Using demographic data matching to uniquely identify individuals for HIV case-based surveillance in low-resource settings, a case study from Kenya

Anthony Waruru1,§
Agnes Natukunda2
Lilly M. Nyaga3
Timothy A. Kellogg2,4
Emily Zielinski-Gutierrez1
Wanjiru Waruiru2,4
Kenneth Masamaro1
Richelle Harklerode4
Jacob Odhiambo5
Eric-Jan Manders6
Peter W. Young1

1) Division of Global HIV & TB (DGHT), U.S. Centers for Disease Control and Prevention (CDC), Nairobi, Kenya; 2) Global Programs for Research and Training, University of California San-Francisco, Kenya; 3) Ministry of Health, National AIDS and STI Control Program (NASCOP), Kenya; 4) University of California, San Francisco, California, USA; 5) Palladium, Nairobi, Kenya; 6) Division of Global HIV & TB (DGHT), U.S. Centers for Disease Control and Prevention (CDC), Atlanta, USA

§Corresponding author:
Anthony Waruru,
P.O Box 606,
Nairobi, 00621,
Kenya
Phone: +254722200179
Email: awaruru@cdc.gov

Running title: Using demographic data matching to uniquely identify individuals for HIV case-based surveillance

Competing interests: The authors have declared that no competing interests exist.

The findings and conclusions in this article are those of the authors and do not necessarily represent the official position of the U.S Centers for Disease Control and Prevention (CDC). This publication was made possible by support from the U.S. President’s Emergency Plan for AIDS Relief (PEPFAR) through cooperative agreements [GH000069-05 and U2GGH001520] from the U.S. Centers for Disease Control and Prevention (CDC), Division of Global HIV & TB (DGHT).
Abstract

Background: HIV case-based surveillance (CBS) is recommended for 2nd generation HIV surveillance. CBS is more accurate than aggregate reporting for monitoring of the HIV treatment cascade including the UNAIDS Fast-Track 90-90-90 targets. In the absence of a universal healthcare identifier (UHID), records matching, linking and deduplication in surveillance and care coordination systems may depend on string-matching algorithms.

Objectives: To compare deterministic and probabilistic algorithms in uniquely identifying HIV cases in a CBS pilot in Kenya and suggest an algorithm for deduplication and demographic data matching of individuals in low-resource settings.

Methods: HIV CBS pilot data were from 124 facilities in two high HIV-burden counties (Siaya and Kisumu) in western Kenya. For efficient processing, data were grouped into three settings: cases within a) HIV testing services (HTS) b) HTS to care; and c) within care. In deterministic matching we compared patient identifiers by direct comparison to determine matches. We used R stringdist package for probabilistic matching and Jaro, Jaro-Winkler, Levenshtein and Damerau-Levenshtein, string edit distance calculation methods. For Jaro-Winkler method, we used a penalty (p) =0.1 and applied four weights (ω) to Levenshtein and Damerau-Levenshtein: deletion ω=0.8; insertion ω=0.8; substitutions ω=1 and transposition ω=0.5.

Results: We abstracted 12,157 cases of which 4,073 (33.5%) were from HTS, 1091 (9.0%) HTS to care and 6,993 (57.5%) within care. Using the deterministic process 435 (3.6%) duplicate records were identified, yielding 11,722 unique cases. Overall, of the probabilistic methods, Jaro-Winkler yielded the least duplicate records, 618 (5.1%) while Jaro, Levenshtein and Damerau-Levenshtein yielded 678(5.6%), 771(6.3%) and 781(6.4%) duplicate records respectively. Specifically, duplicate records yielded by method were: Jaro 266(6.5%) within HTS, 59(5.4%) in HTS-care, and 353(5.0%) within care; Jaro-Winkler 232 (5.7%) within HTS, 50 (4.6%) within HTS-care and 336 (4.8%) within care; Levenshtein 314 (7.7%) within HTS, 72 (6.6%) in HTS-care and 385 (5.5%) within care; and Damerau-Levenshtein 303(7.4%) within HTS, 77(7.1%) in HTS-care 69and 401 (5.7%) within care.

