International Journal of Population Data Science

Journal Website: www.ijpds.org

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], and Basia Zaba[1]

Submission History

Submitted:	07/07/2017
Accepted:	07/11/2017
Published:	15/12/2017

[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 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

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.

Objective
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 up-take 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.

Keywords
data linkage, point-of-contact interactive record linkage, health and demographic surveillance systems, health facility, sub-Saharan Africa

Highlights

- Linking health and demographic surveillance systems (HDSS) to data from a health facility that serves the surveillance population provides a unique opportunity to produce a nascent research infrastructure for better monitoring access to and utilization of health services.
- We implemented our Point-of-contact Interactive Record Linkage (PIRL) software and linked 84% of the individuals who reported residence in the coverage area to one or more of their HDSS records. Characteristics associated with a higher odds of being matched were increased age, a later consent into the study, and fieldworker level of experience.

*Corresponding Author:
Email Address: Christopher.Rentsch@lshtm.ac.uk (CT. Rentsch)

http://dx.doi.org/10.23889/ijpds.v2i1.408

- The main drivers of the linkage algorithm were name, sex, year of birth, village, sub-village, and household member name.
- Automated retrospective linkage would have only correctly identified and linked 55% (1440/2612) of the records with a PPV of 55% (1440/2612).
- Where resources are available, PIRL is a viable approach to link HDSS and other administrative data sources that outperforms purely retrospective approaches.

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[1]. 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[2] and probabilistic[3-5], 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[6]. 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[7].

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.

Methods

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.

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, field-worker 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.

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)[8].

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[9, 10], with match probabilities (m_i) that have been adopted from a similar study in the Agincourt HDSS[7].

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:

$$\text{match probability: } m_i = P(\text{field } i \text{ agrees } |j \epsilon M)$$

$$\text{unmatch probability: } u_i = P(\text{field } i \text{ agrees } |j \epsilon U)$$

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 $w_{di} = \log_2[(1-m_i)/(1-u_i)]$ 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[10, 11]. 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). 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[12]. Previous research has shown the Jaro-Winkler method produces similar results to Double Metaphone and Soundex string comparators in a southern African context[7].

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.

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.

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, the fieldworker cannot access the collected 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.

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 subvillage 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.

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[5, 11, 13-15]. 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[16]. 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). 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 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)[5, 15]. 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.

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 Medicine (LSHTM #8852).

Results

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

Figure 1: Classification diagram of record linkage outcomes against true match status

		True match status		
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

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)

Table 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

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 1). Thus, 2,624 patients reported residence history in the HDSS area and were considered likely to have a record in the community database.

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 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 per-centages 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).

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 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.

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

Table 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												
June - August 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
September - December 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)	
January - June 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)	
July - December 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

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).

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 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). 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.

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). We calculated the sensitivity and PPV of the full algorithm at 10^{th}-, 30^{th}-, 50^{th}-, 70^{th}-, and 90^{th}-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 3).

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). 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).

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.

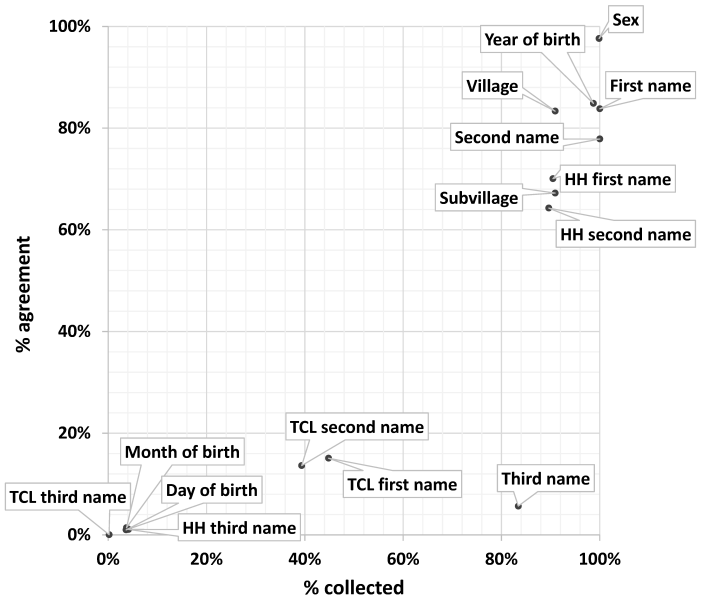
During the study period, we had no refusals to provide informed written consent from clinic attendees who agreed to sit down with a fieldworker. We believe a more likely approach

Table 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				
June - August 2015	1	1	1	1
September - December 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)
January - June 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)
July - December 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: 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

Figure 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

Figure 3: Sensitivity (Se) and positive predictive value (PPV) of automated retrospective record linkage at various match score percentile thresholds, full algorithm

