International Journal of Population Data Science

Journal Website: www.ijpds.org

Record linkage for routinely collected health data in an African health information exchange

Themba Mutemaringa[1,2,3], Alexa Heekes[1,2], Mariette Smith[1,2], Andrew Boulle[1,2], and Nicki Tiffin[4,5,*]

Submission History

Submitted:	19/06/2022
Accepted:	05/01/2023
Published:	01/03/2023

[1]Provincial Health Data Centre, Health Intelligence Directorate, Western Cape Government Health, Western Cape Province, South Africa
[2]Centre for Infectious Disease Epidemiology and Research, School of Public Health and Family Medicine, University of Cape Town, South Africa
[3]Computational Biology Division, Integrative Biomedical Sciences Department, University of Cape Town, South Africa
[4]Wellcome Centre for Infectious Disease Research in Africa, Faculty of Health Sciences, University of Cape Town, South Africa
[5]South African National Bioinformatics Institute, University of the Western Cape, South Africa

Abstract

Introduction
The Patient Master Index (PMI) plays an important role in management of patient information and epidemiological research, and the availability of unique patient identifiers improves the accuracy when linking patient records across disparate datasets. In our environment, however, a unique identifier is seldom present in all datasets containing patient information. Quasi identifiers are used to attempt to link patient records but sometimes present higher risk of over-linking. Data quality and completeness thus affect the ability to make correct linkages.

Aim
This paper describes the record linkage system that is currently implemented at the Provincial Health Data Centre (PHDC) in the Western Cape, South Africa, and assesses its output to date.

Methods
We apply a stepwise deterministic record linkage approach to link patient data that are routinely collected from health information systems in the Western Cape province of South Africa. Variables used in the linkage process include South African National Identity number (RSA ID), date of birth, year of birth, month of birth, day of birth, residential address and contact information. Descriptive analyses are used to estimate the level and extent of duplication in the provincial PMI.

Results
The percentage of duplicates in the provincial PMI lies between 10% and 20%. Duplicates mainly arise from spelling errors, and surname and first names carry most of the errors, with the first names and surname being different for the same individual in approximately 22% of duplicates. The RSA ID is the variable mostly affected by poor completeness with less than 30% of the records having an RSA ID.

The current linkage algorithm requires refinement as it makes use of algorithms that have been developed and validated on anglicised names which might not work well for local names. Linkage is also affected by data quality-related issues that are associated with the routine nature of the data which often make it difficult to validate and enforce integrity at the point of data capture.

Keywords
health information exchange; data linkage; global South; routine health data; Africa; South Africa

*Corresponding Author:
Email Address: ntiffin@sanbi.ac.za (Nicki Tiffin)

https://doi.org/10.23889/ijpds.v8i1.1771

Introduction

In the field of health informatics, the Patient Master Index (PMI) is considered one of the most important tools used in the patient identity management system. A PMI system issues and maintains a unique patient identifier (UPI) which is recorded against patient registration details such as name, surname, date of birth, gender, national registration number and other necessary personal details.

Most healthcare organisations have multiple information systems which can only provide full utility when integrated. The PMI serves to facilitate such integration thereby providing an opportunity for viewing unified data from patient care history regardless of the data coming from disparate sources. The success of integration endeavours, therefore, is highly dependent on the accuracy and quality of the PMI and a good quality PMI is crucial for a high-quality health care delivery system.

Maintaining an accurate PMI is in real-life a very difficult task as it is often affected by multiple issues. The American Health Information Management Association (AHIMA) identifies three major types of issues that affect PMI namely: duplicates, overlay and overlap [1]. Duplicate records occur when the PMI carries multiple instances of records for the same patient; overlap is when a patient's records are spread across facilities, and overlay is when information belonging to two or more patients is incorrectly recorded under a single patient's PMI identifier. Since the PMI helps link patient data across systems, patient safety can be compromised if data are missing or are attributed to the wrong patient, resulting in problems such as misdiagnosis and missed laboratory results [2]. Planning and policy making is also compromised because of under- or over-estimation of the PMI, since existence of duplicates can inflate the PMI size leading to inaccurate computation of vital health measures: the overall burden of duplicates for most organisations is between 8 and 12 percent [1]. Some American studies have estimated the cost of just storing one duplicate record to be more than USD 50, rising to much more for data processing for duplicate records [2, 3].

Record linkage, also known as data linkage, is one of the core aspects in data integration undertakings. It is the process of identifying records that can be attributed to the same individual or entity. When individual data sets contain a UPI, linking records from disparate data sets becomes a simple join exercise. However, in most cases unique UPI are seldom available across all datasets hence the need to use 'quasi-identifiers' to enable record linkage, whereby personal information such as name, surname, date of birth, sex and telephone number can be used for linking person-level data in the absence of unique UPI.

Efficient and accurate record linkage of public health data is important especially in low to middle income countries as it provides the basis for population-based Electronic Health Records (EHR) which can act as a low cost alternative to censuses and other costly large-scale longitudinal surveys [4–6]. Another direct benefit of EHRs is improved clinical care, as patients' historical health records become easily accessible to health service providers during care provisioning as data from different facilities gets linked through the use of a unique patient identifier [7]. Problems such as censoring due to patients migrating from their primary facility are minimised as data get integrated and facilities get connected.

The provincial health data centre, at the western cape government health department, South Africa