Version: 16th Mar., 2018 - Page 2
Conclusions: Without deduplication, over reporting occurs across the care and treatment cascade. Of the probabilistic matching algorithms, Jaro-Winkler was the most conservative in identifying matches. In absence of widespread use of UHID, we suggest Jaro-Winkler probabilistic matching to perform persons-matching using demographic data. A pragmatic estimate of duplicates in health care settings provides useful information that can be used as a corrective factor in estimates and projections models and for targeting and program planning. Without UHID, Kenya needs a standard national deduplication and persons-matching algorithm to improve accuracy in monitoring the Fast-Track 90-90-90 targets.

Key words: deterministic matching; probabilistic matching; HIV case-based surveillance; unique case identification; universal healthcare identifier

Introduction

In sub-Saharan Africa (SSA), HIV case-based surveillance (CBS) has not yet been widely used. CBS provides the ability to link key clinical events over time for an individual patient to demographic attributes. In HIV, these methods are sometimes referred to as “second generation surveillance system” - having progressed beyond initial surveillance approaches that focused on aggregate numbers, [1,2]. Such longitudinal approaches are important for individual tracking of HIV cases from diagnosis, linkage to HIV care and treatment, antiretroviral treatment (ART), viral suppression and other outcomes including retention in care, transfer-out, loss to follow-up or death. This level of follow-up is useful for epidemiological profiles at the smallest geographical units [3], and monitoring of the HIV care and treatment clinical cascades and UNAIDS Fast-Track 90-90-90 targets, [4]. CBS has advantages over aggregate reporting systems since it is patient-based, hence treatment course and outcomes are more accurately tracked. CBS also is more accurate in showing trends and temporal sequences in the HIV epidemic for example, trends of time to linkage to treatment from HIV testing or even changes in the clinical cascade over time [5], and is feasible in low resource settings [6]. However, improved accuracy in CBS is contingent upon unique patient identification and correct record linkage from HIV diagnosis through the treatment course, due to the longitudinal nature of HIV care and multiplicity of data sources. Additionally, correct matching of cases is important for accurate aggregation.
Uniquely identifying cases in HIV case-based surveillance

Record linkage across the care continuum depends on unique patient identification due to chronic patient management within the care setting. Moreover, record linkage is useful for attaching records to locality, for example in demographic and health surveillance systems [7], for data aggregation, and to facilitate correct assessment of program coverage.

Unique record identification and matching of persons receiving health-related services has largely been used in established health programs such as immunization [8,9], and in other settings where unique identification is important such as a national census [10,11]. Though less common in settings such as HIV care and treatment programs, unique patient identification has increasingly become important as patient volume grows and the need for electronic medical records (EMRs) rises. Utility of UHID for correcting misclassification of the final patient outcomes such as loss to followup has been documented. For example, over a third of patients may be misclassified as having been lost to care in highly mobile populations, [12]. The benefits of EMRs over manual records have previously been demonstrated in our study setting [13]. The World Health Organization (WHO) recommends non-name matching for unique patient identification in person-centered patient monitoring and CBS, [14,15]. However, unique patient identification in HIV programs, especially in SSA, is complicated by lack of integration across the health sector, and the chronic nature of HIV requires integration of management of services, [16]. Even when a government identification document is issued at adulthood, use of its unique number for reproductive and healthcare services is limited by acceptance and excludes younger populations. In the absence of a unique persons identifier (UHID) that could be applied in the health sector, biometrics such as fingerprints are recommended [17], and may be used among HIV infected patients receiving care [18]. Other forms of patient identification, for example the comprehensive care clinic (CCC) medical record number assigned to all patients who initiate HIV care services, have limited potential. For example, CCC medical record numbers are unique to facility of origin, and combine a facility-based code with a unique serial patient number. Although there exists a master facility list (MFL) code system to identify each health facility, data management systems may not enforce use of MFL codes to link data across sites. Even within a facility, CCC medical record numbers may be unique to
Service area. National identity numbers have also been tried in HIV care settings but without much success because coverage is limited to patients 18 years and above, and they are subject to poor reception by patients, compromised recall when patients forget to carry the cards and transcription errors.

