POPULATION HEALTH METRICS

RESEARCH Open Access

Evaluation of record linkage of mortality data between a health and demographic surveillance system and national civil registration system in South Africa

Chodziwadziwa W Kabudula[1*], Jané D Joubert[2,3], Maletela Tuoane-Nkhasi[4], Kathleen Kahn[1,5,6], Chalapati Rao[3], Francesc Xavier Gómez-Olivé[1], Paul Mee[1,5], Stephen Tollman[1,5,6], Alan D Lopez[7], Theo Vos[8] and Debbie Bradshaw[2]

Abstract

Background: Health and Demographic Surveillance Systems (HDSS) collect independent mortality data that could be used for assessing the quality of mortality data in national civil registration (CR) systems in low- and middle-income countries. However, the use of HDSS data for such purposes depends on the quality of record linkage between the two data sources. We describe and evaluate the quality of record linkage between HDSS and CR mortality data in South Africa with HDSS data from Agincourt HDSS.

Methods: We applied deterministic and probabilistic record linkage approaches to mortality records from 2006 to 2009 from the Agincourt HDSS and those in the CR system. Quality of the matches generated by the probabilistic approach was evaluated using sensitivity and positive predictive value (PPV) calculated from a subset of records that were linked using national identity number. Matched and unmatched records from the Agincourt HDSS were compared to identify characteristics associated with successful matching. In addition, the distribution of background characteristics in all deaths that occurred in 2009 and those linked to CR records was compared to assess systematic bias in the resulting record-linked dataset in the latest time period.

Results: Deterministic and probabilistic record linkage approaches combined linked a total of 2264 out of 3726 (60.8%) mortality records from the Agincourt HDSS to those in the CR system. Probabilistic approaches independently linked 1969 (87.0%) of the linked records. In a subset of 708 records that were linked using national identity number, the probabilistic approaches yielded sensitivity of 90.0% and PPV of 98.5%. Records belonging to more vulnerable people, including poorer persons, young children, and non-South Africans were less likely to be matched. Nevertheless, distribution of most background characteristics was similar between all Agincourt HDSS deaths and those matched to CR records in the latest time period.

Conclusion: This study shows that record linkage of mortality data from HDSS and CR systems is possible and can be useful in South Africa. The study identifies predictors for death registration and data items and registration system characteristics that could be improved to achieve more optimal future matching possibilities.

Keywords: Health and demographic surveillance system (HDSS), Agincourt HDSS, Record linkage, Civil registration system, Death registration, South Africa, Mortality

* Correspondence: chodziwadziwa.kabudula@wits.ac.za
[1]MRC/Wits Rural Public Health and Health Transitions Research Unit (Agincourt), School of Public Health, Faculty of Health Sciences, University of the Witwatersrand, Johannesburg, South Africa
Full list of author information is available at the end of the article

Background

Reliable and valid statistics on the levels and causes of mortality are widely acknowledged as essential information for monitoring the impact of health interventions and developing public health policies and programs for improving population health [1-4]. An adequate and complete civil registration (CR) system is the ideal source from which to draw such information [2,5].

Well-functioning CR systems do not exist in the majority of African countries [6]. South Africa is one of the few that produce mortality statistics from a CR system [7,8], but previous assessments rated their quality as low [2,9]. In recent years, the country has adopted the Africa Programme on Accelerated Improvement of Civil Registration and Vital Statistics (APAI-CRVS) [6], building on the focused initiatives by Statistics South Africa, the Department of Health, and a group of researchers since the 1990s to improve and strengthen its CR system and cause of death information [10-12]. Therefore, there is a continuous need for assessing the quality of CR mortality data to ascertain the impact of these initiatives and identify remaining gaps and options for further improvement.

A number of criteria, organized into a framework of four quality concepts (generalizability, reliability, validity, and policy relevance), have been proposed for comprehensive assessment of the quality of mortality data in CR systems [13,14]. Although most criteria can be evaluated directly from the mortality data recorded in the CR system and administrative information on the system, data from other sources are also required [13]. Combining vital-event data sources, and cooperation among the custodians of such data sources, was encouraged at the 2012 International Network for the Demographic Evaluation of Populations and Their Health in developing countries (INDEPTH) - African Census Analysis Project (ACAP) Bellagio meeting on using longitudinal INDEPTH data, national censuses, Demographic and Health Surveys, and other national surveys for better health policy in Africa [15].

In South Africa, three INDEPTH Health and Demographic Surveillance Systems (HDSS) collect mortality data in rural populations [16-18]. Such data provide an opportunity for comparison with CR data. However, this requires record linkage between the two data sources, which has not been attempted before. Both data sources are protected by strict data-use clauses to protect the confidentiality of the identity and other information of the deceased. Once linked, comparison would also depend on the quality of the matched records.

This paper describes the practical steps we took to set up and execute record linkage of mortality data and evaluates the quality of the matched records between the CR system and the longest-running of the three HDSS centers in South Africa, the Agincourt HDSS [19,20]. It describes how we overcame the challenges of bringing together data that are kept in secure databases and environments almost 600 kilometers apart, each governed by data-security policies that prohibit the off-site and non-staff use of unit-record data that contain personal identifiers.