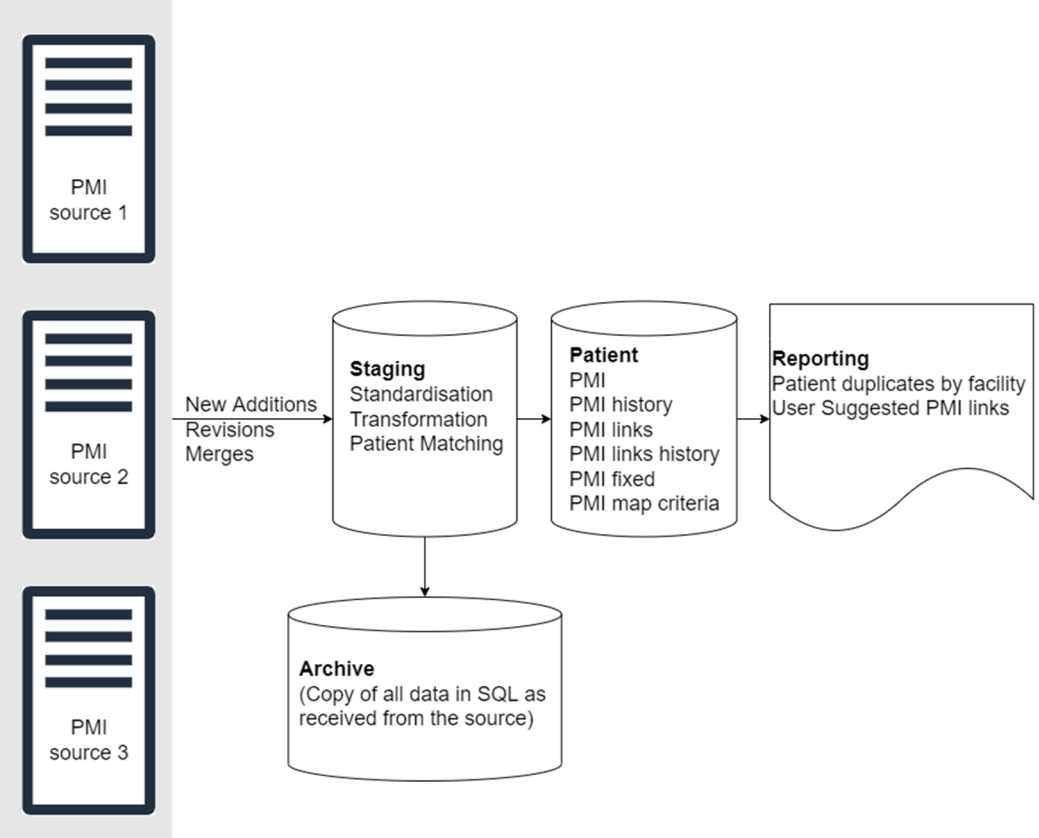
The Western Cape Government Health (WCGH) Department in South Africa created the Provincial Health Data Centre (PHDC) as a hub for hosting and integrating all health data from the multiple health information systems that are currently implemented in the province [7]. The PHDC consolidates data from the disparate administrative and clinical health data systems into an enriched dataset from which various data products and tools are then developed to support clinical care and epidemiological requirements in the department [7].

The WCGH maintains a real-time PMI system that creates a UPI for all new patients who register at public health facilities in the province, which also gets shared with most electronic health information systems. The PMI thus enables interoperability and connection of disparate records in order to improve continuity of care and health service delivery. The PHDC mainly utilises the UPI that is created by the hospital electronic platform in the province (CLINICOM) and shared with the other health systems to integrate data from a variety of source systems. This unique identifier is not, however, universally available across systems and sometimes patients do have multiple UPIs due to a variety of reasons. A record linkage programme has been developed to address this issue by resolving multiple UPIs to single individuals.

On average approximately 1500 new records are added daily to the WCGH PMI. Details from all newly added records, revisions to existing records and information on merged records are extracted and ingested into the PHDC tables via stored procedures as shown in Figure 1. The daily extract of raw data is written to the archive table for storage and to the staging table for further processing. PMI data sent to the staging tables undergo all necessary transformations and linkage before they get sent to the patient database where the records are appended to the PMI at the PHDC. Transformations undertaken include standardisation of dates and other variables to align the data to the specific data types and format consistent with those of the destination tables. A PMI record consists of 3 components kept in different tables namely: demographics, address, and contacts data. The demographic records are kept in the *PMI History* and *PMI* tables, addresses are sent to the *PMI Address* table and contacts to *PMI Contacts* table (as shown in Box 1). The reason for separating addresses and contacts from other patient information is that the former sometimes change and it is important to keep the change history, but doing so in a de-normalised table can result in redundancy because details like name, surname, gender and date of birth seldom change.

In this paper, we explore the underlying causes for multiple UPIs and describe the infrastructure, workflows and processes that are necessary for the functioning of the record linkage scheme as implemented at the WCGH (PHDC). We also analyse the performance of the current linkage algorithms

Figure 1: PMI architecture at the PHDC



using data from patients who registered in the provincial PMI system between 2015 and 2020.

The Western Cape public health system has different levels of care. Primary health care is provided to all residents for free and is obtained at clinics that are owned at different levels of governance, namely municipality/metro and provincial government. The other levels of care are secondary, tertiary, and quaternary and these services are offered at public hospitals where patients get different levels of subsidies depending on their income levels. The PMI links to primary healthcare platforms at clinics and community health centres as well as hospital information systems.

The province PMI size is currently around 15 million records with an estimated duplicates proportion of 10 percent. The province's active healthcare client population is estimated to be just below 7 million and the difference between the PMI figure and this number is due to permanent and transitory migration, duplicates and death [8]. Duplicates generally arise from clerical errors that occur during patient registration. When a patient arrives at a facility, a search is run on the PMI based on UPI or national identity number or patient's demographics (alpha search), and a new record is created if the search does not yield any result. However, an existing patient record may not be found if clerical errors such as misspelling of names or transposition of letters are made on search terms or the original entries, leading to creation of duplicate records. Patients can also contribute to duplicate record creation by providing details that are different to those given on initial registration. Middle names, nicknames and new surnames are some of the patient-side causes of duplicate records. Duplicates also result from technical and systemic challenges such as connectivity problems and limited additional electronic gatekeeping in the source systems when additions are made following a failed search.

Linkage process at the PHDC

The PHDC currently uses a multi-step deterministic (rule-based) approach to link patient data that are routinely collected from health information systems in the Western Cape province of South Africa. The rules or criteria (currently 48) used in the PHDC linkage implementation are as listed in Appendix A. The PMI data and linkage system have been implemented and migrated through different SQL server environments over the years from SQL server 2012, 2014 and 2017. Figure 2 shows a graphical presentation of the PHDC's PMI linkage process.

The process of identifying duplicate records involves comparing pairs of records from different sources to establish if they are related. Two linked records are classified as a match if the values in the assessed fields are the same or if they reach or exceed chosen similarity scores. In this study, registration details of all new patients were compared to those of patients who already existed in the PMI in order to identify duplicate records.

It is not always practically feasible and computationally efficient to compare each record against all records in the other table as this would result in assessing $(n \times m)$ pairs from two tables with n and m rows, respectively. So, the comparison is performed with the aid of blocks which help reduce the