Without these identifiers, name and location-based deduplication of patient records and matching algorithms may be used where patient names and locator information exists [19]. Where patient names and other patient identifiers are documented, algorithms for records matching and deduplication may be used. There are two types of algorithms for deduplication: deterministic - a stepwise procedure in which sets of rules are used to pair up records identifying them as either matches or belonging to different persons, and probabilistic - in which statistical models are used to classify record pairs based on calculating a string distance measure to quantify how dissimilar two strings or words are to one another and applying a decision rule to determine whether two records belong to one individual.

As a chronic condition, HIV care entails use HIV services by patients at multiple locations over a lifetime. While UNAIDS recommends patient-centered co-location and integration of services across care settings such as antenatal care, TB and HIV, [4], co-location is not always feasible and hence tracking patients across the cascade of treatment can be difficult without a UHID and reliable EMR. Additionally, individuals may get an HIV diagnosis at one facility and choose to engage in HIV care at another location, they may receive a diagnosis in more than one care setting, and patients may move HIV care locations with or without notifying health care staff. Person matching using probabilistic methods has been shown to be superior over deterministic methods using simulated datasets [20], and may therefore be a feasible alternative for clinic data.

At facility level, waste of resources may result from duplicate records and uncoordinated HIV care, potentially leading to increased health service provider-patient contact time and multiple prescriptions to the same patient. At a higher regional or national epidemic surveillance e.g. in CBS, duplicate patient records may lead to inaccurate estimates and resource allocation. Hence, patient matching and deduplication is a priority for health care even when the methods of patient identification is not. However, there is lack of data on how probabilistic methods compare with deterministic matching using real HIV
program data in low-resource SSA countries. We used data from a pilot of HIV case-based surveillance in Siaya and Kisumu, two high HIV-burden counties in Western Kenya to: a) compare deterministic and probabilistic patient matching algorithms and b) propose an efficient algorithm for deduplicating and uniquely identifying HIV cases in CBS data collection and reporting in Kenya and similar settings.

Methods

Study setting

This HIV case-based surveillance pilot was conducted between July 2015 and December 2015 in 124 facilities in Kisumu and Siaya counties; facilities were selected to represent a variety of settings such as level, use of an EMR vs paper records, and size of patient population. Data were collected by sub-County AIDS and sexually transmitted infections (STI) Coordinators and Kenya Medical Research Institute (KEMRI) surveillance officers. Data were entered from paper medical records and registers into the customized data entry platform for cases newly diagnosed or newly-enrolled in HIV care in the months January through June 2015 using an Android-based tablets and a standardized HIV case report form. Surveillance officers were trained on data collection using tablets and provided with log-in credentials. All surveillance officers signed a data confidentiality statement. Data contained patient names and other patient identifiers and hence were encrypted before transmission via dedicated virtual private network (VPN) in real-time to an Amazon™ cloud computing service. Management of these data were done by staff at the national AIDS and STI control program (NASCOP), [21].

A case was defined as one with date of diagnosis, age at diagnosis, sex, first and last name. The following three settings relating to the HIV care cascade were used:

1) Within HIV Testing Services (HTS) setting; this included two scenarios, a. Cases found within the same facility - these are cases that were tested at the facility and retested at the same facility hence having different dates of diagnosis, and b. Cases that moved to a different facility - these are cases that tested at one facility and re-tested at a different facility.

2) HTS-to-care; we considered two scenarios, a. HTS-to-care within the same facility and b. HTS-to-care at a different facility. These two scenarios accounted for movement
194of persons diagnosed with HIV and accessing care at the same facility and clients that
195may test at one facility and access care in a different facility. We excluded the MFL code
196in the unique key to identify cases that could be matches in this scenario.
1973) Within care settings; including referrals and linkages from one facility to another.
198Similar to HTS-to-care linkage scenarios, there were cases who had enrolled into care
199in one facility and over the course of care transferred care to another facility. We
200standardized CCC numbers to take into account variations in recording, e.g. use of
201spaces, slashes, dashes, adding leading zeros, and commas. Since patients
202transferring care would need a referral letter with original CCC number, we checked for
203matches utilizing CCC numbers.