Methods

Data sources

Records of individuals who died from 1 January 2006 to 31 December 2009 were extracted from the Agincourt HDSS database and saved under password protection on a portable device. An Agincourt HDSS staff member who is familiar with the collection, processing, and coding of mortality data and the stringent data-use policies at Agincourt, and who had previous experience in electronic record linkage, securely brought the data files to Statistics South Africa's (Stats SA) head office in Pretoria. After confidentiality and data-security agreements were undertaken and signed by the Agincourt HDSS staff member and other members of the record linkage team, the non-Stats SA team members were given access to the secure environment in the Stats SA building where CR data for deaths that occurred within the same period were made available for linkage.

Information about deaths in the Agincourt HDSS was collected as part of annual updates of vital events in a surveillance population occupying 27 villages in Bushbuckridge municipality, Mpumalanga province, South Africa (Figure 1) [19,20].The population is largely Tsonga-speaking, and a third are of Mozambican descent who arrived in the study area in the early to mid-1980s as refugees and/or their descendants. The population has been under surveillance since 1992. Residency status and vital events have been updated at approximately 15- to 18-month intervals between 1993 and 1999, and annually since 1999. During the annual update, an individual who was present at home for at least six months in the last 12 months is considered a permanent resident. Permanent residency status is also assigned to an individual who in-migrated or a child who was born prior to the annual update and is considered as a permanent resident by the household informant. Individuals who were present at home for less than six months in the previous 12 months due to work-related, education-related, and other reasons are assigned a temporary residency status. For each death recorded during the annual update, a verbal autopsy (VA) interview is conducted with caregivers of deceased individuals one to 11 months after death to elicit signs and symptoms of the illness or injury prior to death using a locally validated, local-language VA instrument [20,21]. Because vital events are updated every year, death events missed in one year are captured the following year since the deceased individuals still appear on pre-populated household rosters. Hence, completeness of

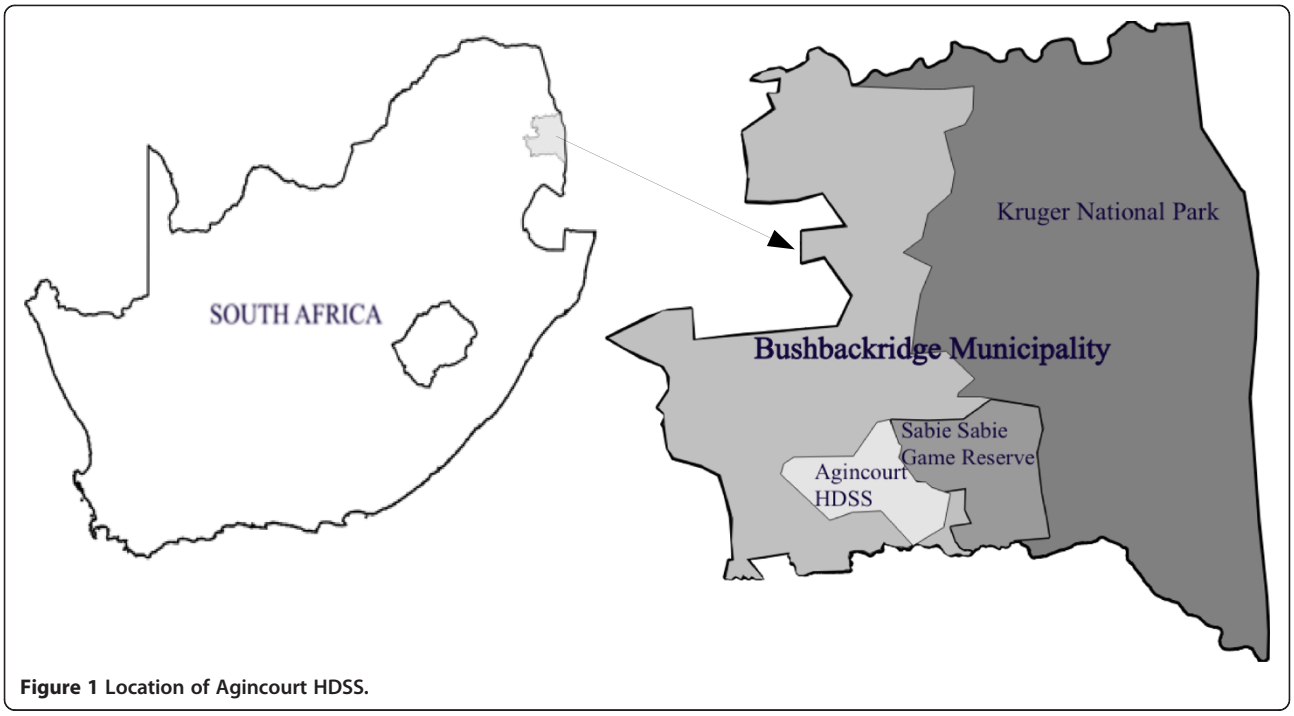

Figure 1 Location of Agincourt HDSS.

recording of deaths into the Agincourt HDSS is very high despite some under-recording of neonatal deaths. In addition, the data items collected pertaining to the characteristics of the event of death and the deceased person cover most of the core topics of themes for vital statistics as recommended by the United Nations [22]. These include date of death, date of death registration when a death certificate is available, place of death, cause of death derived from verbal responses to autopsy interviews, date of birth of the deceased, place of usual residence of the deceased, and marital status of the deceased.