Box 1: Tables used for storing the PMI and running the linkage process

The tables which hold PMI related information are contained in the *Patient* database together with linking-related tables and other event-specific tables which include mortality, births and family links tables.
PMI History – table contains records from different sources, history of any revision done to demographics of original records. Most patients have records across different sources or systems. The systems represent different levels of care. Records for the same patient usually share the same unique UPI and it is useful to keep all these records as they are presented in source systems to enrich the linking process since the values sometimes differ by source and subscribing sources are not always allowed to write updated information back to the main system. PMI linkage utilises both the PMI and PMI history tables. The PMI data are imported from patient registrations; for example, CLINICOM, the Patient Registration and Health Management Information System (PREHMIS), and PHCIS.
PMI – contains patient demographic information where each UPI is represented once, and the best values are selected from the different source for that record as presented in the PMI history.
PMI Addresses - contains all addresses ever recorded for each person as presented in each source. The table also keeps the address change history. When a new address is added the previous one is deactivated, and the updated address becomes the current address.
PMI Contacts – telephonic contact records are kept in the *PMI Contacts* table and values are recorded for each contact type, that is: mobile, home and work contacts. Contacts change history is also recorded and active contacts are presented for each contact type and contact source.
PMI Links (also known as *map.pmi*) – contains the linked pairs for all high confident links as obtained from the linkage process and self-self-mappings for all UPIs. The table also contains dominant UPI which are chosen as identifiers for clusters or groups that are computed on related records based on an algorithm that uses graph theory. The linked pairs often form transitive relationships meaning that some records get linked to others by association, for example, if A links to B and B to C then A is related to C and all of them are placed in the same cluster or grouping, and a dominant UPI is then chosen for the cluster. A dominant UPI should usually be the mostly commonly used and most recently used UPI in the cluster. Besides transitivity, the other constraints that need to be considered when resolving clusters is exclusivity. This occurs in cases where A is related to B then B should not be related to C. The problem posed by the constraints is commonly known as transitivity closure and can be resolved using graph database solutions. We implemented a User-Defined Application , SQL Server, and JSON approach to resolve transitivity closure within linkage clusters ....[9].'
PMI fixed – contains confirmed matches which should always be linked (green list) or unlinked (red list). The confirmed matches include merges confirmed by source systems. The confirmed matches are not 'physically merged' in the PHDC data tables but the details of the permanent relationships are stored, and the match is considered in linkage processing. Duplicate records identified at the time when a patient registers at a facility are merged via a manual process. The merging process can only be initiated once the clerk confirms with the patient that the records are indeed duplicates by verifying identification and address documents which will then be sent to the system people for further verification and processing. Merged records are extracted into the PHDC system with at most a day's delay and are taken to the PMI fixed as whitelisted pairs. It is important to keep a record of the merged pairs in case both records have clinical information.
Map Criteria – stores criteria used by the linkage algorithm
PMI links history – this table stores all matches against source identifiers so that they do not have to be re-linked each time the same patient identifiers are received from a given source. This history is useful for audit, and efficiency.

search space by identifying candidate pairs that have some form of relationship. The linkage criteria act as blocks to enable efficient processing. All new and revised records are run against the whole PMI with each criterion processed as a step in the linkage process.

The linkage criteria are ranked and processed in descending order of strength or 'fidelity'. The processing runs in a stepwise manner with high-fidelity criteria processed first, and a record is removed from being a candidate for subsequent steps once it has found a link on a higher fidelity criterion.

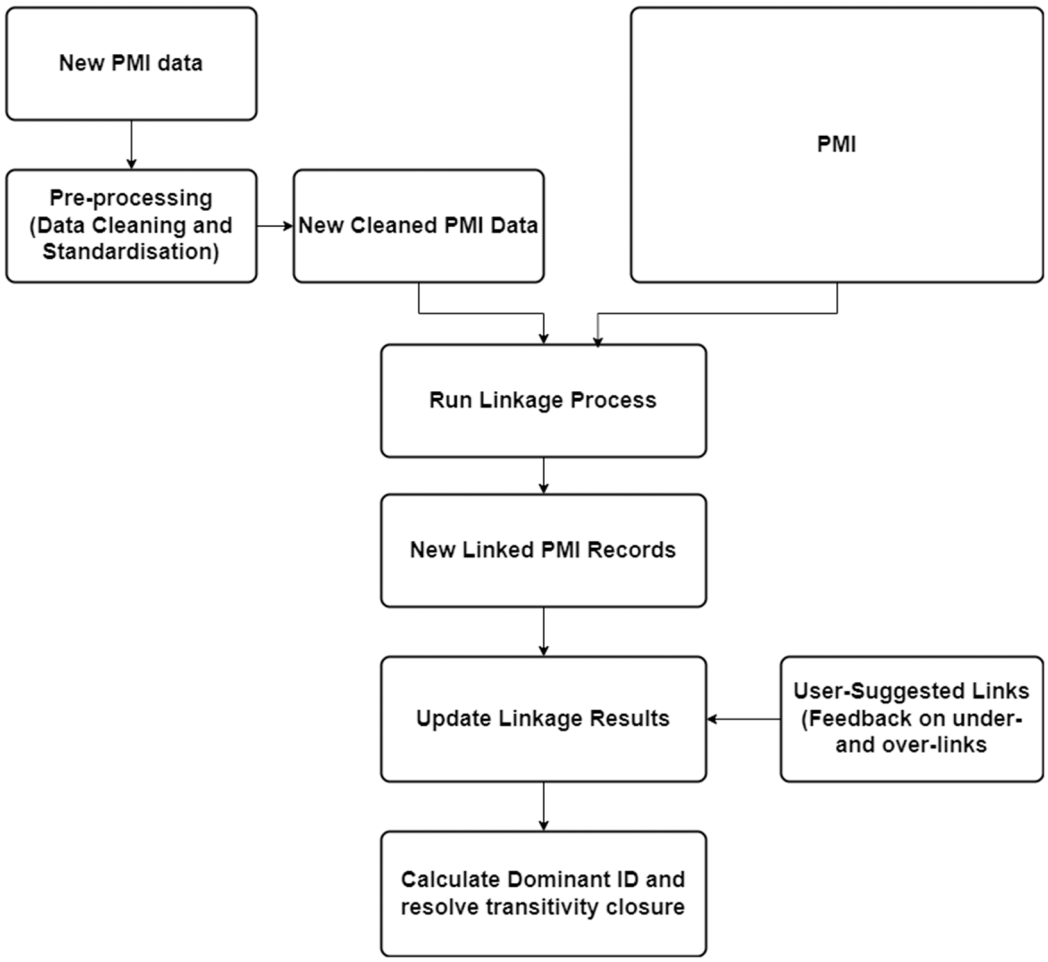
Linkage criteria are classified as follows: exact, highly probable, probable, highly possible, possible and low possible. Only duplicates identified by criteria classified as exact, highly probable, probable, and highly possible are included in this study as they are considered more likely to be high-fidelity matches. During the routine linkage process, all weak linkages are sent for manual review and only the top link established by a higher fidelity criterion for each new PMI record is taken to the linkage results table. No physical merging of the records is done based on the linkage results, but all linked pairs of UPIs are stored in a mapping table. One PMI identifier is chosen to be the dominant identifier among the cluster of PMI identifiers that map to the individual patient. The mapping process is dynamic, in that the dominant identifier for a set of linked patients changes with changes to the cluster.