204Data preparation
205We used HIV diagnosis and settings relating to the cascade of HIV diagnosis described
206in the study setting for grouping – also called “blocking” - before carrying out matching
207analyses and to allow for comparability and faster processing.
208Before carrying out matching processes, we standardized patient identifying fields used
209in matching as follows:

- All blank spaces, commas, apostrophes, and dashes were stripped from
  individuals’ first and last names.
- All string fields were converted to lower case.
- A Soundex [22], was created which consisted of the first names in all records
  since the first names are mostly of English origin. The last name was left “as-
  captured” in the data system and concatenated with the previously-created
  Soundex to create a pseudo name.

An additional variable considered for the de-duplication process was CCC number
which is a unique patient number (UPN) assigned at first clinical encounter once an
HIV-infected patient has gone through triage and is ready for enrolment into a facility-
managed HIV program. The CCC number is an 11 character code comprising a 5 digit
unique facility code followed by a separator and a 5 digit sequentially facility-assigned
unique number. Age at time of diagnosis was re-calculated where date of birth was
present and if missing, it was retained as documented in the source documents.

Methods for data collection are described in the HIV case based surveillance pilot report
[23]. Briefly, it was everyone who was newly diagnosed/enrolled in care in a given 6
month’s period in the participating facilities and subsequent updating of sentinel events for those individuals. We excluded a hundred records which had missing date of diagnosis and three which had missing date of birth prior to matching (Figure 1).

**Deterministic matching**

Out of the cases that were abstracted for this pilot, we used the fields: first name, last name, sex, date of birth, unique master facility list code (MFL) – a permanent five digit code that each health facility in Kenya is assigned - and CCC number. To uniformly capture correct spelling of English first names, we used Soundex of the first name. We then created a “unique key” combining the resulting Soundex values as well as sex and year of birth. We carried out extra steps to determine how to retain a record based on the date of HIV diagnosis as follows: If the retained and duplicate records had conflicting date of diagnosis, we retained the records with the earliest date of diagnosis. For retained records, we maximized completeness of data for all fields by comparing with the duplicate records. Whenever a retained record had missing data that was in duplicate record(s), an append merge was carried out to overwrite missing values with the non-missing value from the matched record.

**Probabilistic matching**

We separated the data according to the “blocking” scenarios described in the deterministic process. We then implemented Jaro-Winkler, Levenshtein and Damerau-Levenshtein string edit distance calculations in R stringdist package [24,25]. String edit distance calculations quantify how dissimilar two strings or words are to one another by counting the minimum number of deletion, insertion, substitution and transposition operations required to transform one string into the other. These probabilistic methodology are based on Fellegi-Sunter linkage rule that classify a record pair as matching or non-matching [11].

When we implemented the Jaro and Jaro-Winkler methods we set the penalty factor – that penalizes matches based on similarity at the beginning of the string as $p=0.1$ according to Winkler and Cohen [11,26]. We considered the four weights ($\omega$) applicable to Levenshtein, Damerau-Levenshtein and optimum string alignment (OSA) - an extended version of Damerau-Levenshtein method) methods as follows: deletion $\omega=0.8$;
Uniquely identifying cases in HIV case-based surveillance

Post processing for matches was performed in Stata version 14.2 (Stata Corporation, Texas USA). Given field experiences on possibilities of age variations, the numeric comparator age, with difference of +/- 2 years was considered sufficiently close to confirm a match. We compared deterministic and probabilistic processes for unique case identification in terms of number of matches yielded and the deduplication extent achieved within the scenarios.

Ethical considerations
Ethical approval was obtained from the Kenya Medical Research Institute and the Office of the Associate Director for Science, Centers for Disease Control and Prevention. Access to data used in these analyses were password protected and all study coordinators, data abstractors and analysts signed a confidentiality form.