The CR data were captured by Stats SA from *Notification of death/still-birth* forms (Form BI-1663) that were submitted to the Department of Home Affairs offices for death registration as required by the country's Births and Deaths Registration Act No 51 of 1992 [23]. As required by the Act, different sections of the form are completed by (i) the person reporting the death, (ii) a medical practitioner (where a medical practitioner is not available, a traditional leader may complete the Death Report (Form BI-1680)), and (iii) a Home Affairs official or member of the South African Police Services if the former is not available [24,25].

Record linkage procedures

We applied deterministic and probabilistic record linkage approaches to link the Agincourt HDSS and CR mortality data. Variables common to both data sources that we used are: national identity number (a unique 13-digit number assigned to South African citizens), surname, sex, day of death, month of death, year of death, day of birth, month of birth, year of birth, institution/place of death, and village name. For village name, village of the household of the deceased individual in the Agincourt HDSS was matched to place of birth, residency, and death in the CR records. Due to the recording of local tribal area names rather than the official village names for some deaths on the CR death registration forms, the place names in the CR records were mapped to their equivalent Agincourt HDSS village names prior to the record linkage exercise.

In deterministic record linkage, a pair of records from two data sources is considered to belong to the same individual if it matches on a unique identifier such as national identification number or a set of conventional personal identifiers (e.g., the combination of first name, last name, and date of birth) [26-29]. We defined 12 deterministic linkage rules based on different combinations of the common variables as presented in Table 1. Record linkage using these rules proceeded iteratively. Records matched by one rule were removed from the pool of records to be matched with subsequent rules in both datasets. The Jaro-Winkler string comparator (JW) [30], which is particularly well-suited for personal names [31], returning values between 0 (complete disagreement) and 1 (exact agreement) as a measure of similarity between two strings [30,32], was used to accommodate typographical errors on surnames. We set a cut-off for designating pairs of surnames as matches to a JW score ≥ 0.85, which is higher than in previous studies [30,33].

Table 1 Deterministic matches

Rule number	Description	Matches in trimmed CR dataset	Matches in full CR dataset	Total matches
1	Match on National ID No	708	161	869
2	Match on Surname, Sex, Date of birth, Date of death	128	28	156
3	Match on Surname, Sex, Date of birth, Year of death, Month of death	88	30	118
4	Match on Surname, Sex, Year of birth, Month of birth, Date of death	34	6	40
5	Match on JW(Surname) > =0.85, Sex, Date of birth, Date of death	39	-	39
6	Match on JW(Surname) > =0.85, Sex, Date of birth, Year of death, Month of death	37	-	37
7	Match on JW(Surname) > =0.85, Sex, Year of birth, Month of birth, Date of death	6	-	6
8	Match on JW(Surname) > =0.85, Sex, Year of birth, Year of death, Agincourt HDSS village = CR place of birth	207	-	207
9	Match on JW(Surname) > =0.85, Sex, Year of birth, Year of death, Agincourt HDSS village = CR place of residence	67	-	67
10	Match on JW(Surname) > =0.85, Sex, Year of birth, Year of death, Agincourt HDSS village = CR place of death	23	-	23
11	Match on JW(Surname) > =0.85, Sex, Year of birth, Date of death, died at hospital	30	-	30
12	Match on JW(Surname) > =0.85, Sex, Date of birth, Year of death, died at hospital	27	-	27
Total		1,394	225	1,619

In probabilistic record linkage, a pair of records from two data sources is classified as a match based on the statistical probability that the values of common variables from the two data sources belong to the same individual [32,34-38]. Each matching variable is assigned a weight that indicates its contribution to the probability of accurately designating a pair of records as a match or non-match [29,32,36]. The weight of a matching variable, i, is calculated from the probability that records belonging to the same individual agree, denoted by m_i, and the probability that records belonging to different individuals agree, denoted by u_i [32,36,38]. Record pairs where variable i agrees receive a weight value of $\log_2 \frac{m_i}{u_i}$, and those where the variable disagrees get a weight value of $\log_2 \frac{1-m_i}{1-u_i}$. A record pair is classified as a match if the sum of the weights on all the matching variables is above a particular threshold value. We estimated m_i and u_i values for all matching variables, except national identity number, using the Expectation Maximization (EM) algorithm [32,39,40]. Only surname pairs with a JW score ≥ 0.85 were considered as matches. Similar to the work of Méray et al. [41] and Tromp et al. [42], the threshold value for determining which record pairs were matches was derived from an estimate of the proportion of true matches among all possible record pair combinations produced by the EM algorithm. The estimated proportion of true matches was multiplied by the total number of all possible record pair combinations to obtain the total number of true matches. Thereafter, all possible record pair combinations were sorted in descending order of the sum of the weights on all matching variables and the top n record pairs, where n equals the calculated number of true matches, were designated as matches.

The number of possible record pair comparisons in two files to be linked is equal to the product of the numbers of records on each file, which can be enormous. Therefore, we used blocking to reduce the number of record pair comparisons [32]. We restricted the comparisons to "blocks" or "pockets" of record pairs with exact matches on one or more variables. We applied the 12 deterministic rules to link the Agincourt HDSS dataset with a trimmed CR dataset that had records for which the deceased was either born, resident, or died within the Bushbuckridge municipality. We repeated the linkage between the Agincourt HDSS dataset with the trimmed CR dataset using the probabilistic approach described above with further blocking on (i) sex and year of birth, and (ii) sex and year of death. Finally, we applied the first four strict deterministic rules in Table 1 to link the thus-far unmatched records in the Agincourt HDSS dataset with records in the full CR dataset (compare Figure 2).