Methods

Patient data

Permission to conduct this study was provided by the Western Cape Department of Health, and ethics approval was given by the University of Cape Town Faculty of Human Research Ethics Committee. The study uses data from all new patients who were registered at public health facilities in the Western Cape Province of South Africa between 1 December 2015 and 31 October 2020. The study period was selected to start at

Figure 2: Overview of the PHDC PMI linkage process



the same time that the PHDC patient matching was fully implemented until the time of data extraction. This study was conducted by employees of the Western Cape Government, working within the secure PHDC environment hosted by the Western Cape Department of Health. No individualised data were downloaded or used outside of the routine health data environment managed by the department. The reported linkage algorithms were run within the secure environment and only aggregated findings reported for publication and open sharing.

Data used in the linkage process are obtained at the time of registering a new patient in one of the province's three main PMI registration systems which all write to the CLINICOM system which maintains the province's PMI. The three systems where PMI registration is done are: CLINICOM – a platform mainly built for managing patients at provincial hospitals, the Primary Health Care Information System (PHCIS) and Patient Registration and Health Management Information System (PREHMIS) – purpose built platforms for primary health care facilities managed at provincial and city level, respectively.

Linkage variables

The fields used in the linkage process are: RSA ID – which is a National Identifier assigned to each South African individual on registration of birth by the Department of Home Affairs; date of birth; year of birth – this is derived from data provided in the date of birth field; month of birth – this is also derived from data provided in the date of birth field; day of birth - derived from data provided in the date of birth field; surname; first names; gender; address line 1; address line 2; address line 3; address line 4; post code; home phone; work phone, and mobile phone.

Descriptive analyses

All data used in this study were collected from the PMI registration systems and stored in a SQL Server database and were retained within the secure PHDC system for the analyses. Computer codes written in T-SQL were run against the data to produce aggregates and descriptive analyses.

Monthly duplicate trends

The proportions of duplicate records for each month were calculated by dividing the number of duplicate records created in the month by the total of new PMI records created in that month.

Proportion duplicates by linkage criteria

Different combinations of these criteria are used to assess likelihood of duplication, and each combination is assigned an

estimation of the strength of the assessment of duplication. The proportion of duplicates for each linking criterion is calculated by dividing the total number of linked pairs for each criterion by the total number of linked pairs.

Completeness of linkage variable

Variable completeness is important in informing the choice of which variables to use for linkage. The proportion completeness for each variable used in the linkage process was calculated by dividing the total number 'non-empty' records by the total number of registered PMI records.

Number of duplicate records per patient

The distributions of patients by the number of UPIs were calculated by first creating clusters of each record and the corresponding duplicate(s) if they exist then expressing the number of clusters of the same size as percentage of the total PMI.

Estimating extent of errors among linkage variables

Field values for all linked records were compared between the original and duplicate record for all variables used in the linkage process to determine the causes and extent of errors. The proportion of duplicate pairs with mismatching values are given by dividing the total count of linked pairs with mismatching values by the total number of linked pairs.

Results

There were 2,107,930 PMI records created between 1 December 2015 and 30 October 2020 and a total of 290,249 probable duplicates were identified by the PHDC algorithm in the same period. The graph in Figure 3 shows a steady decline in the proportion of duplicates from just below 16.8% in December 2015 to 9.6% in October 2020.

The distribution of duplicates by linking criteria (table 1) showed 60.4% highly possible matches with 23.1% of these based on exact date of birth, exact first names and a similarity score on surnames that is greater or equal to 0.85. Probable matches had 35.6% with majority (27.7%) of these matches based on exact first names, exact surname and exact date of birth and same values on South African national identity number. Highly probable matches amounted to 4.0%, with the majority based on the exact South African national identity number, and similarity scores for surname and date of birth greater or equal to 0.85 and 0.95 respectively.

The results in Figure 4 show the extent of completeness of patient attributes which are used in the linking process. Folder number (100%), first names (100%), surname (100%), sex (100%), date of birth (100%), postal code (94%) and address line4/town (94%) all had completeness proportions of over 95%. Only 29% of patient records have a South African national identity number.

The graph in Figure 5 shows the proportion of discrepancies among field values: field values were compared for each linked pair to check for the extent of discrepancies among linking variables. The graph shows that most variations are in the first name and surname with approximately 22% of name values in linked pairs not being the same. The third highest proportion of discrepancy was in the date of birth with 14%. Of the pairs, 19% did not show any difference in any of the fields considered in the analysis.

Discussion

In this study, we found that the proportion of duplicates within the period December 2015 to October 2020 is between 16.8 and 9.6 percent. We also showed that most common errors in patient records are misspellings of patient first name and surname and that improvements in patient registration and presence of a unique patient identifier helps to facilitate linkage since only around 30 percent of the records have the RSA ID. Besides the RSA ID, most of the fields used in patient linkage have good completeness, for example, first name, surname, sex, date of birth, and folder number are all 100 percent complete although the quality of some field values is poor due to use of placeholder values.

Between 2015 and 2020, the observed downward trend in monthly duplicates presented in Figure 1 could be due to several factors: there is ongoing oversight of the integration of PMI data in the province by a technical team with representatives from each of the contributing platforms and data systems. The level of engagement around PMI issues has improved and around mid-2016 a monthly report of duplicates created at each public health facility was distributed by email to facilities managers by this team. This report was developed based on the results from the PHDC matching process and is used to raise awareness around duplicates issues as well as identifying training needs, as it also shows the user account responsible for creating the duplicate allowing for targeted interventions as they are required. The duplicates report was subsequently also made available on demand through the health department SharePoint site at the end of 2016. The frequency of feedback from system users has also generally increased, as seen from increased traffic in support emails coming to the PHDC.

The uneven data in some months can be directly linked to known events: the spikes in the months of October 2016 and September 2017 were due to the rollout of the CLINICOM system at two district hospitals. These facilities had elevated duplicate rates as they were not previously online hence it had been difficult to manage records with no access to the centralised PMI system. One system alone had 39% PMI duplicates before integration, thus skewing the aggregated duplicate proportion for the month. In 2017, for example, when one of these district hospitals was brought online, more than 25,000 new PMIs were uploaded in a batch causing a spike in duplicates to 25% for that month, but in the subsequent month the duplicate proportions stabilised again at approximately 14%.

