Real-time record linkage between demographic surveillance and health facility data in rural Tanzania

Christopher Rentsch, Georges Reniers, Richard Machemba, Baltazar Mtenga, Chodziwadziwa Kabudula,

Denna Michael, Mark Urassa, Jim Todd, Basia Zaba

Paper for UAPS 2015

ABSTRACT

Most analyses of health service use and its outcomes are limited to patients who are enrolled into clinical care and lack a population perspective. In contrast, health and demographic surveillance systems (HDSS) often rely on self or proxy reports of health services use and usually lack the necessary detail and accuracy about clinical events. Since June 2015, we have been linking patients of three clinics to their HDSS record in real-time, to produce an augmented data source for better monitoring access to and utilization of health facility services in a rural community in Tanzania. So far, we have consented 2,719 patients, and among those who reportedly live in the area covered by the HDSS, we have matched 84 percent to their HDSS record. In this manuscript, we describe the record linkage methodology, report record linkage statistics, and discuss some of the patient attributes that are correlated with (un)successful record linkage.

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) exhaustively measure vital events, but rely on self or proxy reports of health services use and these usually lack detail and accuracy about the clinical events and services received. Linking an individual's HDSS record to their clinic records would provide a unique opportunity to produce an augmented data source for better monitoring access to and utilization of health facility services in the study area.

The Kisesa open cohort study located in northwest Tanzania includes population-based HDSS and HIV sero-surveillance surveys since 1994. Additionally, within the Kisesa HDSS catchment area is a government-run health centre. Previously, data linkage between the health centre and HDSS data was only done retrospectively and based on a probability that any two records between the datasets were a true match.[1,2] In addition to names, the probabilities took into account other personal information such as sex, birthdate, and residence information. Missing data, spelling errors, use of nicknames, name changes after marriage, poor recall of birth dates, and changes of residence complicate retrospective record linkage. While there has been some evidence that analyses using retrospectively matched datasets produced similar results to analyses using truly matched datasets,[3] no comparison has ever been conducted in Kisesa.

In this study, we report on prospective, or "real-time," record linkage between the Kisesa HDSS and multiple clinics located in the Kisesa health centre. This prospective approach differs from deterministic[4] and other probabilistic record linkage approaches,[5,6] which are usually done retrospectively. Real-time record linkage occurs in the presence of the patient, which has the advantage that uncertainty surrounding his or her identity can be resolved and offers an opportunity for seeking informed consent. Using specially designed record linkage software for Kisesa, a fieldworker enters a patient's demographic information in software that uses a probabilistic algorithm to search and rank possible matches from the HDSS database. Then, the interviewer manually searches through each of the returned HDSS records to view detailed information, such as other household members, in order to make a more informed match. We introduced real-time record linkage in Kisesa health centre on 1 June 2015. In this manuscript, we report on fieldwork experiences, initial record linkage statistics, and patient attributes correlated with (un)successful record linkage.

METHODS

Community databases

The first Kisesa HDSS commenced in 1994 in a rural ward located in the Magu district of Mwanza region in northwest Tanzania. The HDSS databases include biannual rounds (29 to date) of household-based surveys that collect information on births, pregnancies, deaths, in- and out-migration, and spousal and parent-child relationships. A limitation of the HDSS data is the inability to link together HDSS records of individuals who move households within the HDSS area. Therefore, an individual may have multiple HDSS records if s/he resided in more than one household in the HDSS area since 1994. These HDSS data are linked to seven rounds of HIV serological surveillance conducted by Tazama staff every three years, with a detailed questionnaire on sexual behaviour and partnership factors, fertility outcomes, HIV-related knowledge, and use of health services.

Clinic databases

Three clinics located in Kisesa health centre were initially selected as real-time 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 these 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. Each patient seen in the CTC receives a unique CTC number. In the ANC and HTC, the Tazama project team has developed electronic databases to store the data collected in the paper logbooks. In the ANC, women receive a unique ANC number that identifies a particular pregnancy.