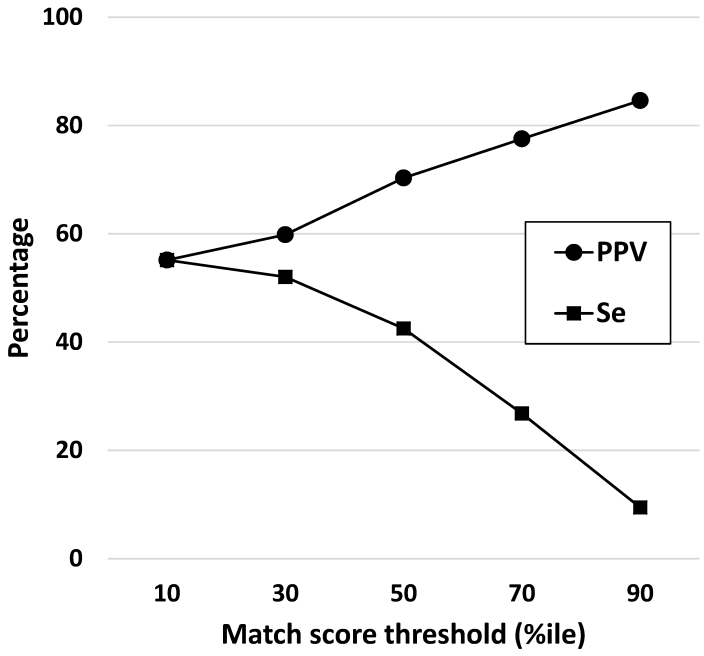


Table 4: Distribution of individual characteristics in the dataset matched using point-of-contact interactive record linkage (PIRL) compared with those matched using a purely automated probabilistic approach using the full algorithm, by match score threshold

		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)	
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 ($\chi 2$) 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

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[7]. 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[15]. 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.

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[17]. 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%[18]. 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, subvillage, 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 cost-effective solution for smaller-scale research projects where data quality is a principal concern.

Conclusion

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.

Acknowledgements

The authors thank David Beckles and Jason Catlett for providing technical support for the development of the PIRL software, the field team for conducting the interviews and data collection, and the participating communities. 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].

Conflicts of Interest

None declared

Supplementary Appendices

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

References

1. Sankoh O, Network I. CHESS: an innovative concept for a new generation of population surveillance. Lancet Glob Health. 2015;3(12):e742.
2. 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.
3. Jaro MA. Probabilistic linkage of large public health data files. Stat Med. 1995;14:491-8.
4. 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. J Clin Epidemiol. 2007;60(9):883-91.
5. Sayers A, Ben-Shlomo Y, Blom AW, Steele F. Probabilistic record linkage. Int J Epidemiol. 2015.
6. 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.
7. 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).
8. Kabudula C, Rentsch C, Catlett J, Beckles D, Masilela N, Zaba B, et al. PIRL - Point-of-contact Interactive Record Linkage software. https://doi.org/10.5281/zenodo.998867; 2017.
9. Newcombe H, Kennedy J, Axford S, James A. Automatic Linkage of Vital Records. Science. 1959;130(3381):954-9.
10. Fellegi IP, Sunter AB. A Theory for Record Linkage. J Am Stat Assoc. 1969;64(328):1183-210.
11. Herzog TN, Scheuren FJ, Winkler WE. Data quality and record linkage techniques: Springer Science & Business Media; 2007.
12. Winkler WE. String Comparator Metrics and Enhanced Decision Rules in the Fellegi-Sunter Model of Record Linkage. 1990.
13. Winkler WE. Overview of Record Linkage and Current Research Directions. Washington, DC: US Bureau of the Census; 2006.
14. Christen P. Data matching: concepts and techniques for record linkage, entity resolution, and duplicate detection: Springer Science & Business Media; 2012.
15. Harron K, Goldstein H, Dibben C. Methodological developments in data linkage: John Wiley & Sons; 2015.

16. Newcombe H. Strategy and art in automated death searches. Am J Public Health. 1984;74(12).

17. 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.

18. 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. Trop Med Int Health. 2015.

Abbreviations

AIC	Akaike information criterion
ANC	antenatal clinic
CI	confidence interval
CTC	HIV care and treatment centre
HDSS	health and demographic surveillance system
HTC	HIV testing and counselling clinic
IQR	interquartile range
LSHTM	London School of Hygiene & Tropical Medicine
OR	odds ratio
PIRL	point-of-contact interactive record linkage
PMTCT	prevention of mother-to-child transmission
PPV	positive predictive value