Evaluation of record linkage results

Since we set strict deterministic matching rules with very narrow margins for error, evaluation of the record linkage results focused on matches generated by the

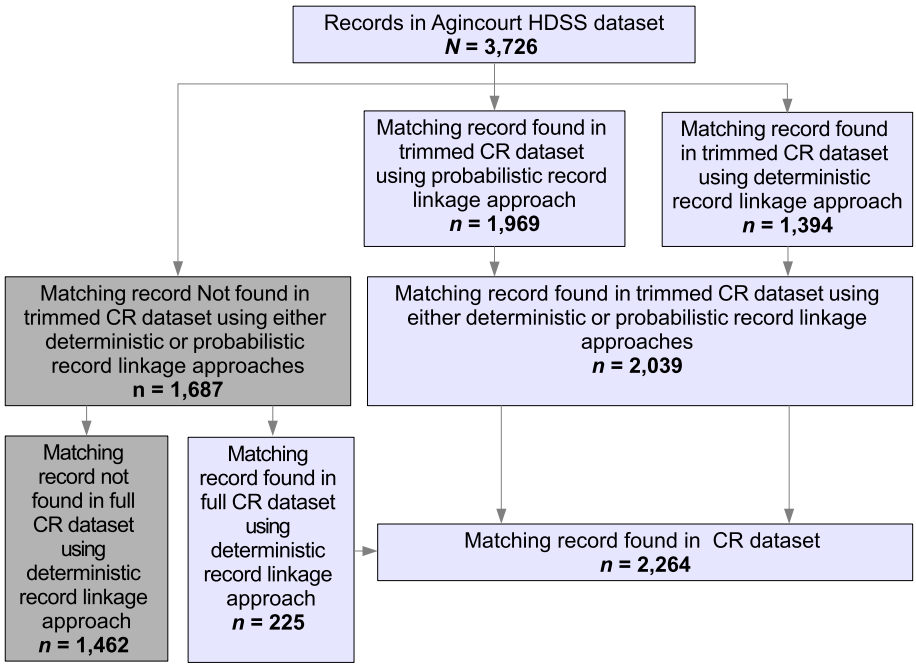

Figure 2 Agincourt HDSS records matched with CR records.

probabilistic record linkage approach. Their quality was evaluated using sensitivity and PPV calculated from a subset of records that were linked by means of national identity number. This is justifiable because national identity numbers contain a check digit that prevents incorrect matching. We also compared characteristics of the deceased individuals in the Agincourt HDSS dataset whose records were matched and unmatched to records in the CR dataset in logistic regression models to identify characteristics associated with successful matching. Variables selected for analysis included sex, age, nationality, having a national identity number, residency status, level of education, wealth quintile, year of death, and place of death. Wealth quintiles were derived from data on ownership of assets such as cattle, a car, and cell phone and access to amenities including drinking water and sanitation using principal component analysis [43]. In addition, the distribution of background characteristics in all deaths that occurred in 2009 and those linked to CR records was compared using Pearson Chi squared tests to assess systematic bias in the resulting record-linked dataset in the latest time period.

Software

The record linkage of the data between the two data sources was done using Microsoft SQL Server 2008 which had the EM algorithm implemented in Microsoft C# progamming language, integrated in it as a common language runtime (CLR) function. The JW algorithm we used is part of the SimMetrics library [44]. It was also integrated in Microsoft SQL Server 2008 as a CLR function. Stata (version 11.2, Stata Corporation, Texas, USA) was used for data analysis.

Ethics

The study received ethical approvals from the University of Queensland School of Population Health Research Ethics Committee (approval no. JJ010911), the South African Medical Research Council Ethics Committee (EC008-6/2011), and the University of the Witwatersrand Human Research Ethics Committee (Medical) (M120106).

Results

The Agincourt HDSS and full CR datasets, respectively, contained 3,726 and 2,464,915 unit records of individuals who died from 2006 to 2009. Place of birth, residence, or death in 29,416 of the records in the CR dataset was within the Bushbuckridge municipality. Overall, 2,264 (60.8%) of the 3,726 records from Agincourt HDSS were matched to records in the CR dataset (Figure 2). The matched record pairs comprised of 2,039 record pairs that were obtained by linking Agincourt HDSS records to records in the trimmed CR dataset, and 225 record pairs obtained by linking the remaining Agincourt HDSS records with records in the full CR dataset. Deterministic and probabilistic record linkage approaches, respectively and independently, produced 1,394 and 1,969 of the record pairs that were obtained by linking Agincourt HDSS records to records in the trimmed CR dataset, and 1,324 (95.0%) of the records that were matched using the

deterministic approach were also matched using the probabilistic approach. The weights computed for probabilistic record linkage for each of the variables in the two blocking schemes are presented in Tables 2 and 3. The weights indicate that village had the highest agreement weight followed by surname, and for the block on sex and year of death, record pairs that agreed on all the variables were assigned an overall weight score of 32.91344, whereas those that disagreed on all the variables were assigned an overall weight score of −14.02270.

Most of the record pairs that were generated by linking the remaining Agincourt HDSS records with records in the full CR dataset had Hazyview, a town about 40 km away from the Agincourt HDSS, as the reported place of birth, residence, or death in the CR dataset. There were also a few cases for which the reported place of birth, residence, or death in the CR dataset is indeed within the Agincourt HDSS study site, such as Belfast and Somerset, but had not been assigned to the Bushbuckridge municipality in the CR system. For example, one of the death records from Somerset village in the Agincourt HDSS dataset was in the CR dataset assigned to Somerset West, a town in the Western Cape province. Over half (53.7%) of the combined deterministic matches were found via the deceased's identity number (Table 1).

