SOFTWARE TOOL ARTICLE

REVISED Point-of-contact Interactive Record Linkage (PIRL): A software tool to prospectively link demographic surveillance and health facility data [version 2; referees: 2 approved]

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 Zaba [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]SELECT Star, Atlanta, GA, 30309, USA
[4]Independent Researcher, London, UK
[5]The Tazama Project, National Institute for Medical Research, Mwanza, Tanzania
[6]District Medical Officer, Ministry of Health Tanzania, Magu District, Tanzania

v2 **First published:** 06 Nov 2017, **1**:8 (doi: 10.12688/gatesopenres.12751.1)
Latest published: 11 Jan 2018, **1**:8 (doi: 10.12688/gatesopenres.12751.2)

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

Open Peer Review

Referee Status: ✓ ✓

	Invited Referees 1	2
REVISED **version 2** published 11 Jan 2018		
version 1 published 06 Nov 2017	✓ report	✓ report

1 **Duncan Smith**, University of Manchester, UK

2 **Hye-Chung Kum**, Texas A & M University, USA
Texas A & M University, USA

Discuss this article

Comments (0)

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.

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

Corresponding author: Christopher T. Rentsch (christopher.rentsch@lshtm.ac.uk)

Author roles: Rentsch CT: Conceptualization, Data Curation, Funding Acquisition, Investigation, Methodology, Project Administration, Resources, Software, Supervision, Validation, Visualization, Writing – Original Draft Preparation, Writing – Review & Editing; **Kabudula CW**: Conceptualization, Data Curation, Methodology, Resources, Software, Supervision, Validation, Writing – Original Draft Preparation, Writing – Review & Editing; **Catlett J**: Data Curation, Methodology, Resources, Software, Supervision, Writing – Original Draft Preparation, Writing – Review & Editing; **Beckles D**: Conceptualization, Data Curation, Methodology, Resources, Software, Supervision, Writing – Original Draft Preparation, Writing – Review & Editing; **Machemba R**: Investigation, Methodology, Project Administration, Resources, Software, Supervision, Writing – Review & Editing; **Mtenga B**: Data Curation, Investigation, Methodology, Project Administration, Resources, Software, Supervision, Writing – Review & Editing; **Masilela N**: Conceptualization, Data Curation, Methodology, Resources, Software, Writing – Review & Editing; **Michael D**: Investigation, Project Administration, Resources, Supervision, Writing – Review & Editing; **Natalis R**: Investigation, Project Administration, Resources, Supervision, Writing – Review & Editing; **Urassa M**: Conceptualization, Data Curation, Funding Acquisition, Investigation, Methodology, Project Administration, Resources, Software, Supervision, Validation, Visualization, Writing – Review & Editing; **Todd J**: Conceptualization, Data Curation, Funding Acquisition, Investigation, Methodology, Project Administration, Resources, Software, Supervision, Validation, Visualization, Writing – Review & Editing; **Zaba B**: Conceptualization, Data Curation, Funding Acquisition, Investigation, Methodology, Project Administration, Resources, Software, Supervision, Validation, Visualization, Writing – Original Draft Preparation, Writing – Review & Editing; **Reniers G**: Conceptualization, Data Curation, Funding Acquisition, Investigation, Methodology, Project Administration, Resources, Software, Supervision, Validation, Visualization, Writing – Original Draft Preparation, Writing – Review & Editing

Competing interests: No competing interests were disclosed.

How to cite this article: Rentsch CT, Kabudula CW, Catlett J *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 Research 2018, **1**:8 (doi: 10.12688/gatesopenres.12751.2)

Grant information: 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].
The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

First published: 06 Nov 2017, **1**:8 (doi: 10.12688/gatesopenres.12751.1)

REVISED **Amendments from Version 1**

We thank the reviewers for their comments and suggestions. We have carefully reviewed them, and incorporated them into the updated version of our paper. We feel that the comments have resulted in positive changes to our manuscript, which include a clearer definition of existing record linkage methodology, a brief call to other implementations of interactive record linkage, a description of how we calculated our u_i probabilities, an additional description of privacy during the record linkage interview, and a few other relatively minor changes as suggested by the reviewers.

