%ile		Threshold=90%ile		Threshold=10%ile		Threshold=50%ile		Threshold=90%ile	
Characteristic	**n (%)**	**n (%)**	***p*-value***	**n (%)**	***p*-value***	**n (%)**	***p*-value***	**n (%)**	***p*-value***	**n (%)**	***p*-value***	**n (%)**	***p*-value***
Separated/Widowed	249 (16.5)	187 (16.5)		116 (14.6)		21 (14.2)		185 (16.4)		121 (15.5)		22 (13.9)	
Pregnant at last HDSS round[b]													
No	1,057 (95.7)	758 (95.5)	0.8425	529 (95.0)	0.5292	101 (98.1)	0.3094	769 (95.2)	0.6166	531 (95.5)	0.8862	93 (92.1)	0.1310
Yes	48 (4.3)	36 (4.5)		28 (5.0)		2 (1.9)		39 (4.8)		25 (4.5)		8 (7.9)	
Enrolled in school at last HDSS round[c]													
No	378 (72.0)	282 (67.6)	0.1454	185 (68.3)	0.2725	25 (52.1)	0.0038	295 (67.7)	0.1438	186 (69.9)	0.5422	21 (60.0)	0.1288
Yes	147 (28.0)	135 (32.4)		86 (31.7)		23 (47.9)		141 (32.3)		80 (30.1)		14 (40.0)	

Abbreviations: HDSS - health and demographic sentinel surveillance

*Statistical differences tested for significance with chi-square (χ^2) or Fisher's Exact tests

[a] This question was only given to individuals aged 15 years or older

[b] This question was only given to females between 15 and 49 years of age

[c] This question was only given to individuals between 5 and 25 years of age

10.9.2 Paper D

Supplemental Table 1. Characteristics among participants of population-based HIV serological surveys in Magu, Tanzania who tested HIV-negative, by whether they had evidence of a previous diagnostic HIV test

	Sero 6 (2010)			**Sero 7 (2013)**			**Sero 8 (2016)**		
	No previous test	**Previous test**		**No previous test**	**Previous test**		**No previous test**	**Previous test**	
	n=6,729	**n=722**	**p-value**	**n=6,040**	**n=1,010**	**p-value**	**n=5,160**	**n=1,511**	**p-value**
Demographic characteristic									
Sex									
Female	4,075 (61)	413 (57)	0.0544	3,682 (61)	598 (59)	0.2730	3,138 (61)	956 (63)	0.1023
Male	2,609 (39)	308 (43)		2,328 (39)	408 (41)		1,997 (39)	551 (37)	
Age, years									
15-29	3,743 (56)	200 (28)	<0.0001	3,284 (54)	216 (21)	<0.0001	2,845 (55)	385 (25)	<0.0001
30-49	1,743 (26)	389 (54)		1,551 (26)	535 (53)		1,248 (24)	696 (46)	
50+	1,243 (18)	133 (18)		1,205 (20)	259 (26)		1,067 (21)	430 (28)	
Education level									
No primary	1,955 (29)	147 (20)	<0.0001	1,752 (29)	236 (23)	0.0002	1,380 (27)	365 (24)	0.1285
Some primary	1,109 (16)	93 (13)		786 (13)	122 (12)		571 (11)	170 (11)	
Primary or higher	3,665 (54)	482 (67)		3,502 (58)	652 (65)		3,209 (62)	976 (65)	
Sub-village of residence, type									
Rural	3,885 (58)	418 (58)	0.0352	3,877 (64)	493 (49)	<0.0001	2,817 (55)	767 (51)	0.0030
Peri-urban	1,415 (21)	175 (24)		1,136 (19)	276 (27)		1,092 (21)	380 (25)	
Urban	1,429 (21)	129 (18)		1,027 (17)	241 (24)		1,251 (24)	364 (24)	
Sub-village of residence, has road									