The true duplicate proportion could be a bit higher than the level presented as most baby records are not included in the analysis. Records for new-born babies are initially assigned place holder values as they get created before the infant is named. Most of these records end up being 'orphaned' because another record is usually created when the infant returns to health facilities and registers on proper names. The system could be modified to force integrity between all infant and

Figure 3: New patient registrations and proportion of duplicates created at public health facilities between December 01, 2015 and October 31, 2020

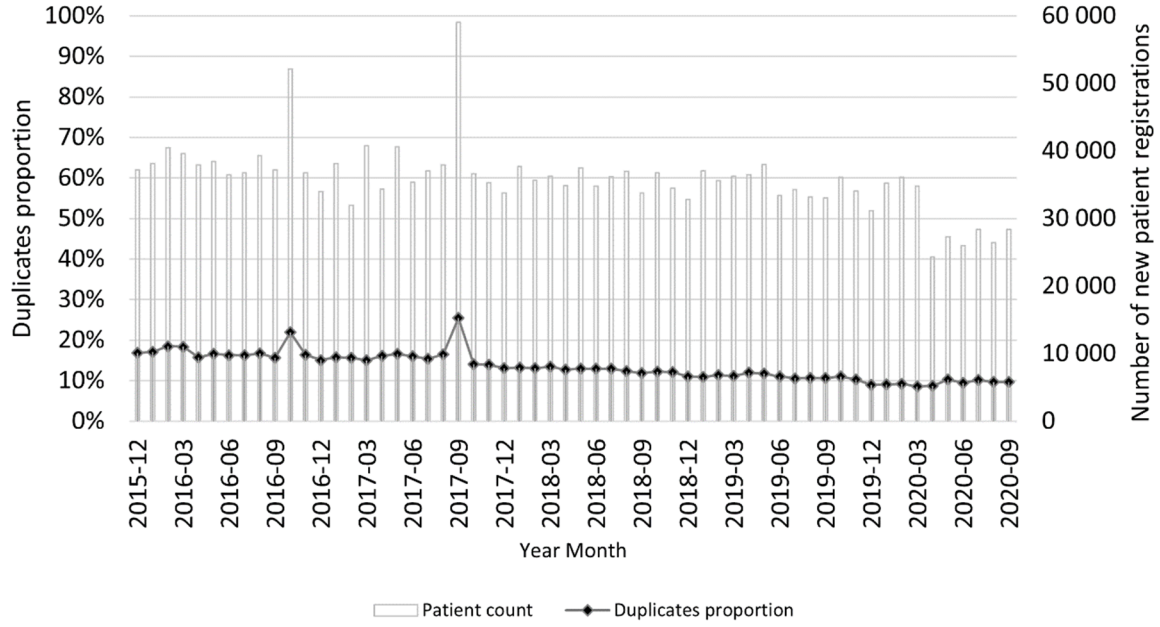


Table 1: Distribution of duplicates identified between 2015 and 2020 by linkage criteria

Mapping criteria	**Matching pairs**		
High probable	**N**	**Within category (%)**	**Overall (%)**
exact RSA ID, *simscore(surname) >= 0.85, simscore(date of birth) >= 0.95	10 070	86.21	3.47
exact RSA ID, simscore(firstnames) >= 0.85	748	6.40	0.26
exact RSA ID, simscore(date of birth) >= 0.95	640	5.48	0.22
exact RSA ID, switch(exact surname, exact firstnames)	206	1.76	0.07
exact RSA ID, simscore(surname) >= 0.85	17	0.15	0.01
sub totals	**11 681**	100.00	4.02
Probable			
exact firstnames, exact surname, exact date of birth (incl RSA ID check)	80 386	77.86	27.70
switch (exact surname, exact firstnames), exact date of birth	22 861	22.14	7.88
sub totals	**103 247**	100.00	35.57
High possible			
exact date of birth, exact firstnames, simscore(surname) >= 0.85 (incl RSA ID check)	67 102	38.27	23.12
exact surname, exact date of birth, simscore(firstnames) >= 0.85 (incl RSA ID check)	59 622	34.01	20.54
exact surname, exact firstnames, simscore(date of birth) >= 0.95 (incl RSA ID check)	42 324	24.14	14.58
exact firstnames, simscore(surname) >= 0.85, simscore(date of birth) >= 0.95 (incl RSA ID check)	6 273	3.58	2.16
sub totals	**175 321**	100.00	60.40
Grand Total	**290 249**		**100.00**

child records by ensuring that records for children below a certain age, for example 12 years, carry details, especially folder number or RSA ID of the mother to ensure linkage with maternal, infant and child record.

The causes of duplicates for linkage fields values that match exactly could be due to challenges faced by registrars with retrieving existing folders such as shortage of records management personnel and the length of time it takes

Figure 4: Percent completeness among linkage variables using data from new patients who registered between 2015 and 2020

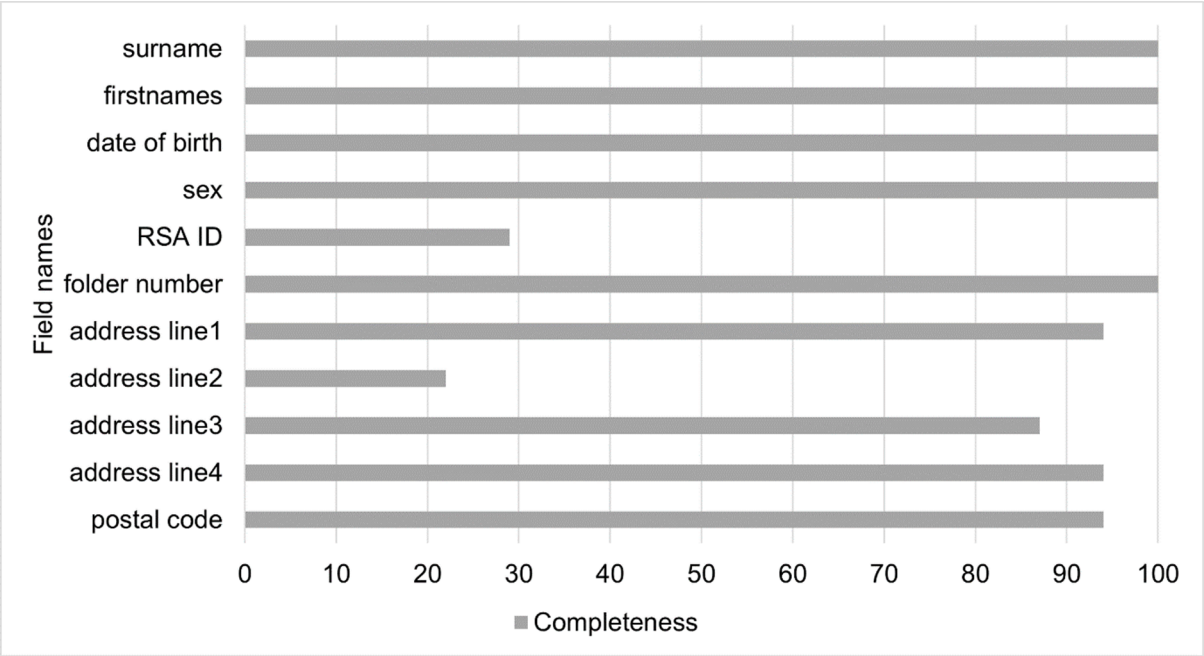
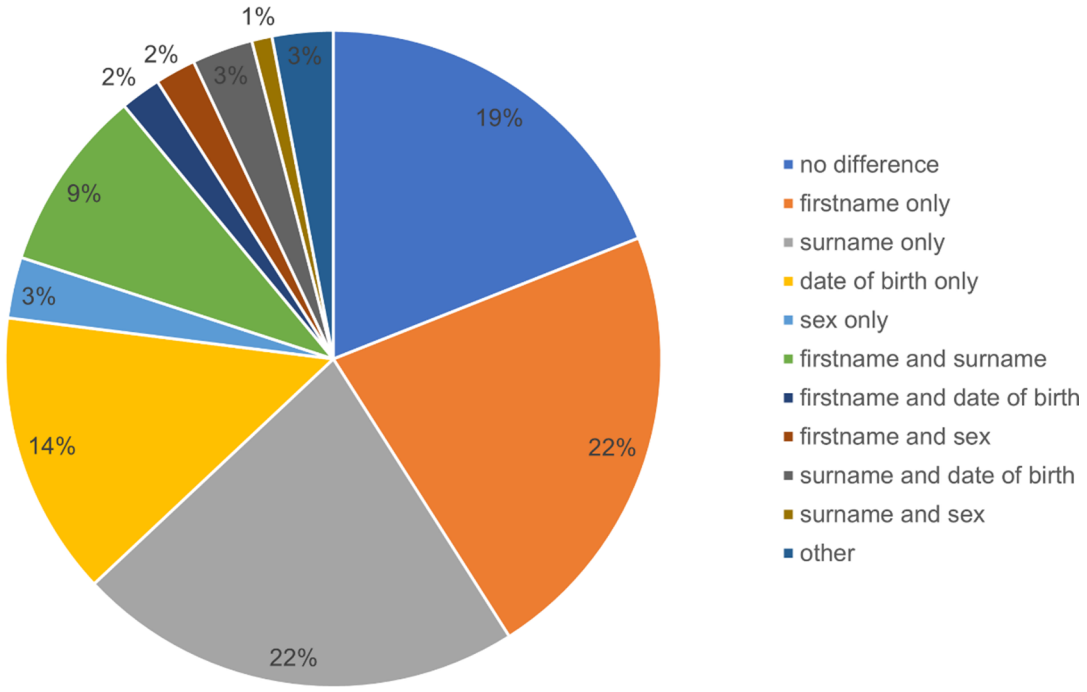