For any subsequent pregnancy a woman may have, she receives a new ANC number that is unlinked to her previous care in the clinic. Similarly, HTC clients obtain an HTC number that identifies their test number. Attendees to the HTC collect their unique ID number on a piece of cardstock paper to bring back on subsequent visits in order to link multiple tests for the same individual in the clinic.

Interviews with clinic attendees

The subject population 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. As individuals arrived to the clinics, a fieldworker introduced him/herself and the study. The fieldworker then invited the patient to a desk located within the clinic but out of the way of normal clinic operations in order to conduct the brief real-time record linkage interview. The goal of the interview was for the fieldworker and clinic attendee to work together to locate an attendee's record(s) in the HDSS database. Fieldwork started in Kisesa health centre on 1 June 2015 and results presented in this paper included all data collected through 31 October 2015.

Record linkage algorithm

Our computer software utilized a probabilistic record linkage algorithm to search through the HDSS database for potential matches. The algorithm incorporated the following parameters or data fields: up to three names; sex; year, month, and day of birth; village and sub-village; and up to three names for the ten-cell leader of the patient. The algorithm used for searching possible matches and ranking them was based on the probabilistic record linkage model by Newcombe et al.[7] and formalized by Fellegi and Sunter.[5] Let M be a set of true matches and U be a set of true non-matched record pairs. If there were n total data fields used for matching, then two individual agreement probabilities was defined for each field i (with i = 1…n) as follows:

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

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

The higher the ratio m_i/u_i, the more useful a variable or 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} = ln[m_i/u_i] for fields where both datasets agree, and w_{di} = ln[(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 the fields.

What was considered an agreement between two fields differed for each parameter (Supplemental Table 1). Spelling errors, use of nicknames, and name changes after marriage complicated locating an exact match between any two names in these databases. Thus, we used the Jaro-Winkler string comparator approach to compare the name fields between two records.[8] A name-pair resulting in a Jaro-Winkler score of greater than or equal to 0.8 was considered an agreement. For year of birth to agree, the difference between the two compared years needed to be less than 2 years. All other parameters (sex, month and day of birth, village, and subvillage) needed to agree exactly.

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. While searching through these potential matches, the fieldworker was able to view the full list of household members associated with each HDSS record. The fieldworker then worked with the patient to determine which HDSS record, if any, was a true match.

Statistical analyses

For this paper, we calculated preliminary matching statistics including the percentage of individuals whose HDSS record was found. A match percentage was considered the proportion of patients who were matched to at least one HDSS record (numerator) out of the number of patients seen (denominator). We stratified these percentages were stratified by clinic and month of operation. We also calculated match percentages that excluded patients who were unlikely to have an HDSS record – either the patient claimed to never live in the HDSS catchment area or they recently moved or were born into Kisesa ward after the last HDSS round (end of 2014). The quality of the matching algorithm was also assessed by investigating the rank of the matches made.

Chi-square (χ^2) tests were used to assess if the match percentage and patient characteristics differed among the three clinics. Wilcoxon rank-sum tests were used to examine differences in Jaro-Winkler and age differences between record-pairs. Patient characteristics included sex, age, when they were consented in this study (during or after initial training months), and whether their claimed village was more or less rural. More rural villages were considered those which over 60% of the population were classified as rural; less rural villages were those that consisted of less than 60% of the population classified as rural. The first two months of real-time record linkage operations were considered to be "training months." Multivariable logistic regression models were fitted to identify characteristics that were associated with a successful match to an HDSS record among individuals who claimed residence history in the Kisesa HDSS area. As a sensitivity analysis, we also ran similar regression models for each clinic separately to determine if the same conclusions were reached. The Tanzanian National Institute for Medical Research (NIMR) Lake Zone Institutional Review Board (LZIRB) and the London School of Hygiene & Tropical Medicine (LSHTM) granted ethical approval for real-time record linkage.