See referee reports

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[1]. 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[2] and 1960s[3]. Two popular methods of record linkage have been used to combine data sources. Deterministic record linkage[4] is a rule-based approach that typically requires exact matching on a set of identifiers existing in all data sources. Probabilistic methods[5–7] 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[8]. 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[9], cardiovascular risk after acute infection[10], and the association between body mass index and cancer[11].

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[12].

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[13,14], 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.*[15] adapted a comprehensive checklist[16] 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. Many of the available software packages are designed for batch linkages, such as those used in purely automated retrospective linkage[17,18]. 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[19]. This paper describes our implementation of this software, and we attach a GitHub link[20] to the full source code for others to download and amend to their own research needs.

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 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.

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[2,3], with match probabilities (m_i) that have been adopted from a pilot study in the Agincourt HDSS[21]. 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[7]. 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:

$$\text{match probability: } m_i = \text{P(field } i \text{ agrees} \mid j \in M) \quad (1.1)$$

$$\text{unmatch probability: } u_i = \text{P(field } i \text{ agrees} \mid j \in U) \quad (1.2)$$

For a given field with match probability m_i and unmatch probability u_i, the software calculates the matching weights w_{ai} as $= \log_2[m_i/u_i]$ for fields where both datasets agree, and w_{di} as $= \log_2[(1-m_i)/(1-u_i)]$ where they disagree. Assuming independence of observations across the fields, the match score is computed by summing the weights across all fields[3,15].

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[22]. Previous research has shown the Jaro-Winkler method produces similar results to Double Metaphone and Soundex string comparators in a southern African context[21]. 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.

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 1). Information from most of these fields contribute to the linkage algorithm described in the Implementation section above.

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.