Results
Matches identified
Using deterministic methods, of the 12,157 records, 67 (1.64%) were matches in HTS alone, 164 (15.03%) in HTS-to-care and 204 (2.92%) in care setting. This yielded a total of 435 (3.58%) matches and 11,722 unique cases across the testing, care and treatment cascade (Table 1).

Table 1. Scenarios in HIV diagnosis, care and treatment cascade and deduplication yield, HIV case-based surveillance, Kenya 2015.

<table>
<thead>
<tr>
<th>Scenarios, N a</th>
<th>Deterministic</th>
<th>Probabilistic</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Jaro</td>
<td>Jaro-Winkler</td>
</tr>
<tr>
<td>All c</td>
<td>12,157</td>
<td>435(3.58)</td>
</tr>
<tr>
<td>A</td>
<td>4,073</td>
<td>67(1.64)</td>
</tr>
<tr>
<td>B</td>
<td>1,091</td>
<td>164(15.03)</td>
</tr>
<tr>
<td>C</td>
<td>6,993</td>
<td>204(2.92)</td>
</tr>
</tbody>
</table>

Unique cases
11,722 | 11,479 | 11,539 | 11,386 | 11,376

a A = HTS alone; B = HTS to care; C = Within care; D = Combined
b The optimum string alignment option yielded similar number of matches - 779(6.4%)
c Summed up for all the scenarios.
Overall, of the probabilistic methods, *Jaro-Winkler* yielded the least duplicate records, 282618 (5.08%) while *Jaro, Levenshtein* and *Damerau-Levenshtein* incrementally yielded 283678(5.58%), 771(6.34%) and 781(6.42%) duplicate records respectively. Specifically, duplicate records yielded by method were: *Jaro* 266(6.53%) within HTS, 59(5.41%) in HTS-care, and 353(5.05%) within care; *Jaro-Winkler* 232 (5.70%) within HTS, 50 286(4.58%) within HTS-care and 336 (4.80%) within care; *Levenshtein* 314 (7.71%) within 287HTS, 72 (6.60%) in HTS-care and 385 (5.51%) within care; and *Damerau-Levenshtein* 288303(7.44%) within HTS, 77(7.06%) in HTS-care and 401 (5.73%) within care.

**HIV case records and matching fields**

A total of 12,260 records were collected (Figure 1). The final data set used for the matching exercise included 12,157 records representing adult and pediatric cases. Out of these records and prior to data de-duplication, 4,073 (33.50%), 1,091 (8.97%) and 6,993 (57.52%) corresponded to HTS, HTS-to-care and within care settings respectively. In HTS setting, 247 (6.06%) surnames had exactly the same spelling. First names were less unique than surnames, with 407 (9.99%) having the same spelling. When Soundex was applied to the English first names, 321 (7.88%) resulted in an identical Soundex string. In the entire dataset, the four match fields (first name, last name, sex and age) were 100% complete while CCC number was 60% complete. Only 0.13% of adult records had the national identity number and 44.6% had a mobile number. Similarity of names varied by setting (Table 2).

### Table 2. Percent complete or identical records within matching fields, HIV case-based surveillance, Kenya 2015.

<table>
<thead>
<tr>
<th>Fields</th>
<th>HTS n=4,073</th>
<th>HTS-care, n=1,091</th>
<th>Care, n=6,993</th>
<th>All, n=12,157</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Complete</td>
<td>Identical</td>
<td>Complete</td>
<td>Identical</td>
</tr>
<tr>
<td>First name</td>
<td>100.0</td>
<td>407(9.99)</td>
<td>100.0</td>
<td>195(14.57)</td>
</tr>
<tr>
<td>Soundex of first name</td>
<td>100.0</td>
<td>321(7.88)</td>
<td>100.0</td>
<td>143(13.11)</td>
</tr>
<tr>
<td>Last name</td>
<td>100.0</td>
<td>247(6.06)</td>
<td>100.0</td>
<td>85(7.79)</td>
</tr>
<tr>
<td>Soundex of last name</td>
<td>100.0</td>
<td>245(6.02)</td>
<td>100.0</td>
<td>111(10.17)</td>
</tr>
<tr>
<td>Sex</td>
<td>100.0</td>
<td>-</td>
<td>100.0</td>
<td>-</td>
</tr>
<tr>
<td>CCC number</td>
<td>100.0</td>
<td>-</td>
<td>95.0</td>
<td>8(0.73)</td>
</tr>
</tbody>
</table>
Comparison of deterministic and probabilistic matching