RESULTS

As of 31 October 2015, we conducted brief interviews with 2,721 clinic attendees. None of these patients refused to provide written consent. Of the 2,721 patients, 889 (33%) claimed to have never lived in the catchment area for the HDSS, and 420 (15%) were recent residents (either born in or moved to the area after the last HDSS round). Thus, 1,421 patients claimed to have residence history in the HDSS area and were considered likely to have a record in the community database. Of these 1,421 patients, 1,188 (84%) were matched to one or more HDSS records. By clinic, the match percentage was 86% in the CTC, 83% in the HTC, and 83% in the ANC (p=0.341). Each of the three clinics improved its match percentage since real-time record linkage was initiated, which was likely due to the fieldworkers gaining understanding of the computer software as well as having additional opportunities to match patients to their HDSS record(s) during repeat visits. The month-specific match percentage increased from 76% in June to its highest level of 93% in September.

Patient characteristics differed significantly between the three clinics (Table 1). Sixty-three percent of the HTC patients, 67% of the CTC patients, and 98% of the ANC patients were female (the other 2% of ANC patients were male infants) (p<0.001). The CTC and HTC patients had a more varied age distribution than ANC patients. In the CTC, 67% of the patients were aged between 15 and 49 and 27% were older than 50. In the HTC, 80% of the patients were aged between 15 and 49 and 18% were older than 50. However, in the ANC, the high majority of patients seen (94%) were aged between 15 and 49. One ANC patient was aged 50 at her first visit and was dropped from the multivariable analyses below. Lastly, a higher proportion of ANC patients claimed to live in less rural villages while CTC and HTC patients were more evenly spread across both village types (p<0.001).

A multivariable model including all patients with a residence history in the HDSS area suggested those who were older, lived in more rural villages, and were first seen after the training months had around double the odds of being matched to an HDSS record than those who were younger, lived in less rural villages, and were seen during the training months, respectively (Table 2). Only two of these associations were detected in the multivariable model including CTC patients only – older patients and patients who reside in more rural villages had over twice the odds to be matched compared to those who were younger and those who lived in less rural villages, respectively. In both the ANC and HTC models, the only significant association detected was that patients seen after the training months had around three times the odds to be matched than those seen during the training months. There was no evidence of an association between a patient's sex and finding an HDSS record overall and by clinic.

Our probabilistic algorithm performed well in this setting. HDSS records were found for 1,188 individuals who claimed to have residence history in the HDSS area. Interestingly, HDSS records were also found for 139 (11%) of the patients who claimed to have no residence history in the HDSS area (the name "Kisesa" refers to a ward, a village within the ward, and a subvillage within that, which makes it conceivable that patients may have different concepts of how the HDSS catchment area was defined). Additionally, some of the individuals claimed to have multiple residency episodes within the HDSS area, thus qualifying them to have more than one HDSS ID record. In total, we matched 1,764 HDSS records to 1,327 individuals. Of these individuals, 983 (74%) were matched to only one HDSS record, 269 (20%) were matched to two HDSS records, 60 (5%) were matched to three HDSS records, 12 (1%) were matched to four HDSS records, and 3 (<1%) were matched to five HDSS records. Of the 1,764 matches made, 1,342 (76%) were ranked with the highest score by the matching algorithm, and 237 (13%) were ranked with the second highest score. The remaining 185 (11%) matched records were ranked between third and eleventh by the computer algorithm. No matches resulted in scores ranked lower than eleventh.

A table that shows the completeness of all parameters as well as the match probabilities (m_i) overall and by clinic can be found in the supplemental material (Supplemental Table 1). Among all 1,764 matches made, the fields with the highest completeness (greater than 98%) were first name, second name, sex, and year of birth. Based on the Jaro-Winkler scores, the match probability was 0.95 among first name comparisons and 0.88 among second name comparisons. Both of these match probabilities differed significantly by clinic (first name p=0.03; second name p<0.01). For both names, the match probabilities were highest in the CTC and lowest in the ANC. The match probability for sex was 0.99 and did not significantly differ by clinic (p=0.537). The match probability for year of birth (based on the condition that it was less than two years difference) was 0.55 and did not significantly differ by clinic (p=0.550). After further examining the distribution of the age differences, 85% of the matches made had an age difference of 2 years or less, but this increased to 95% with an age difference of 6 years or less and 98% with an age difference of 10 years or less.