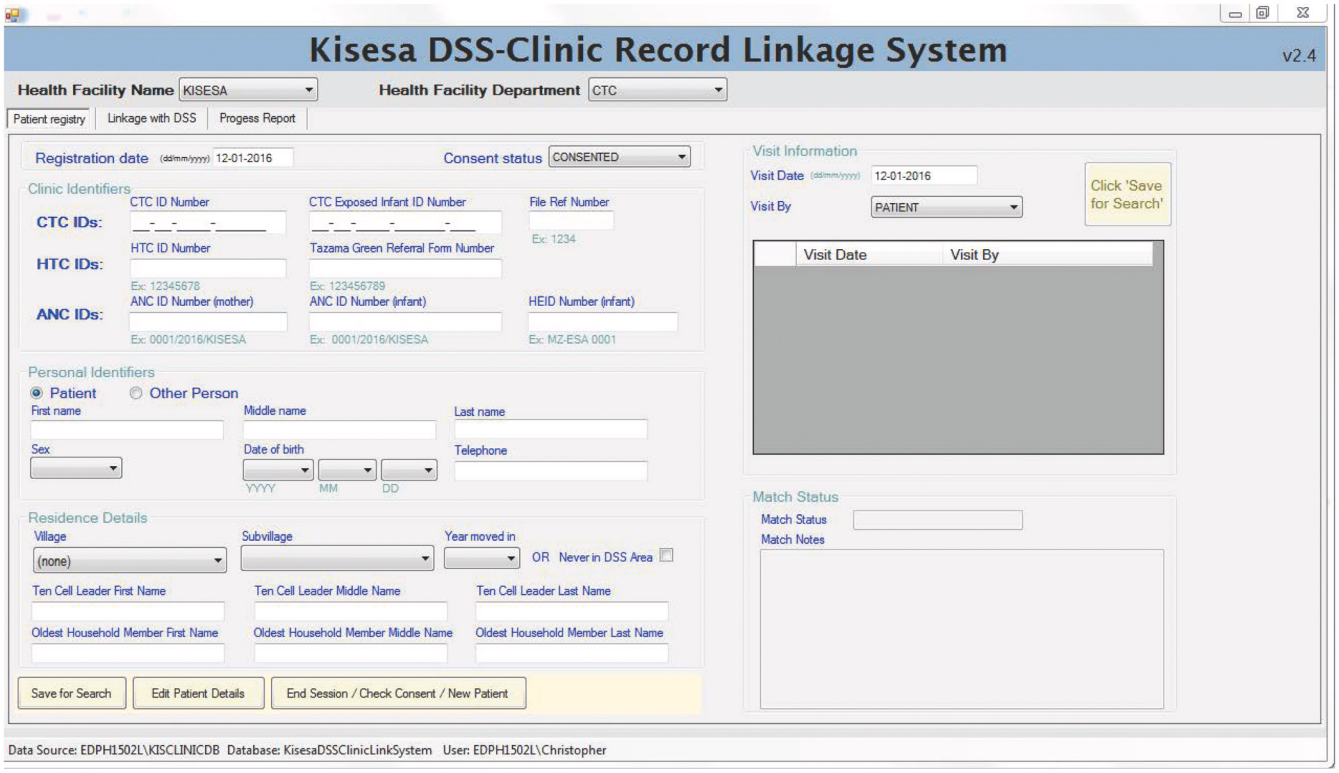

Figure 1. User interface of Point-of-contact Interactive Record Linkage (PIRL) software.

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.

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).

Use cases

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).

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 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).

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 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.

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 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.

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 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

Table 1. Personal identifiers used for three case patients sampled from the fake dataset 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: 44618061		
		HTC: 44447050			
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
Subvillage	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

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.

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.

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[19]. 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.

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[23]

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).

Competing interests

No competing interests were declared.

Grant information

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].

The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Acknowledgements

The authors thank 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).

Supplementary material

Supplementary File 1. Kisesa-HDSS record linkage user guide.

Click here to access the data.

Supplementary File 2. Fake input dataset with codebook.

Click here to access the data.

Supplementary File 3. Script to create fake input dataset.

Click here to access the data.

Supplementary File 4. Output datasets for case patients with codebook.

Click here to access the data.