A comparison of deterministic and the most conservative probabilistic matching (Jaro-Winkler) yield is presented in Figure 2. In the HTS scenario, probabilistic methods yielded 5.7% matches compared to 1.6% using deterministic matching. In HTS-to-care scenario, deterministic matching yielded more matches (15.0%) than probabilistic (4.6%). Within care, deterministic matching yielded (2.9%) and probabilistic matching (4.8%). Overall, probabilistic matching yielded 42% more matches than deterministic process \[\frac{(5.1-3.6)}{3.6} = 42\%\].

Discussion

Main findings

In this study, we have demonstrated that probabilistic methods succeeded in patients matching and identifying more matches compared to deterministic process. Our analyses demonstrate that it is possible to match cases and combine into single case report. This would greatly improve accuracy in target setting, estimation of denominators and calculation of coverage gaps in the HIV care cascade and tracking the UNAIDS Fast-track 90-90-90 targets. The utility of the process also serves the dual purpose of better care coordination at the facility level and improved HIV surveillance at a higher sub-national or national level. We also demonstrate that we can do enough matching in the absence of a UHID to move ahead with CBS implementation in low-resource settings in Kenya. As such, lack of a UHID should not stifle movement towards use of CBS for surveillance.

Caveats

The choice of string distance probabilistic algorithm to use largely depends on the nature of the strings to compare and the nature of typographic errors [24]. Choice of the matching pseudo ID is therefore important. For example, deterministic matching yielded more duplicates for the HTS to care scenario (15%) compared to 4.6 to 7.1% across the probabilistic methods. This may be due to the fact that rigorous manual assessment of possible matches was done using the CCC numbers such that within care transfers for matches within the HTS to care scenario were more efficiently captured. Minimalistic
demographic fields were used in probabilistic matching across all scenarios and the CCC number was not included in the process.

Probabilistic and deterministic matching

Our study compared four variants of probabilistic string-distance matching methods. Although three of these yielded more matches, the Jaro-Winkler distance method was found to be best in probabilistic matching in terms of speed and would be a better choice since it is more conservative. In developed countries, it has been shown that about 5 to 10% of medical records may be duplicate [27]. Although there is no such estimate in sub-Saharan Africa, our study showed that in this CBS pilot, the Jaro-Winkler method yielded a conservative 5.1% duplicate records. Jaro-Winkler has been proposed as a method over other string matching algorithms since it was designed with relatively short strings in mind [24], hence may be best suited to our setting. In addition, it works well when the name beginnings are the same, [28]. For that reason, we standardized beginning of the match strings by using a Soundex of the English names.

Application considerations

We used R - an open-source software - in our study since it provided programming flexibility. However, we also used Stata to perform some post-match management tasks. These however can be incorporated in an R workflow. Exploratory use of available open source software such as CDC's Registry Plus™ Linkplus [29], which was originally developed for cancer registries has been done in low-resource HIV care settings for example in Haiti [30]. Other web-based applications that have utility for fuzzy matching and record cleaning do exist, for example Freely Extensible Biomedical Record Linkage (Febrl), [6] may have potential. However, post-match processing is necessary to achieve a high degree of true matches. Users of off-the-shelf solutions such as Linkplus should take caution since many mismatches may be likely to be true, [30]. The use of current English-names based Soundex algorithms is not appropriate for Kenyan names. In creating unique identifiers that contain a Soundex component, variations of the first name can yield a different Soundex since the first character is always part of the Soundex, [22]. A visual inspection of matches based solely on Soundex of first and last name showed a high false-positive rate. Hence, research on how to construct a Soundex algorithm for Kenyan names will be useful as has been successfully done in...
Japan, India and South Africa, [31–33]. A potential identifier for patients in care is the CCC number since more than 95% of cases in care had CCC numbers. However, although the procedure for creating a CCC number is clearly laid out, and the number is meant to be permanent for patients enrolled in care, the allocation of this number in practice presents challenges. The most common challenge is the various ways in which the number is recorded.