Figure 5: Proportion of comparison pairs with value mismatches among linkage variables for duplicates identified between 2015 and 2020



to retrieve an existing folder which means they resort to creating new records. There have also been reports of duplicate submissions occurring due to system connection issues that cause a slow response from the web service leading to administrators making multiple clicks to submit the registration. The high proportion of duplicates could drop drastically if the patient registration systems run a 'deep search' to check again before committing a new patient to the database.

Duplication is caused by variety of reasons which include clerical errors like spelling errors on first names and surnames, use of nicknames for example "Ntombi" instead of "Ntombifuthi", first name and surname switches, middle name omission in one record and use of placeholder values. Surnames are almost always affected by similar problems as first names, however, one issue that mostly affect surnames, although at a small scale, is surname changes mainly due to marriage as well as some other reasons. One of the main causes of discrepancies

in first name and surname fields is phonetic errors which arise because of how the names are typed depending on how they sound to the registrar or clerk. Common spelling errors like "Khumalo" instead of "Kumalo" are a result of the phonetics and pronunciation of names. Date of birth is mainly affected by digit transposition and switching day and month values, for example, 2002-03-02 and 2002-02-03. South African Identity Number is the least affected in terms of showing errors that might lead to duplicate creation, but it is not populated in most of the records. The issue is compounded by limited use of identity documents which would drastically improve linkage in the absence of the unique UPI if they were used more widely.

In the private sector, billing systems and healthcare insurance reimbursements mean that record keeping is much more stringent; but especially at primary care level in the public sector no billing takes place and record keeping is not as stringent. Interestingly, in tertiary care there is a sliding scale for payment based on income of the health care client, so data capture in the CLINICOM system is slightly more robust as there is need to capture information for client billing.

Linkage at the PHDC is also affected by completeness of important variables like folder number, first name, surname, sex, and date of birth, which all have 100% completeness because they are required fields. However, not all the information entered is valid or is of good quality, for example, placeholder values such as "baby of", and "unknown", are observed in the first name field. It is also difficult to assess the quality of numeric values, except for folder number and South African identity number which can be validated using a checksum test on the last digit [10]. The South African national identity number has poor completeness because it is not mandatory for a patient to provide one when they need to access healthcare. There is suspicion that one of the reasons patients are reluctant to share identity details is partly because of the history of distrust with government systems in South Africa, arising from notorious pre-Democracy "pass laws" and abuse of identity documents by the prior Apartheid government. Also, poor South African identity number completeness is due to undocumented individuals, both those who are eligible for SA citizenship but were never registered with the Department of Home Affairs (DHA), but also a large number of immigrants and refugees in the country coming from the rest of Africa who do not yet have documentation in the country. The numbers of undocumented individuals may be small, but these are often vulnerable individuals who may have particular health needs, and we cannot afford for them to 'slip through the cracks' of health care because of registration and data linkage issues.

Most address fields suffer from incompleteness because most of the time the whole address is entered in the first address field instead of populating each address component in the appropriate field. Colloquial or unofficial names for geographical locations are also in common use especially in areas of informal settlements. The quality is especially poor for historical addresses because most systems predated the availability of APIs to normalise and verify addresses in real-time, and system enhancements are currently underway to resolve addresses at the time of registration wherever possible.

The current linkage process mainly uses Jaro-Winkler algorithm for text comparison as it was previous shown to work well [6] and since the available phonetic encoding techniques are not well suited for non-English names [11, 12] but further research will be done on the data to find the best combination of both phonetic encoding and text edit distance techniques. We have developed a machine learning model and a probabilistic record linkage scheme, based on the Fellegi-Sunter algorithm [13], which is undergoing final review before deployment. PHDC linkage is also affected by data quality-related issues that are associated with the routine nature of the data which is often difficult to validate and enforce integrity at the point of data capture. Missingness in some linkage variables as well as the need to link datasets without PMI identifiers support the need for a probabilistic approach.

Reporting

The PHDC has so far created reports on potential duplicates that are available for use by facility officers. The main one is the *patient duplicates by facility* report which provides details of all potential duplicates created per day at each public health facility and is available to approved users on a SharePoint site. The report gives details for all high confidence duplicate candidates, facility where the duplicate was created and the user who created the record. This provides a mechanism for evidence-based decision support for facility managers to see, for example, if there is need for further training for information personnel. Another report is the suggested links report which allows users to notify the PHDC of over- or under-linking cases they are able to confirm.