In a subset of 708 records from the Agincourt HDSS that were deterministically linked by means of national identification number, the probabilistic approaches yielded sensitivity of 90.0% and a positive predictive value of 98.5%.

Table 4 presents a number of characteristics of the deceased in the Agincourt HDSS dataset and their association with successful matching to individuals in the CR dataset. In a univariate model, higher matching likelihood was associated with age (age > =5 years), nationality (South African), having a national identity number recorded in the VA system, residency status prior to death (permanently residing in the Agincourt HDSS study area), education (at least some primary education, but higher likelihood with secondary), wealth (from second to highest wealth quintile), year of death (more recent year), and place of death (place of death specified). In a multivariable model, matched and non-matched cases differed significantly in terms of age, nationality, having a national identity number recorded, residency status, wealth (from middle to highest wealth quintile), and place of death (place of death specified). Having adjusted for all variables in the model, having a national identity number recorded increased the odds to be matched by almost 14 times.

Although a number of characteristics prevented successful record linkage of mortality between the Agincourt HDSS and CR system for the period considered in this study, in the latest time period (2009), except for infants and non-South Africans, there were no significant differences in the distribution of background characteristics in all Agincourt HDSS deaths compared to those matched with CR records (Table 5).

Discussion

In South Africa, there are no comprehensive systems of pre-linked health data covering large or entire populations such as the Manitoba Population Health Information System in Canada [45] or systems that routinely or periodically link data at any level of jurisdiction. In this study, we have assessed the feasibility of setting up and executing record linkage of mortality data and evaluated the quality of the matched records between the Agincourt HDSS and the CR system. The study was motivated by the unexplored potential of HDSS as sources of independent mortality data for assessing the quality of mortality data in CR systems in low-and middle-income countries.

Using deterministic and probabilistic approaches, our study yielded a matching rate of 60.8% for mortality records from 2006 to 2009, with sensitivity of 90% and PPV of 98.5% for the probabilistic linkage. This matching rate was influenced by a number of limitations relating to the amount, accuracy, completeness, and consistency of information available for the linkage process [46]. First, we had a small number of common variables in the two datasets. Second, collection of the ideal unique-identifier variable, national identity number, was introduced gradually in the Agincourt HDSS over the period of our investigation,

Table 2 Weights for the probabilistic linkage approach with blocking on sex and year of death

Variable	m_i	u_i	Agreement weight	Disagreement weight
Surname	0.80987	0.01581	5.67838	−2.37191
Day of birth	0.62776	0.03498	4.16563	−1.37431
Month of birth	0.68986	0.08655	2.99473	−1.55840
Year of birth	0.80879	0.01643	5.62171	−2.36291
Month of death	0.80330	0.08366	3.26334	−2.21987
Day of death	0.53200	0.03253	4.03141	−1.04770
Village	0.60572	0.00832	6.18540	−1.33066
Institution/place of death	0.82912	0.42244	0.97283	−1.75695

Table 3 Weights for the probabilistic linkage approach with blocking on sex and year of birth

Variable	m_i	u_i	Agreement weight	Disagreement weight
Surname	0.81578	0.01588	5.68252	−2.41737
Day of birth	0.69855	0.03605	4.27643	−1.67704
Month of birth	0.75919	0.08734	3.11968	−1.92216
Year of death	0.98414	0.28645	1.78059	−5.49130
Month of death	0.80748	0.08367	3.27068	−2.25083
Day of death	0.52955	0.03286	4.01026	−1.03967
Village	0.60579	0.00808	6.22918	−1.33126
Institution/place of death	0.82785	0.43602	0.92498	−1.71199

starting only in 2007. However, it is worth noting that as of 2013, national identity number was available on 68% of the individuals still under surveillance in the Agincourt HDSS. Therefore, national identity number has an increased future potential as a unique matching variable. Third, we set strict deterministic matching rules with narrow margins for error, such as in the case of the spellings of surnames. Fourth, there has been a particular problem with the reporting of tribal area names instead of village names for some deaths in the death registration system. As more than one village is contained in a tribal area, it is not possible to correct this data entry. Last, the use of proxy respondents, inevitably, in both VA and CR systems, and that VA interviews are conducted one to 11 months after death, may also have reduced the accuracy of individual-level information.

While the record linkage approach employed in this study would typically allow the assessment of completeness using a standard two-source capture-recapture analysis [47,48], it is not possible in our study. This stems from difficulties in identifying CR deaths that occurred within the Agincourt HDSS borders due to the recording of local tribal area names rather than the official village names on the CR death registration forms for some deaths. The three tribal areas containing the study site additionally include areas not covered by the Agincourt HDSS. Furthermore, the places of birth, death and residence in the CR data, reported by the relative or friend of the deceased, were not verified against the StatsSA official or Agincourt HDSS colloquial place names. Valuable lessons were learned in this regard, and recommendations are offered in the Conclusion.