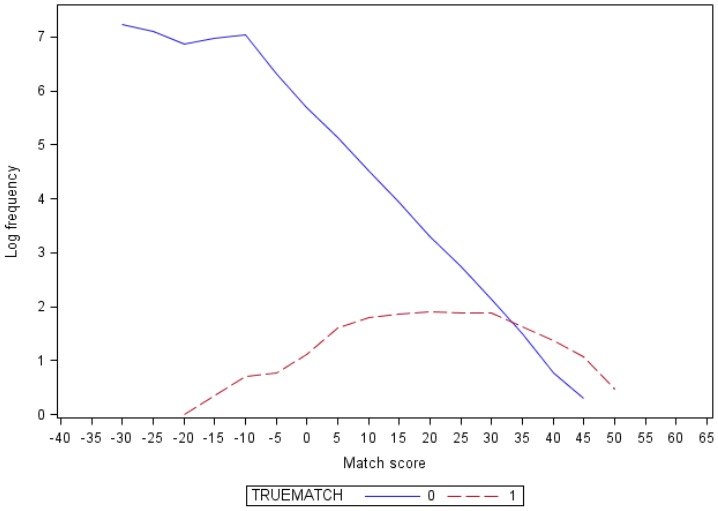
Supplemental Table 1: Agreement conditions, match (m) probabilities, proportion collected, and proportion of records with agreement for each field (i) in the probabilistic algorithm, by first and matched search attempts, n_M=2,612

			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 $\geq$ 0.8	0.87	100.0%	83.8%	100.0%	94.1%	0.0%	10.3%
Second name	Jaro-Winkler $\geq$ 0.8	0.87	100.0%	77.9%	100.0%	87.9%	0.0%	10.1%
Third name	Jaro-Winkler $\geq$ 0.8	0.85	83.4%	5.7%	82.0%	5.3%	-1.4%	-0.3%
TCL first name	Jaro-Winkler $\geq$ 0.8	0.87	44.8%	15.1%	65.8%	42.9%	20.9%	27.8%
TCL second name	Jaro-Winkler $\geq$ 0.8	0.87	39.4%	13.6%	60.8%	40.9%	21.5%	27.3%
TCL third name	Jaro-Winkler $\geq$ 0.8	0.85	0.2%	0.0%	0.2%	0.2%	0.0%	0.1%
HH first name	Jaro-Winkler $\geq$ 0.8	0.52	90.5%	70.1%	93.2%	75.2%	2.7%	5.1%
HH second name	Jaro-Winkler $\geq$ 0.8	0.52	89.6%	64.3%	92.2%	70.8%	2.6%	6.5%
HH third name	Jaro-Winkler $\geq$ 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%

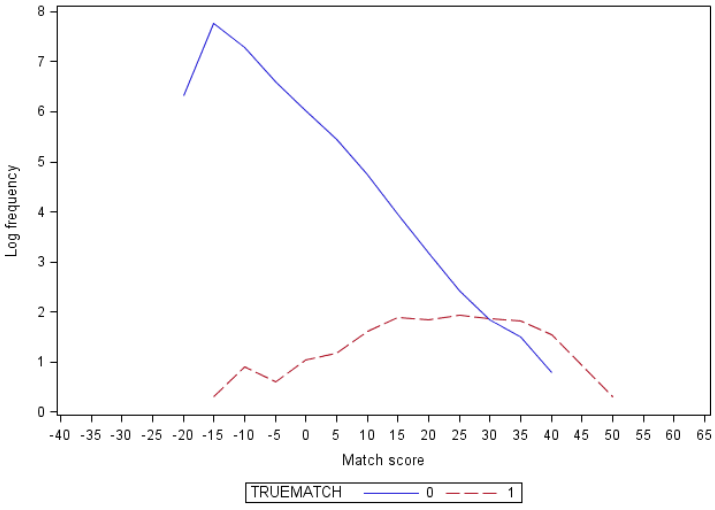
Abbreviations: HDSS = health and demographic surveillance surveys; n_M = number of matches; m-prob = match probability; TCL = ten-cell leader; HH = household member

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

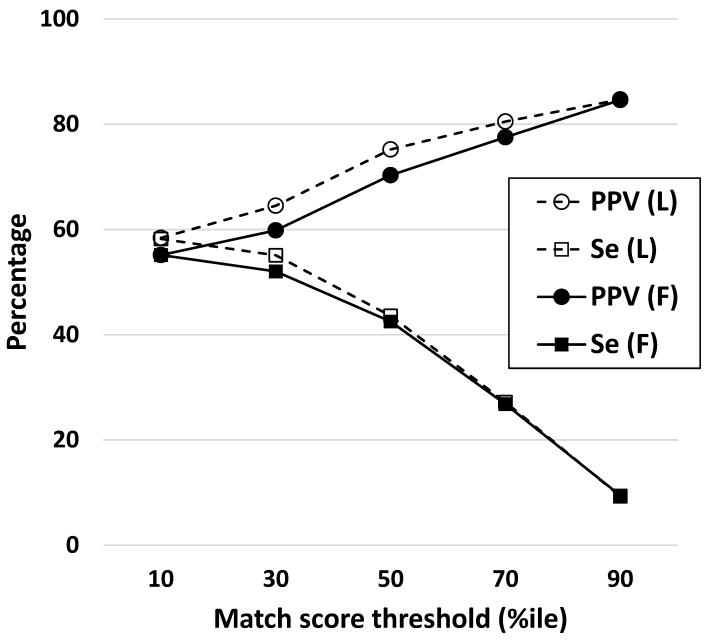
Supplemental Figure 1: Log frequency of match scores calculated for all pairwise comparisons using full algorithm, by true match status



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													
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)	
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 ($\chi 2$) 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