There are a limited number of health exchanges doing similar linkage at this scale for example, in Canada and Australia. Some of the successes from better data linkage in Australia, for example, are noted by Smith and Clark [14] in the 50-year review of the progress made in the area of data linkage including important contributions in public health such as establishing the teratogenic effects of maternal diet [15]. Record linkage programs are well advanced in Australian states and key strides have been made in establishing a national linkage resource via the Population Health Research Network [14]. In the global South, however, these kinds of initiatives are not as common. In many cases there is local provision of health care using local electronic administrative platforms, but electronic medical records (EMRs) tend to be based on District Health Information Software (DHIS2) and do not integrate multiple data resources with a daily frequency as done by the PHDC. Another South African study applied a novel graph-based linkage scheme to link nationwide laboratory data to create a national HIV cohort [16].

The data linkage achieved at the PHDC is ground-breaking for an LMIC country, and some of the reasons it is possible are: Firstly, the data have been managed at a provincial level which has been slightly more manageable than national scope data integration; secondly, there is a long term legacy of a single health identifier in the province because of the CLINICOM system as well as integration of this PMI with other health information systems such as PHCIS and PREHMIS combined with a mature civil registration system that provides a reliable RSA ID, date of birth and capture of standardised name and other personal information from birth (people have a legal name and captured date of birth); thirdly, an existing electronic health data infrastructure in the province has provided a secure environment to build the linkage systems; and finally there has

been an opportunity to build in-house capacity using research-related funding to develop processes such as linkage to improve the PHDC data offering.

Conclusion

Overall, data linkage at the PHDC is improving over time, as seen by the trend of reducing duplicates over the past six years. This improvement in the potential for record linkage could be a result of improved patient registration. We can see disruptions caused by changes to the system, but we anticipate these will have less impact over time as the systems mature and undergo further refinements including leveraging facility visit information and clinical information to help strengthen confidence in matches. Figure 5 shows clearly that our weakest link is first names and surname matching. This highlights that how methods capture and link South African names in the data is key to improving linkage algorithms, and understanding the characteristics of data linkage successes and challenges at the PHDC can inform interventions by the WCGH to improve data collection within public health facilities. Linkage algorithms currently in use attempt to deal with spelling errors in strings which often use the way the strings sound as a basis for linkage by developing a code based on the consonant of the string; but these algorithms were developed and validated based on anglicized names and do not always work for local names. Therefore, it is worth exploring ways to improve phonetic matching for local names.

Record linkage has recently increased in popularity as organisations become data driven in their approach to daily operations, highlighting the large challenge affecting implementation of record linkage endeavours through the lack of universally available unique identifiers across disparate datasets. It is therefore imperative to explore other available variables that can be used as *quasi-identifiers*. Most variables that are available are, however, compromised with many errors arising largely from data capturing errors.

This study explored how a record linkage programme is set up in a public health environment and highlighted the extent of errors affecting linkage variables as used in health record linkage implementation. The work also provides an understanding of record linkage in the PHDC and will provide a basis for further explorations on the impact of data quality on our ability to successfully link and deduplicate records. Understanding the current characteristics of data linkage will provide a baseline against which to assess the future success of ongoing interventions to improve data collection at source. Furthermore, this linkage process in the PHDC has demonstrated that data linkage in a health information exchange can be realised in LMICS, especially where conditions include existing electronic health data systems alongside the implementation of a robust health identifier separate to civil registration and other national identifiers. Although the Western Cape Province's unique patient identifier (folder number) is imperfect it has assisted tremendously in the linkage process and the process can only improve as more functionality is added, especially the functionality to perform prospective linkage which will allow for a deeper search at the point of registration. The national patient identifier also known as the health patient registration system (HPRS) will also go a long way towards improving patient identity and consequently patient linkage when it gets fully implemented [14, 17].

Ethical approval

The study received approval from the Health Research Unit at the Western Cape Provincial Health Department and from the Human Research Ethics Committee at the University of Cape Town (HREC REF: 141/2020).

Data availability statement

The analysis presented here was undertaken within the secure environment of the Western Cape Government Health Department, and no independent data set was generated because of the uniquely identifying nature of these data. For queries about the dataset, please contact Prof Andrew Boulle andrew.boulle@westerncape.gov.za.

Acknowledgements and funding

NT acknowledges receiving funding from the Bill & Melinda Gates Foundation (The African Data and Biospecimen Exchange, *INV-037558*), the CIDRI-Africa Wellcome Trust grant (*203135/Z/16/Z*) and the NIH H3ABioNET award (*U24HG006941*). AB and TM are supported by funding from the US National Institutes of Health (R01HD080465, U01AI069911), Bill and Melinda Gates Foundation (1164272; 1191327; INV-004657, INV-017293), the Wellcome Trust (203135/Z/16/Z), the United States Agency for International Development (72067418CA00023). The authors thank and acknowledge the contributions of our Western Cape Government colleagues, especially Boldrin Erasmus and Teresa Kruger.

Conflict of interest statement

The authors declare no conflicts of interests.

References

1. How to Maintain a Clean Master Patient Index. [cited 15 Mar 2021]. Available: https://www.ironmountain.com/resources/general-articles/h/how-to-maintain-a-clean-master-patient-index.

2. Joffe E, Bearden CF, Byrne MJ, Bernstam EV. Duplicate Patient Records – Implication for Missed Laboratory Results. AMIA Annu Symp Proc. 2012;2012: 1269–1275.

3. Biddle M. Maintaining the Master Patient Index: The impact of patient registration processes on data integrity. MS, University of Tennessee Health Science Center. 2015. https://doi.org/10.21007/chp.hiim.0016

4. Howe GR. Printed in U.S.A. Use of Computerized Record Linkage in Cohort Studies.

5. Saraswat L, Ayansina DT, Cooper KG, Bhattacharya S, Miligkos D, Horne AW, et al. Pregnancy outcomes in women with endometriosis: a national record linkage study. BJOG Int J Obstet Gynaecol. 2017;124: 444–452. https://doi.org/10.1111/1471-0528.13920

6. 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: 1.

7. Boulle A, Heekes A, Tiffin N, Smith M, Mutemaringa T, Zinyakatira N, et al. Data Centre Profile: The Provincial Health Data Centre of the Western Cape Province, South Africa. Int J Popul Data Sci. 2019;4. https://doi.org/10.23889/ijpds.v4i2.1143

8. StatsSa; Mid-year population estimates 2020. Available: http://www.statssa.gov.za/publications/P0302/P03022020.pdf

9. Mauri D. Transitive Closure Clustering with T-SQL, SQLCLR and JSON: yorek/non-scalar-uda-transitive-closure. 2018. Available: https://github.com/yorek/non-scalar-uda-transitive-closure

10. Decoding your South African ID number. In: Western Cape Government [Internet]. [cited 7 Sep 2021]. Available: https://www.westerncape.gov.za/general-publication/decoding-your-south-african-id-number-0