Supplemental Table 1. Characteristics among participants of population-based HIV serological surveys in Magu, Tanzania who tested HIV-negative, by whether they had evidence of a previous diagnostic HIV test

	Sero 6 (2010)			**Sero 7 (2013)**			**Sero 8 (2016)**		
	No previous test	**Previous test**		**No previous test**	**Previous test**		**No previous test**	**Previous test**	
	n=6,729	**n=722**	**p-value**	**n=6,040**	**n=1,010**	**p-value**	**n=5,160**	**n=1,511**	**p-value**
No	4,277 (64)	457 (63)	0.8885	4,187 (69)	562 (56)	<0.0001	3,062 (59)	849 (56)	0.0286
Yes	2,452 (36)	265 (37)		1,853 (31)	448 (44)		2,098 (41)	662 (44)	
Current marital status									
Never married/cohabitated	2,453 (36)	86 (12)	<0.0001	2,272 (38)	120 (12)	<0.0001	2,067 (40)	192 (13)	<0.0001
Ever married/cohabitated	4,276 (64)	636 (88)		3,768 (62)	890 (88)		3,093 (60)	1,319 (87)	
Behavioural characteristic									
Number of sex partners in last 12 months									
Don't know/refused	1,446 (21)	14 (2)	<0.0001	1,382 (23)	31 (3)	<0.0001	1,342 (26)	62 (4)	<0.0001
0	1,006 (15)	82 (11)		926 (15)	122 (12)		661 (13)	190 (13)	
1	3,741 (56)	489 (68)		3,343 (55)	741 (73)		2,880 (56)	1,135 (75)	
2 or more	536 (8)	137 (19)		389 (6)	116 (11)		277 (5)	124 (8)	
Condom use at last sex									
Don't know	2,454 (36)	95 (13)	<0.0001	4,852 (80)	792 (78)	0.2158	1,765 (34)	167 (11)	<0.0001
No	3,910 (58)	581 (80)		987 (16)	175 (17)		3,162 (61)	1,270 (84)	
Yes	365 (5)	46 (6)		201 (3)	43 (4)		233 (5)	74 (5)	
Clinical characteristic									

Supplemental Table 1. Characteristics among participants of population-based HIV serological surveys in Magu, Tanzania who tested HIV-negative, by whether they had evidence of a previous diagnostic HIV test

	Sero 6 (2010)			**Sero 7 (2013)**			**Sero 8 (2016)**		
	No previous test	**Previous test**		**No previous test**	**Previous test**		**No previous test**	**Previous test**	
	n=6,729	**n=722**	**p-value**	**n=6,040**	**n=1,010**	**p-value**	**n=5,160**	**n=1,511**	**p-value**
Visited health provider in last 12 months									
No	1,102 (16)	96 (13)	0.0322	1,226 (20)	166 (16)	0.0043	1,618 (31)	422 (28)	0.0110
Yes	5,627 (84)	626 (87)		4,814 (80)	844 (84)		3,541 (69)	1,089 (72)	

Abbreviations: HIV - human immunodeficiency virus; sero - HIV serological survey

Note: all statistics are given in n(row %); differences tested for significance with chi-square (χ^2) and Fisher's exact tests

Supplemental Table 2. Characteristics among participants of population-based HIV serological surveys in Magu, Tanzania who tested HIV-positive, by whether they had evidence of a previous diagnostic HIV test