References

1. Trust W: **Enabling Data Linkage to Maximise the Value of Public Health Research Data: full report.** 2015.
Reference Source
2. Newcombe HB, Kennedy JM, Axford SJ, *et al.*: **Automatic linkage of vital records.** *Science.* 1959; **130**(3381): 954–9.
PubMed Abstract | Publisher Full Text
3. Fellegi IP, Sunter AB: **A Theory for Record Linkage.** *J Am Stat Assoc.* 1969; **64**(328): 1183–210.
Publisher Full Text
4. Roos LL Jr, 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.
PubMed Abstract | Publisher Full Text
5. Jaro MA: **Probabilistic linkage of large public health data files.** *Stat Med.* 1995; **14**(5–7): 491–8.
PubMed Abstract | Publisher Full Text
6. Meray N, Reitsma JB, Ravelli AC, *et al.*: **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.
PubMed Abstract | Publisher Full Text
7. Sayers A, Ben-Shlomo Y, Blom AW, *et al.*: **Probabilistic record linkage.** *Int J Epidemiol.* 2016; **45**(3): 954–64.
PubMed Abstract | Publisher Full Text | Free Full Text
8. Herrett E, Gallagher AM, Bhaskaran K, *et al.*: **Data Resource Profile: Clinical Practice Research Datalink (CPRD).** *Int J Epidemiol.* 2015; **44**(3): 827–36.
PubMed Abstract | Publisher Full Text | Free Full Text
9. Smeeth L, Cook C, Fombonne E, *et al.*: **MMR vaccination and pervasive developmental disorders: a case-control study.** *Lancet.* 2004; **364**(9438): 963–9.
PubMed Abstract | Publisher Full Text
10. Smeeth L, Thomas SL, Hall AJ, *et al.*: **Risk of myocardial infarction and stroke after acute infection or vaccination.** *N Engl J Med.* 2004; **351**(25): 2611–8.
PubMed Abstract | Publisher Full Text
11. Bhaskaran K, Douglas I, Forbes H, *et al.*: **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.
PubMed Abstract | Publisher Full Text | Free Full Text
12. Sankoh O, INDEPTH Network: **CHESS: an innovative concept for a new generation of population surveillance.** *Lancet Glob Health.* 2015; **3**(12): e742.
PubMed Abstract | Publisher Full Text
13. Fure E: **Interactive Record Linkage: The Cumulative Construction of Life Courses.** *Demogr Res.* 2000; **3**(11).
Publisher Full Text
14. Kum HC, Krishnamurthy A, Machanavajjhala A, *et al.*: **Privacy preserving interactive record linkage (PPIRL).** *J Am Med Inform Assoc.* 2014; **21**(2): 212–20.
PubMed Abstract | Publisher Full Text | Free Full Text
15. Herzog TN, Scheuren FJ, Winkler WE: **Data quality and record linkage techniques.** Springer Science & Business Media; 2007.
Publisher Full Text
16. 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.
Reference Source
17. Christen P, Churches T, Hegland M: **Febrl-a parallel open source data linkage system.** *Lect Notes Comput Sc.* 2004; 638–47.
Publisher Full Text
18. Jurczyk P, Lu JJ, Xiong L, *et al.*: **Fine-grained record integration and linkage tool.** *Birth Defects Res A Clin Mol Teratol.* 2008; **82**(11): 822–9.
PubMed Abstract | Publisher Full Text
19. Rentsch CT, Reniers G, Kabudula C, *et al.*: **Point-of-contact interactive record linkage (PIRL) between demographic surveillance and health facility data in rural Tanzania.** *International Journal of Population Data Science.* 2017; **2**(1).
Publisher Full Text
20. Kabudula C, Rentsch C, Catlett J, *et al.*: **PIRL - Point-of-contact Interactive Record Linkage software.** 2017.
Publisher Full Text
21. Kabudula CW, Clark BD, Gómez-Olivé FX, *et al.*: **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.
PubMed Abstract | Publisher Full Text | Free Full Text
22. Winkler WE: **String Comparator Metrics and Enhanced Decision Rules in the Fellegi-Sunter Model of Record Linkage.** 1990.
Reference Source
23. LSHTM-ALPHAnetwork: **LSHTM-ALPHAnetwork/PIRL_RecordLinkageSoftware: Initial public release of Point-of-contact Interactive Record Linkage (PIRL) software.** *Zenodo.* 2017.
Data Source

Open Peer Review

Current Referee Status: ✔ ✔

Version 1

Referee Report 27 December 2017

doi:10.21956/gatesopenres.13811.r26151

✔ **Hye-Chung Kum** [1,2]
[1] Department of Health Policy & Management, Texas A & M University, College Station, TX, USA
[2] Department of Computer Science and Engineering, Texas A & M University, College Station, TX, USA

This paper is a case study of using an interactive record linkage software at point of contact in Tanzania. Interactive record linkage at point of contact has benefits over retrospective linkages, so makes sense to do this when possible.

The paper has some good points, but below are some suggestions for improvement.

* [CRITICAL] The following sentence is incorrect and should be edited
- "Deterministic record linkage is a rule-based approach that 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."

- Deterministic RL does not require exact match between identifiers. For example, same soundex of the name is not an exact match but rather an approximate match and can be used in determistics methods. And determinististic methods can be an effective method to link data when common unique identifiers are not available. There are pros & cons to both the determinististic and probablistic approach. A more relevent distinction is between exact match and approximate match. There is a section in the paper about "Agreement conditions vary for each of the parameters." which discuss the degree of approximate match, either as a field or a full record. Calling exact match based algorithms determinististic match is a common but confusing nomenclature. Both determistics and probablistic match can be based on exact match on fields, or approximate match on fields. Due to many issues in real data, approximate matched based algorithms (both determinististic and probabilistic) do better. It is important to not confuse exact match with determinististic methods for this reason. The quality of matching results are comparable for both deterministic and probabilistic methods as long as the process for linkage is well developed (Antonie 2014, Zhu 2015). More importantly, data standardization, cleaning, flexibility on approximate matches are important in both approaches.