Even though the matching rate in this study is low and it is not possible to assess completeness of death registration using a standard two-source capture-recapture analysis due to the limitations above, the similarity in the distribution of most of the background characteristics in all Agincourt HDSS deaths compared to those matched with CR records in the latest time period (2009) suggests that the record-linked data can enhance understanding of death registration practices into the CR system through identifying subgroups likely to be underrepresented in the CR data. For example, the finding that after adjusting for other variables, matching rates are significantly lower for records belonging to more vulnerable people, including poorer persons, children <5 years, and non-South Africans could possibly be interpreted to mean that their deaths are less likely to be registered. In addition, adding cause of death data to the record-linked data can also allow cause attribution and leading cause of death comparisons between the data sources. Such analyses, accompanied by careful interpretation, can form a useful basis from where to adjust cause of death data according to observed biases. At the individual level, misclassification patterns can be identified, which can offer insight into newly identified and re-occurring biases in cause of death attribution. Cause of death analyses using the record-linked data generated in this linkage study are presented in a forthcoming paper [49].

Conclusion

Despite strict policies to protect the confidentiality and safety of the data reported into each system, record linkage of mortality data between a CR system and an HDSS was possible in our study. To our knowledge, our study is the first in South Africa and possibly in sub-Saharan Africa to assess the feasibility and utility of linking HDSS and CR mortality data. The resultant data are useful for assessing selected population and individual health measures as referred to above, and hold potential to improve rural data quality.

We suggest the following five crucial contributions for further fruitful linkage exercises: the routine collection of national identity number in all the South African HDSSs; collaborative efforts to address place-name inconsistencies; recording of actual village/town/suburb names on death notification forms instead of tribal area names or adequate provision to provide for both; the development of an electronic place-name database, linked to detailed maps, against which to verify place names reported into the CR system, for use by Home Affairs registration

Table 4 Factors predictive of successful matching of death records between Agincourt HDSS and South African CR system

Variable	n	Matched n (%)	Univariate Odds ratio (95% confidence interval)	Multivariable Odds ratio (95% confidence interval)
	3726	2264 (60.8)		
Sex				
Female	1771	1104 (62.34)	1.00	1.00
Male	1955	1160 (59.34)	0.88 (0.77-1.01)	0.97 (0.82-1.14)
Age (years)				
<5	555	213 (38.38)	1.00	1.00
5-14	106	67 (63.21)	2.76 (1.79-4.24)***	2.83 (1.64-4.90)***
15-49	1729	1126 (65.12)	3.00 (2.46-3.65)***	2.89 (2.11-3.96)***
50-64	575	368 (64)	2.86 (2.24-3.63)***	2.24 (1.65-3.05)***
65+	761	490 (64.39)	2.90 (2.31-3.64)***	1.87 (1.41-2.49)***
Nationality				
Other nationality	1191	569 (47.77)	1.00	1.00
South African	2531	1695 (66.97)	2.22 (1.93-2.55)***	2.05 (1.70-2.48)***
National Identity number recorded in VA system				
Not available	2722	1324 (48.64)	1.00	1.00
Available	1004	940 (93.63)	15.51 (11.91-20.2)***	13.65 (10.12-18.43)***
Residency status				
Temporary and other	1211	642 (53.01)	1.00	1.00
Permanent	2515	1622 (64.49)	1.61 (1.40-1.85)***	1.28 (1.06-1.54)*
Education				
None	1720	936 (54.42)	1.00	1.00
Primary	1070	694 (64.86)	1.55 (1.32-1.81)***	0.92 (0.72-1.17)
Post primary	733	512 (69.85)	1.94 (1.61-2.33)***	1.02 (0.75-1.39)
Wealth quintile				
Lowest	605	308 (50.91)	1.00	1.00
Second	631	375 (59.43)	1.41 (1.13-1.77)**	1.21 (0.93-1.58)
Middle	647	406 (62.75)	1.62 (1.3-2.04)***	1.44 (1.1-1.88)*
Fourth	752	494 (65.69)	1.85 (1.48-2.3)***	1.59 (1.22-2.07)**
Highest	745	501 (67.25)	1.98 (1.59-2.47)***	1.73 (1.32-2.27)***
Year of death				
2006	885	453 (51.19)	1.00	1.00
2007	901	518 (57.49)	1.29 (1.07-1.55)**	1.16 (0.94-1.44)
2008	1024	655 (63.96)	1.69 (1.41-2.03)***	0.99 (0.79-1.24)
2009	916	638 (69.65)	2.19 (1.8-2.65)***	0.83 (0.64-1.06)
Place of death				
Hospital	1759	1117 (63.5)	1.00	1.00
Health center	42	25 (59.52)	0.85 (0.45-1.58)	1.07 (0.51-2.23)
Clinic	26	19 (73.08)	1.56 (0.65-3.73)	1.23 (0.41-3.68)

Table 4 Factors predictive of successful matching of death records between Agincourt HDSS and South African CR system *(Continued)*

Home	1525	937 (61.44)	0.92 (0.79-1.06)	1.12 (0.93-1.34)
Vehicle accident site	102	59 (57.84)	0.79 (0.53-1.18)	0.79 (0.48-1.29)
Other	255	98 (38.43)	0.36 (0.27-0.47)***	0.43 (0.31-0.6)***
Unknown	17	9 (52.94)	0.65 (0.25-1.68)	1.13 (0.31-4.06)

Statistical significance: ***P < 0.001; **P < 0.01; *P < 0.05.