Limitations

Our study has several limitations. Firstly, the choice of unique key as a combination of several fields may not be optimal. However, we developed the matching string taking advantage of existing identifiers in our data. The first names in Kenya are usually English baptismal or anglicized. We took advantage of this to standardize names that are misspelled using Soundex. Other challenges include: manual transcription errors during patient transfers and assigning of new numbers for transfer-in patients. In spite of these limitations, we were able to merge the cases based on the names, sex, date of birth and CCC number in the within care scenario and hence identify potential matches in the deterministic process. Finally, many studies have applied common measures of validity such as positive predictive value, sensitivity and specificity, [34]. Unlike those studies, we did not have a gold-standard to compare with.

Conclusions and recommendations

There has been an ongoing discussion and developing of steps and approach to developing a universal health ID for countries, [35]. In Kenya, the process of developing a universal healthcare identifier assigned at birth and collected as part of receiving health services has been going on for a while. If and when implemented, a UHID will have the highest potential to mitigate challenges with unique identification and record linkage for an expanded national CBS system. Other opportunities for matching using national identity card number and mobile phone number may have limited application where completeness and accuracy are high. In well-defined populations such as in a health and demographic surveillance systems (HDSS), the assigned HDSS number has potential if collected by the health facility but its utility would only be limited to a HDSS region. Names should also clearly be designated as either first, middle or last. Where a person has only two names, the second name should be designated as either last,
family or surname. In the interim, a unique patients’ deduplication algorithm based on available identifiers is necessary. We propose a 10-step matching algorithm that is easy to apply to quickly match and link unique cases across HIV care settings, (Figure 3). The proposed algorithm would improve timeliness and accurate monitoring of the HIV epidemic in Kenya including the Fast-Track 90-90-90 targets.

The recent guidelines for patient-centered monitoring advocate for use of unique patient identifiers instead of names [14]. For HIV case-based surveillance in Kenya, we propose a process that builds up step-wise from identifying data sources, blocking scenarios, examination of data quality using completeness as a measure coupled with quality improvement measures through routine data quality audits, developing of a matching key, deduplication at a lower level and finally cross-examination, validation and sending of CBS data to national level for surveillance. Although validation of probabilistic approach is a necessary extra step, this may be best done in settings where “gold standard” data sets such as those utilizing biometric finger vein technology for patient identification. Given that these settings are rare, we suggest that programs identify a percentage that best suits their setting and resources for validation purposes.

Comparing probabilistic matches to “gold standard” data in Kenya and similar settings offers an opportunity for future work in search of alternatives for patient matching. In the meantime, probabilistic demographic data matching has utility to improve quality of data in monitoring the 90-90-90 cascade and in other healthcare settings.

Acknowledgements
We acknowledge the Kisumu and Siaya county and sub-county HIV/AIDS and STI coordinators (S/CASCOs) and the Kenya Medical Research Institute (KEMRI) surveillance officers who helped in data collection.

Authors contributions
AW conceived the idea for this manuscript and prepared the concept, data analyses and wrote the first and subsequent drafts of the manuscript. AN, also helped with data analyses. PY, AN, TK, WW, KM, and EZ provided extensive comments on the concept and manuscript drafts. LN, JO, PY, and EM provided insights on policy implications and
recommendations. RH, and KM supervised the data abstraction process. All authors read the manuscript, provided feedback, and approved the final version.

References

Centers for Disease Control and Prevention. Guidelines for national human immunodeficiency virus case surveillance, including monitoring for human immunodeficiency virus infection and acquired immunodeficiency syndrome. vol. 48. 1999.


Otieno J. HIV patients to be recorded biometrically. Star 2014.


Loo M van der, Laan J van der, Logan N. Package “stringdist” 2016.