Antonie L, Inwood K, Lizotte DJ, Andrew Ross J. Tracking people over time in 19th century Canada for longitudinal analysis. Mach Learn. 2014;95(1):129-146. doi:10.1007/s10994-013-5421-0.
Zhu Y, Matsuyama Y, Ohashi Y, Setoguchi S. When to conduct probabilistic linkage vs. deterministic linkage? A simulation study. J Biomed Inform. 2015;56:80-86. doi:10.1016/j.jbi.2015.05.012.

* [IMPORTANT] An explanation of the role of household as a unit, as well as how it relates to the linkage task would make the paper more clear. My read of the paper, the record linkage task is for people, yet there is many mention of the word household and it seems there is some important aspect of the household used in the linkage process but not described anywhere in the paper.

* [IMPORTANT] This might be related to the above point. A better description of the EXACT linkage goal would improve the paper. The following are some sections that need better clarification.

- "One major weakness of the Kisesa HDSS is the lack of reconciling records of individuals who move households within the HDSS area. Therefore, 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."
> what does it mean to move households? Do you mean a the composition of the household changes? For example, a daughter from household A, marries and moves to a different household as a wife? Or is this a simple family moved to a new location? What is a household in this context? Is HDSS IDs a person level identifier, that is if there are multiple IDs per person, are these duplicate records that need to be cleaned out of the database? If not, what is the unit of the HDSS IDs? Are those household person IDs (meaning, when a person is in a different household, they should have another ID, even if it is the same person) ?

- The RL process diagram and other explanations in appendix 1 should be better summerized to be included in the main text, as it is important that the reader understand this process, and the paper should be understandable without having to fully read the appendix.

- The real time RL occurs when a patient visits a clinic. Thus, is the goal to identify the correct record for the patient in the HDSS at the time of visit? (which sounds like correct record retrival task). Or is the goal to clean the HDSS of duplicate records at the point of patient visit? Strictly speaking this is a deduplication task, and identifying the duplicate records in only the first step. How to 'clean' the database after identification is more important but not discuss much in this paper. Or maybe it is to identify ALL records relating to the patient in the HDSS at point of visit, and link these records within the HDSS system, leaving the duplicate records along. If this is correct, what id the unit of HDSS ID and why do you need it smaller than a person and keep duplicate records per person.

* The backup process description could be more clear. Again, the goal of backup is unclear in the paper. Is the goal to consolidate records from all computers in a local clinic then have the local databases synced to the master HDSS database on the cloud once a day?

* Given the sensitive nature of HIV, a brief discussion on the issue of privacy and what the patient can and cannot see during the process would be good to include in the main paper. Maybe a discussion of future work to improve privacy.

* Although this paper is about interactive RL, there is no review of the literature on interactive RL. A discussion of the general pros and cons of interactive record linkage along with references would frame the paper better. A focus on the role of the person in the process, what the person needs from the automatic process to do a good job, and how the software meets those need might work well. Below are some references that might help you get started

Martha Bailey, Connor Cole, Morgan Henderson, and Catherine Massey. 2017. How Well Do Automated Linking Methods Perform in Historical Samples? Evidence from New Ground Truth. Technical Report.

Working Paper.

Gordon Darroch. 2002. Semi-Automated Record Linkage with Surname Samples: a Regional Study of Case LawLinkage, Ontario 1861–1871. History and Computing 14, 1-2 (2002), 153–183.

Hyunmo Kang, Lise Getoor, Ben Shneiderman, Mustafa Bilgic, and Louis Licamele. 2008. Interactive entity resolution in relational data: A visual analytic tool and its evaluation. IEEE transactions on visualization and computer graphics 14, 5 (2008), 999–1014.

Hye-Chung Kum, Ashok Krishnamurthy, Ashwin Machanavajjhala, Michael K Reiter, and Stanley Ahalt. 2014b. Privacy preserving interactive record linkage (PPIRL). Journal of the American Medical Informatics Association 21, 2 (2014), 212–220.