Table 5 Background characteristics of all 2009 Agincourt HDSS deaths compared to those matched with CR records

Variable	All deaths in Agincourt HDSS (n = 846)	Deaths matched with CR records (n = 618)	
	n (%)	*n* (%)	*p*-value
Sex			
Female	411 (48.58)	307 (49.68)	
Male	435 (51.42)	311 (50.32)	0.679
Age (years)			
1-4	48 (5.67)	22 (3.56)	
5-14	25 (2.96)	20 (3.24)	
15-49	405 (47.87)	299 (48.38)	
50-64	138 (16.31)	109 (17.64)	
65+	230 (27.19)	168 (27.18)	0.431
Nationality			
Other	246 (29.08)	145 (23.46)	
South African	599 (70.80)	473 (76.54)	0.038
Residence status			
Permanent	623 (73.64)	467 (75.57)	
Temporary and other	223 (26.36)	151 (24.43)	0.404
Highest level of education			
None	350 (41.37)	238 (38.51)	
Primary	236 (27.90)	190 (30.74)	
Post primary	199 (23.52)	147 (23.63)	0.622
Place of death			
Hospital	439 (51.89)	327 (52.91)	
Health center	8 (0.95)	5 (0.81)	
Clinic	4 (0.47)	4 (0.65)	
Home	329 (38.89)	244 (39.48)	
Vehicle accident site	19 (2.25)	13 (2.10)	
Other	46 (5.44)	24 (3.88)	
Unknown	1 (0.12)	1 (0.16)	0.894
Wealth quintile			
Lowest	127 (15.01)	72 (11.65)	
Second	134 (15.84)	92 (14.89)	
Middle	150 (17.73)	111 (17.96)	
Fourth	184 (21.75)	142 (22.98)	
Highest	180 (21.28)	148 (23.95)	0.469

offices; and aligning study site borders with established official borders when setting up or extending HDSS sites.

Given our success in matching with surnames in this study and other studies' successes in using names [50,51], we additionally recommend that in addition to the surname, the deceased's full names (which are already captured on notice of death/stillbirth forms) be included in StatsSA datasets. Finally, concerted action among the governmental departments involved, health researchers, and relevant health data advisory committees is suggested to revitalize/modify the data fields on the notification form such that it is possible to identify the place of death, death registration, most recent employment prior to death, and residence of the deceased.

From a broader perspective, the methods and findings from this study are also of interest given the potential for application in other HDSS sites. Currently there are more than 45 HDSS sites across Africa, Asia and Oceania [16,18]. Conducting similar studies could serve to evaluate CR data where available, help identify gaps in national or sample CR systems, and where feasible, guide improved mortality and cause of death estimates. Of special interest would be the conduct of a similar study using data from an urban HDSS, such as DodaLab in Vietnam [52], to obtain empirical evidence for or against the general assumption that death registration is more complete in urban compared to rural areas and to help identify under-registered groups in urban areas. Such an empirical approach has potential to strengthen the evidence base for population health assessment and policy in developing countries where CR systems are weak.

Finally, our study provides scarce empirical evidence about factors affecting death registration, which has implications for strategies to accelerate death registration in countries with deficient CR systems.

Competing interests

The authors declare that they have no competing interests.

Authors' contributions

JJ wrote the study proposal and ethics applications and coordinated the collaboration, data preparation, and matching exercise. DB conceptualized the paper with inputs from JJ and CWK. CWK extracted the Agincourt HDSS data, and with inputs from DB and JJ, did the electronic matching, created the base analytic dataset, analyzed and interpreted the data, created the tables and figures, and wrote the first draft of the Introduction, Methods, and Results. JJ wrote the first draft of the Discussion and critically appraised the other sections for structure and intellectual content. CWK and JJ integrated the comments from co-authors and external referees. ADL, CR, DB, and TV consistently supplied critical inputs during all study phases. ADL, DB, KK, MT-N, ST, and TV made substantial contributions to acquisition of the data. MT-N led the vital registration data extraction and data security arrangements. All authors contributed to interpreting the data, critically reviewed the drafts, and approved the final manuscript.

Acknowledgments

We thank Statistics South Africa and the MRC/Wits Rural Public Health and Health Transitions Research Unit (Agincourt) for making available data used in this study. We are also grateful to Statistics South Africa for housing the matching exercise at the head office in Pretoria; Ms Ramadimetja Matji, Ms Aletia Barkley, and Ms Kerotse Mmatli for their participation in meetings and contributions to securing the data prior to the matching exercise; Ms Marlanie Moodley for preparing maps of the Agincourt and Bushbuckridge areas; and Ms Ria Laubscher for her assistance during the matching exercise.

Funding/Sponsorship

The study was supported by the MRC/Wits Rural Public Health and Health Transitions Research Unit (Agincourt), the South African Medical Research Council (MRC), and Statistics South Africa. The study was conducted while the second author held a University of Queensland Research Scholarship and the Endeavour International Postgraduate Research Scholarship at the University of Queensland, Brisbane, Australia. The Agincourt HDSS is funded by the Medical Research Council and University of the Witwatersrand, South Africa, Wellcome Trust, UK (grant no. 058893/Z/99/A, 069683/Z/02/Z, 085477/Z/08/Z), and National Institute on Aging of the NIH (grants 1R24AG032112-01 and 5R24AG032112-03). This paper was first presented at the INDEPTH Scientific Conference, October 2013, and was supported by an INDEPTH travel award. The funders had no role in study design, data collection, analysis, decision to publish, or preparation of the manuscript.