11. Ndyalivana Z, Shibeshi Z. Development of Soundex algorithm for isiXhosa language. 2014.

12. Christen P. A Comparison of Personal Name Matching: Techniques and Practical Issues. Sixth IEEE International Conference on Data Mining - Workshops (ICDMW'06). Hong Kong, China: IEEE; 2006. pp. 290–294. https://doi.org/10.1109/ICDMW.2006.2

13. Fellegi IP, Sunter AB. A Theory for Record Linkage. J Am Stat Assoc. 1969;64: 1183–1210. https://doi.org/10.1080/01621459.1969.10501049

14. Smith M, Flack F. Data Linkage in Australia: The First 50 Years. Int J Environ Res Public Health. 2021;18: 11339. https://doi.org/10.3390/ijerph182111339

15. Bower C, Stanley FJ. Dietary folate as a risk factor for neural-tube defects: evidence from a case-control study in Western Australia. Med J Aust. 1989;150: 613–619. https://doi.org/10.5694/j.1326-5377.1989.tb136723.x

16. Bor J, MacLeod W, Oleinik K, Potter J, Brennan AT, Candy S, et al. Building a National HIV Cohort from Routine Laboratory Data: Probabilistic Record-Linkage with Graphs. bioRxiv. 2018 [cited 3 Sep 2019]. https://doi.org/10.1101/450304

17. Beck EJ, Shields JM, Tanna G, Henning G, de Vega I, Andrews G, et al. Developing and implementing national health identifiers in resource limited countries: why, what, who, when and how? Glob Health Action. 2018;11: 1440782. https://doi.org/10.1080/16549716.2018.1440782

Appendix A: Patient linking rules used at the PHDC

No	Criteria strength	Match description
1	exact	Exact match on folder number, surname and date of birth
2	highly probable	Exact match on folder number, (year of birth and day of birth) or (month of birth and day of birth) and (year of birth and month of birth) and JW(surname) >= 0.85
3	highly probable	Exact match on SA ID, JW(surname) >= 0.85 and JW(date of birth) >= 0.95
4	probable	Exact match on folder number, firstnames, surname
5	highly possible	Exact match on folder number, surname, JW(firstnames) >= 0.85
6	probable	Exact match on folder number, JW(surname) >= 0.85, JW(firstnames) >= 0.85
7	possible	Exact match on SA ID, surname, firstname
8	possible	Exact match on SA ID, surname, firstname (switched comparison of surname and firstnames)
9	highly probable	Exact match on SA ID, JW(surname) >= 0.85
10	highly probable	Exact match on SA ID, JW(firstnames) >= 0.85
11	probable	Exact match on SA ID, JW(date of birth) >= 0.95
12	exact	Exact match on folder number, fullname (firstnames + surname), date of birth
13	highly probable	Exact match on folder number, date of birth, JW(fullname (firstnames + surname)) >= 0.85
14	highly probable	Exact match on folder number, JW(date of birth) >= 0.85, JW(fullname (firstnames + surname)) >= 0.85
15	highly probable	Exact match on folder number, fullname (firstnames + surname)
16	probable	Exact match on folder number, JW(fullname (firstnames + surname)) >= 0.85
17	probable	Exact match on folder number, fullname contains surname (wildcard command)
18	probable	Exact match on folder number, date of birth
19	probable	Exact match on folder number, surname
20	probable	Exact match on folder number, firstnames
21	low possible	Exact match on folder number, surname, SA ID (whether null or not null)
22	low possible	Exact match on folder number, surname (switched firstnames and surname), SA ID (if or not null)
23	low possible	Exact match on date of birth, surname, JW(firstnames) >= 0.85, SA ID (if or not null)
24	probable	Exact match on SA ID (if not null), JW(surname) >= 0.85, JW(firstnames) >= 0.85
25	low possible	Exact fullname = firstnames + surname
26	possible	Exact match on date of birth, firstnames, JW(surname) >= 0.85, SA ID (if not null), JW(address) >= 0.95 (concatenate address fields from address line 1 to 4)
27	highly probable	Exact match on firstnames, surname, SA ID (if not null), (JW(date of birth) >= 0.95 OR (exact YOB and (exact MOB OR switched MOB/DOB), JW(address) >= 0.95 (concatenate address fields from address line 1 to 4)
28	highly probable	Exact match on date of birth, firstnames, SA ID (if not null), JW(surname) >= 0.95
29	exact	Exact match on surname, firstnames, YOB, (SA ID = SA ID if not null), [JW(date of birth) >= 0.95 OR (exact MOB OR switched MOB/DOB)]
30	highly probable	Exact match on date of birth, surname, firstnames
31	highly possible	Exact match on date of birth, surname, JW(firstnames) >= 0.85
32	highly possible	Exact match on date of birth, firstnames, JW(surname) >= 0.85
33	highly possible	Exact match on surname, firstnames, [JW(date of birth) >= 0.95 OR (exact YOB, exact MOB OR switched MOB/DOB)]
34	highly possible	Exact match on surname, [JW(firstnames) >= 0.85, [JW(date of birth) >= 0.95) OR exact YOB, (exact MOB OR switched MOB/DOB)]
35	highly possible	Exact match on firstnames, JW(surname) >= 0.85, [JW(date of birth) >= 0.95) OR exact YOB, (SA ID = SA ID if not null), (exact MOB OR switched MOB/DOB)]
36	possible	Exact match on date of birth, surname, JW(firstnames) >= 0.85
37	possible	Exact match on fullname vs (firstnames + surname), exact date of birth
38	possible	Exact surname, firstnames, [JW(date of birth) >= 0.85,
39	possible	Exact Surname, JW(firstnames) >= 0.85, JW(date of birth) >= 0.85 and (Exact YOB and [(Exact MOB) OR (Switched Exact DOB/MOB and Switched Exact MOB/DOB)])
40	possible	Exact firstnames, SA ID if not null and JW(surname) >= 0.85 and (JW(date of birth) >= 0.95 and (Exact YOB and [(Exact MOB) or (Exact switched DOB/MOB and Exact switched MOB/DOB)])
41	probable	Exact date of birth, firstnames and JW(surname) >= 0.85 and SA ID = SA ID if not null and JW(address) >= 0.95
42	probable	Exact surname, firstnames, SA ID if not null, and JW(address) >= 0.85 and [JW(date of birth) >= 0.95 OR (Exact YOB and [(Exact MOB) OR (Exact switched DOB/MOB and exact switched MOB/DOB)])
43	highly probable	Exact SA ID and switch on exact firstnames and surname and exact date of birth

Continued

Appendix A: Continued

No	Criteria strength	Match description
44	highly probable	Exact SA ID, JW(surname, firstnames) >= 0.85, JW(firstnames, surname) >= 0.85
45	probable	Exact switch on firstnames and surname and exact date of birth
46	highly probable	Exact SA ID, surname, firstnames
47	probable	Exact firstnames, surname, date of birth match on different lines with the same folder number
48	manual	User notifies links