Eric Ragan, Hye-Chung Kum, et al. 2018. Balancing Privacy and Information Disclosure in Interactive Record Linkage with Visual Masking. ACM SIGCHI 2018.

Qiaomu Shen, Tongshuang Wu, Haiyan Yang, Yanhong Wu, Huamin Qu, and Weiwei Cui. 2017. NameClarifier: a visual analytics system for author name disambiguation. IEEE transactions on visualization and computer graphics 23, 1 (2017), 141–150.

Is the rationale for developing the new software tool clearly explained?
Partly

Is the description of the software tool technically sound?
Partly

Are sufficient details of the code, methods and analysis (if applicable) provided to allow replication of the software development and its use by others?
Partly

Is sufficient information provided to allow interpretation of the expected output datasets and any results generated using the tool?
Partly

Are the conclusions about the tool and its performance adequately supported by the findings presented in the article?
Partly

Competing Interests: No competing interests were disclosed.

I have read this submission. I believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.

Referee Report 08 November 2017

doi:10.21956/gatesopenres.13811.r26086

Duncan Smith
School of Social Sciences, University of Manchester, Manchester, UK

The paper describes a record linkage application that has been used to link patient records in Tanzania. The difference between this and other linkage applications is that new records are entered with the patient present and therefore able to assist in identifying correct matches. It is only the current patient's new record that is linked against the other records in the database.

The paper is generally well written. The use cases are useful for illustrating how the system is used in practice. But I do have a few questions / suggestions.

- Where was Reference 17 (Rentsch CT, Reniers G, Kabudula C, et al.) published? The reference is incomplete.

- "The higher the ratio m_i / u_i , the more useful a field is for matching purposes."
  I wouldn't put it like that. A very low ratio is also very useful. (The sentence is probably superfluous anyway.)

- "A Jaro-Winkler score ≥ 0.8 was considered a match."
  Perhaps re-word to make it clear this means a match on the field rather than on the record pair. It is not entirely clear from the description how names are handled. Is a match on name declared if at least one of the Jaro-Winkler scores are ≥ 0.8, or something else?
  Maybe the above sentence could be something like "One or more Jaro-Winkler scores ≥ 0.8 was considered a match on name."

- The paper explains that the m_i are derived from a pilot study. But where do the u_i come from? They (and the m_i) could be estimated from the database itself. Have the authors considered this?

- "The software automatically detects when a patient has been seen during a previous clinic visit and displays the match status (e.g., matched, not matched) to the user. The dates of all follow-up visits are automatically logged into the system."
  How does this happen? Automatically suggests without input from the individual. What match status? Automatically logged? (This is explained in more detail later, but it is not clear at this point in the paper.)

- .bak is commonly used as a file extension for backups of arbitrary file types, so what is ".bak format"?

- "The use of nick-names and interchangeable name order (exemplified in Case 2) is accounted for in the linkage algorithm by allowing all pairwise comparisons between reported names and names found in the HDSS data source."
  There is no need to say this twice in a short paper.

- The software is released under the MIT licence and made available via GiHub. This is a good thing. However, it does use proprietary technologies that might limit its applicability.

- "The software has been developed for machines running a Windows 7 operating system."
  Does it run on other Windows versions? What are the options (if any) for potential users without

available Windows machines?
(I won't comment on the code here as it's probably not particularly relevant to the paper itself.)

- I think a little more could be added on the privacy aspect. How much of the data in the potential matches is the patient allowed to see / know? Potential matches could easily relate to people who live near the patient or are close relatives. It seems that great care would need to be taken to avoid revealing the identity of others in the database.

Is the rationale for developing the new software tool clearly explained?
Yes

Is the description of the software tool technically sound?
Partly

Are sufficient details of the code, methods and analysis (if applicable) provided to allow replication of the software development and its use by others?
Yes

Is sufficient information provided to allow interpretation of the expected output datasets and any results generated using the tool?
Partly

Are the conclusions about the tool and its performance adequately supported by the findings presented in the article?
Partly

Competing Interests: No competing interests were disclosed.

Referee Expertise: Statistics, record linkage, graphical models

I have read this submission. I believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.