DISCUSSION

Real-time record linkage – which combines probabilistic methods with a relatively simple human intervention – shows promise for prospectively linking multiple sources of data in rural Tanzania. The percentage of individuals who were matched among those who claimed to have residence history in the

HDSS area was similar to match percentages found in other African HDSS areas. The Agincourt HDSS located in South Africa currently has a match percentage between 70% and 80% (Chodziwadziwa Kabudula, personal communication). Also, a predominately paediatric clinic located in the Kilifi HDSS area in Kenya has a current match percentage at around 65% (Greg Fegan and Eduard Sanders, personal communication).

All three clinics in Kisesa experienced a substantial increase in match percentage in all but one month of record linkage operations. However, the increase in match percentage was more pronounced in the ANC and HTC than it was in the CTC. The lowest match percent in the CTC was 75% while the lowest in the ANC was 59% and 46% in the HTC. This is likely due to the varied previous work experiences between the fieldworkers. All fieldworkers had previous experience with cleaning data in Kisesa health centre; however, the fieldworker assigned to the CTC during training had over a year of additional experience working with HDSS data. This experience likely equipped this fieldworker with more advanced tactics on how to best search through the HDSS database using the record linkage software. These strategies were shared with the other fieldworkers during daily, end-of-day meetings. Indeed, the multivariable results showed that while there were no increased odds for patients to be matched after training compared with during training in the CTC, there was about a three-fold increase in the odds of finding a match after the training months compared with during training in both the ANC and HTC.

The multivariable results also suggested that older patients had higher odds to be matched to an HDSS record. These older patients may have spent more time in an established residence in the HDSS area, which may have made it more likely to find their HDSS record than for more transient or younger individuals. However, this association was not detected when looking at HTC patients alone (and this cannot be tested in the ANC as only one patient was seen in this older category). Another HDSS record linkage site located in South Africa found that older individuals were significantly less likely to be matched. Further investigation into this discrepancy is required.

In the model combining all patients as well as in CTC patients alone, patients who claimed to live in a more rural village had about twice the odds to be matched than those living in less rural villages. One explanation of this phenomenon could potentially be due to the higher rate of migration within and into the less rural areas, which have a higher density of households than in more rural villages. Meanwhile, those who cl aim to live in a village that are more rural may likely have a more stable residence history.

The probabilistic algorithm worked sufficiently well in this setting. A high majority of the matches were ranked as the most likely or second most likely match based on the scores output by the probabilistic algorithm. Several of the individual parameters used in the algorithm had both high percentages of completeness as well as match probabilities (first and second name, sex, year of birth, village, and subvillage). Nevertheless, the level of completeness and match probabilities significantly varied on several of these and other parameters among the three clinics. First and second names for both the patient and their ten-cell leaders had significantly higher match probabilities in the CTC than in the ANC and HTC. Additionally, the match probabilities for village and subvillage were significantly higher in the CTC than in the ANC and HTC. As previously mentioned, the fieldworker initially assigned to the CTC likely has an advantage over the other fieldworkers due to his previous work experiences. As a way to control for this "fieldworker effect," we began rotating the fieldworkers among the clinics. We aim to examine this effect in future analyses.

CONCLUSION

Real-time record linkage is a promising tool for linking multiple sources of data in rural Tanzania. After five months of operation, we have achieved similar levels of success to other African HDSS sites conducting real-time record linkage. However, continued monitoring of these record linkage statistics is required. There are some discrepancies arising among the clinics in terms of completeness and quality of data collection. Improvements have already been made in the software to allow us to control for these discrepancies in future analyses. Further testing of the probabilistic algorithm is also required to maximise the full potential of real-time record linkage in these settings.

REFERENCES

1. Gourlay A, Wringe A, Todd J, 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.
2. Cawley C, Wringe A, Todd J, 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.
3. 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).
4. 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.
5. Fellegi IP, Sunter AB. A Theory for Record Linkage. *J Am Stat Assoc.* 1969;64(328):1183-1210.
6. Jaro MA. Probabilistic linkage of large public health data files. *Stat. Med.* 1995;14:491-498.
7. Newcombe H, Kennedy J, Axford S, James A. Automatic Linkage of Vital Records. *Science.* 1959;130(3381):954-959.
8. Winkler WE. *String Comparator Metrics and Enhanced Decision Rules in the Fellegi-Sunter Model of Record Linkage.* 1990.