	Sero 6 (2010)			**Sero 7 (2013)**			**Sero 8 (2016)**		
	No previous test n=392	**Previous test n=138**	**p-value**	**No previous test n=335**	**Previous test n=222**	**p-value**	**No previous test n=215**	**Previous test n=275**	**p-value**
Demographic characteristic									
Sex									
Female	269 (69)	97 (71)	0.5562	219 (66)	164 (74)	0.0483	147 (69)	195 (71)	0.5957
Male	123 (31)	39 (29)		113 (34)	58 (26)		67 (31)	80 (29)	
Age, years									
15-29	120 (31)	30 (22)	0.1177	87 (26)	48 (22)	0.4379	49 (23)	35 (13)	0.0012
30-49	216 (55)	83 (60)		188 (56)	128 (58)		125 (58)	156 (57)	
50+	56 (14)	25 (18)		60 (18)	46 (21)		41 (19)	84 (31)	
Education level									
No primary	135 (34)	48 (35)	0.5940	130 (39)	74 (33)	0.4197	92 (43)	108 (39)	0.6587
Some primary	62 (16)	17 (12)		37 (11)	26 (12)		24 (11)	29 (11)	
Primary or higher	195 (50)	73 (53)		168 (50)	122 (55)		99 (46)	138 (50)	
Sub-village of residence, type									
Rural	208 (53)	56 (41)	0.0382	194 (58)	101 (46)	0.0049	111 (52)	131 (48)	0.3348
Peri-urban	99 (25)	42 (30)		66 (20)	68 (31)		44 (20)	72 (26)	
Urban	85 (22)	40 (29)		75 (22)	53 (24)		60 (28)	72 (26)	
Sub-village of residence, has road									

Supplemental Table 2. Characteristics among participants of population-based HIV serological surveys in Magu, Tanzania who tested HIV-positive, by whether they had evidence of a previous diagnostic HIV test

	Sero 6 (2010)			Sero 7 (2013)			Sero 8 (2016)		
	No previous test n=392	Previous test n=138	p-value	No previous test n=335	Previous test n=222	p-value	No previous test n=215	Previous test n=275	p-value
No	237 (60)	75 (54)	0.2096	221 (66)	131 (59)	0.0953	123 (57)	156 (57)	0.9148
Yes	155 (40)	63 (46)		114 (34)	91 (41)		92 (43)	119 (43)	
Current marital status									
Never married/cohabitated	34 (9)	7 (5)	0.1733	41 (12)	20 (9)	0.2321	23 (11)	22 (8)	0.3048
Ever married/cohabitated	358 (91)	131 (95)		294 (88)	202 (91)		192 (89)	253 (92)	
Behavioural characteristic									
Number of sex partners in last 12 months									
Don't know/refused	15 (4)	4 (3)	0.5641	7 (2)	6 (3)	0.2332	7 (3)	10 (4)	0.3461
0	63 (16)	29 (21)		59 (18)	54 (24)		47 (22)	78 (28)	
1	261 (67)	89 (64)		239 (71)	146 (66)		148 (69)	168 (61)	
2 or more	53 (14)	16 (12)		30 (9)	16 (7)		13 (6)	19 (7)	
Condom use at last sex									
Don't know	78 (20)	32 (23)	0.6910	264 (79)	174 (78)	0.9899	39 (18)	57 (21)	0.6070
No	287 (73)	96 (70)		61 (18)	41 (18)		164 (76)	199 (72)	
Yes	27 (7)	10 (7)		10 (3)	7 (3)		12 (6)	19 (7)	
Clinical characteristic									

Supplemental Table 2. Characteristics among participants of population-based HIV serological surveys in Magu, Tanzania who tested HIV-positive, by whether they had evidence of a previous diagnostic HIV test

	Sero 6 (2010)			**Sero 7 (2013)**			**Sero 8 (2016)**		
	No previous test	**Previous test**		**No previous test**	**Previous test**		**No previous test**	**Previous test**	
	n=392	**n=138**	**p-value**	**n=335**	**n=222**	**p-value**	**n=215**	**n=275**	**p-value**
Visited health provider in last 12 months									
No	49 (13)	10 (7)	0.0915	49 (15)	25 (11)	0.2519	59 (27)	63 (23)	0.2495
Yes	343 (88)	128 (93)		286 (85)	197 (89)		156 (73)	212 (77)	

Abbreviations: HIV - human immunodeficiency virus; sero - HIV serological survey

Note: all statistics are given in n(row %); differences tested for significance with chi-square (χ^2) and Fisher's exact tests