TABLES

Table 1. Comparisons of real-time record linkage participants by clinic, n=2,721

Exclusion criteria	**CTC** **(n=846)**	**ANC** **(n=960)**	**HTC** **(n=915)**	***P*[a]**
Total excluded	404 (47.8)	328 (34.2)	568 (62.1)	<0.001
Never lived in HDSS area	383 (45.3)	117 (12.2)	389 (42.5)	<0.001
Recently born or moved into HDSS area	22 (2.6)	212 (22.1)	186 (20.3)	<0.001
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				
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[b]				
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 sentinal surveillance

Note: all statistics are given in n(%)

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

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

Table 2. Associations of being matched to an HDSS record among those with residence history in the Kisesa HDSS area, overall and by clinic

	Overall[a]	**By clinic**		
	(n=1,421)	**CTC (n=442)**	**ANC (n=631)**	**HTC (n=347)**
Covariate	**OR (95% CI)**	**OR (95% CI)**	**OR (95% CI)**	**OR (95% CI)**
Male sex, *(ref=female)*	0.97 (0.65, 1.46)	1.33 (0.72, 2.48)	0.33 (0.06, 1.78)	0.82 (0.45, 1.49)
Age				
<15	0.66 (0.35, 1.24)	0.98 (0.31, 3.10)	1.35 (0.37, 4.89)	0.13 (0.03, 0.59)
15-49	ref	Ref	ref	ref
50+	**2.28 (1.28, 4.07)**	**2.77 (1.26, 6.09)**	[b]	2.08 (0.84, 5.20)
Claimed village of residence				
More rural	**1.79 (1.25, 2.55)**	**2.63 (1.41, 4.93)**	1.32 (0.74, 2.35)	1.90 (0.98, 3.69)
Less rural	ref	Ref	ref	ref
When first seen[c]				
After training months	**2.49 (1.80, 3.43)**	1.15 (0.61, 2.17)	**3.31 (2.08, 5.28)**	**2.61 (1.38, 4.94)**
During training months	ref	ref	ref	ref

Abbreviations: HDSS - health and demographic sentinal 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] Overall model also was adjusted for department

[b] One ANC patient who was 50 years of age removed from multivariable analyses

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

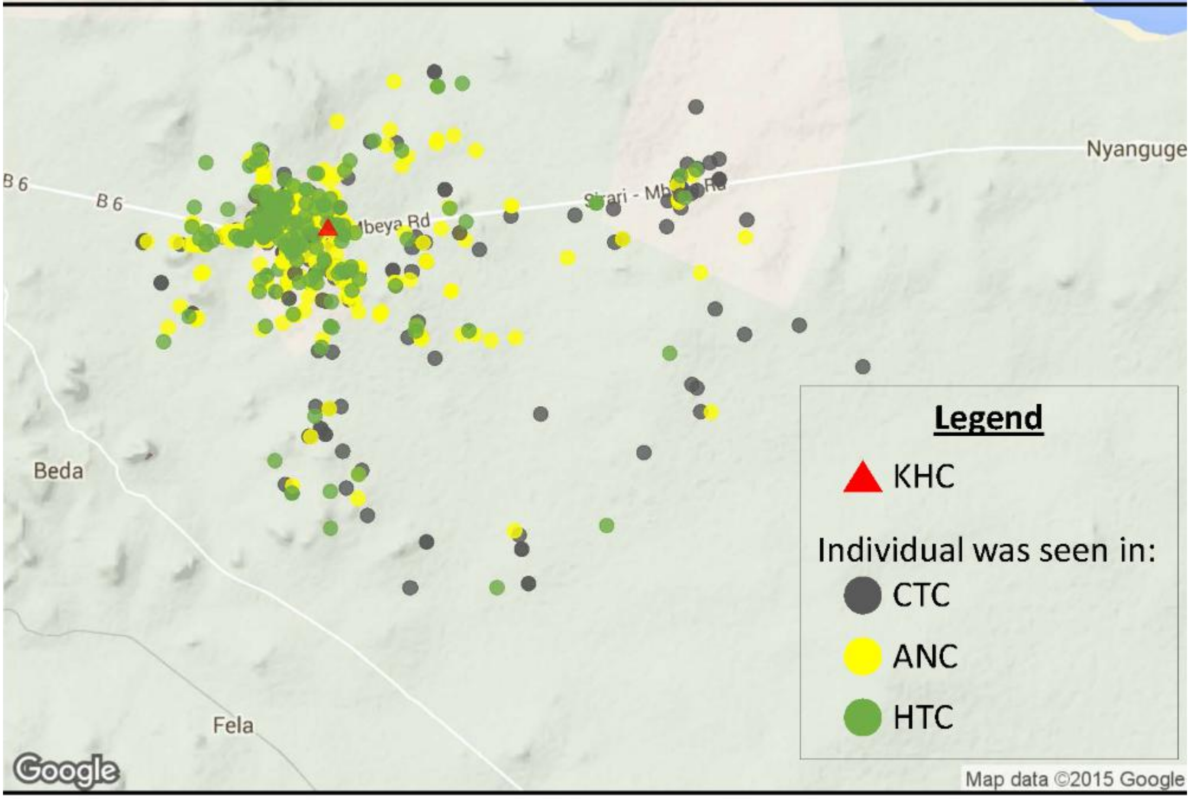
SUPPLEMENTAL MATERIAL

Supplemental Table 1. Match probabilities for each parameter in the probabilistic algorithm used to search the HDSS database for potential matches, n_M=1,764 matches

		Overall		**By clinic**					
		(n_M=1,764)		**CTC (n_M=630)**		**ANC (n_M=714)**		**HTC (n_M=420)**	
Parameter *i*	**Agreement condition**	**% collected**	**m_i**	**% collected**	**m_i**	**% collected**	**m_i**	**% collected**	**m_i**
First name	Jaro-Winkler ≥ 0.8	100.0%	0.9546	100.0%	0.9619	100.0%	0.9342	100.0%	0.9405
Second name	Jaro-Winkler ≥ 0.8	99.9%	0.8825	100.0%	0.9206	100.0%	0.8375	99.5%	0.9019
Third name	Jaro-Winkler ≥ 0.8	76.2%	0.0565	97.1%	0.0490	91.7%	0.0458	18.3%	0.2078
TCL first name	Jaro-Winkler ≥ 0.8	77.5%	0.4704	90.6%	0.6445	78.0%	0.3501	56.9%	0.3347
TCL second name	Jaro-Winkler ≥ 0.8	68.3%	0.4896	85.6%	0.6698	65.7%	0.3625	46.9%	0.2995
TCL third name	Jaro-Winkler ≥ 0.8	0.3%	0.0000	0.0%	0.0000	0.3%	0.0000	0.7%	0.0000
Sex	exact match	99.9%	0.9892	100.0%	0.9921	99.9%	0.9860	100.0%	0.9905
Year of birth	within 2 years	98.6%	0.8476	98.3%	0.8207	98.7%	0.8879	98.8%	0.8193
Month of birth	exact match	4.5%	0.4250	4.8%	0.6000	1.3%	0.3333	9.8%	0.3171
Day of birth	exact match	4.4%	0.3247	4.4%	0.5000	1.3%	0.2222	9.5%	0.2250
Village	exact match	95.6%	0.9484	96.3%	0.9769	95.9%	0.9270	93.8%	0.9416
Subvillage	exact match	95.6%	0.7888	96.3%	0.8468	95.9%	0.7533	93.8%	0.7614

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

Figure 1. Spatial mapping of the households of matched individuals in Kisesa ward, by clinic



Abbreviations: KHC – Kisesa Health Centre; CTC – HIV care and treatment centre; ANC – antenatal clinic; HTC – HIV testing and counselling clinic