Author details

[1]MRC/Wits Rural Public Health and Health Transitions Research Unit (Agincourt), School of Public Health, Faculty of Health Sciences, University of the Witwatersrand, Johannesburg, South Africa. [2]Burden of Disease Research Unit, South African Medical Research Council, Parow Vallei, Western Cape, South Africa. [3]School of Population Health, The University of Queensland, Herston, Brisbane, Queensland, Australia. [4]Health and Vital Statistics, Statistics South Africa, Pretoria, South Africa. [5]Umeå Centre for Global Health Research, Division of Epidemiology and Global Health, Department of Public Health and Clinical Medicine, Umeå University, Umeå, Sweden. [6]INDEPTH Network, Accra, Ghana. [7]Melbourne School of Population and Global Health, The University of Melbourne, Carlton, Victoria, Australia. [8]Institute of Health Metrics and Evaluation, University of Washington, Seattle, USA.

Received: 17 April 2014 Accepted: 11 August 2014
Published: 30 August 2014

References

1. Lozano R, Naghavi M, Foreman K, Lim S, Shibuya K, Aboyans V, Abraham J, Adair T, Aggarwal R, Ahn SY, AlMazroa MA, Alvarado M, Anderson HR, Anderson LM, Andrews KG, Atkinson C, Baddour LM, Barker-Collo S, Bartels DH, Bell ML, Benjamin EJ, Bennett D, Bhalla K, Bikbov B, Abdulhak AB, Birbeck G, Blyth F, Bolliger I, Boufous S, Bucello C: **Global and regional mortality from 235 causes of death for 20 age groups in 1990 and 2010: a systematic analysis for the Global Burden of Disease Study 2010.** *Lancet* 2012, **380:**2095–2128.
2. Mahapatra P, Shibuya K, Lopez AD, Coullare F, Notzon FC, Rao C, Szreter S: **Civil registration systems and vital statistics: successes and missed opportunities.** *Lancet* 2007, **370:**1653–1663.
3. Wang H, Dwyer-Lindgren L, Lofgren KT, Rajaratnam JK, Marcus JR, Levin-Rector A, Levitz CE, Lopez AD, Murray CJ: **Age-specific and sex-specific mortality in 187 countries, 1970–2010: a systematic analysis for the Global Burden of Disease Study 2010.** *Lancet* 2012, **380:**2071–2094.
4. Carter KL, Rao C, Lopez AD, Taylor R: **Mortality and cause-of-death reporting and analysis systems in seven pacific island countries.** *BMC Public Health* 2012, **12:**436.
5. Rao C, Osterberger B, Anh TD, MacDonald M, Chúc NTK, Hill PS: **Compiling mortality statistics from civil registration systems in Viet Nam: the long road ahead.** *Bull World Health Organ* 2009, **87:**58–65.
6. United Nations, African Union Commission, African Development Bank: *Africa Programme on Accelerated Improvement of Civil Registation and Vital Statistics (APAI-CRVS).* New York: Economic Commission for Africa, United Nations; 2011.
7. Setel PW, Macfarlane SB, Szreter S, Mikkelsen L, Jha P, Stout S, AbouZahr C: **A scandal of invisibility: making everyone count by counting everyone.** *Lancet* 2007, **370:**1569–1577.
8. Bradshaw D, Groenewald P, Laubscher R, Nannan N, Nojilana B, Norman R, Pieterse D, Schneider M, Bourne DE, Timaeus I, Dorrington R, Johnson L: **Initial burden of disease estimates for South Africa, 2000.** *SAMJ* 2003, **93:**682–688.

9. Mathers CD, Ma Fat D, Inoue M, Rao C, Lopez AD: **Counting the dead and what they died from: an assessment of the global status of cause of death data.** *Bull World Health Organ* 2005, **83:**171–177c.
10. Bradshaw D, Kielkowski D, Sitas F: **New birth and death registration forms - a foundation for the future, a challenge for health workers?** *SAMJ* 1998, **88:**971–974.
11. Rao C, Bradshaw D, Mathers CD: **Improving death registration and statistics in developing countries: lessons from sub-Saharan Africa.** *South Afr J Demogr* 2004, **9:**81–99.
12. Bah S: **Multiple forces working in unison: the case of rapid improvement of vital statistics in South Africa post-1996.** *World Health Popul* 2009, **11:**50–59.
13. Rao C, Lopez AD, Yang G, Begg S, Ma J: **Evaluating national cause-of-death statistics: principles and application to the case of China.** *Bull World Health Organ* 2005, **83:**618–625.
14. Joubert J, Rao C, Bradshaw D, Vos T, Lopez AD: **Evaluating the quality of national mortality statistics from civil registration in South Africa, 1997–2007.** *PLoS ONE* 2013, **8:**e64592.
15. INDEPTH Network: *Using Longitudinal INDEPTH Data, National Censuses, DHS, and Other National Surveys for Better Health Policy in Africa. Report of Meeting nr 1352.* Bellagio, Italy: INDEPTH Network; 2012.
16. Sankoh O: **Global health estimates: stronger collaboration needed with low-and middle-income countries.** *PLoS Med* 2010, **7:**e1001005.
17. Ye Y, Wamukoya M, Ezeh A, Emina J, Sankoh O: **Health and demographic surveillance systems: a step towards full civil registration and vital statistics system in sub-Sahara Africa?** *BMC Public Health* 2012, **12:**741.
18. Sankoh O, Byass P: **The INDEPTH Network: filling vital gaps in global epidemiology.** *Int J Epidemiol* 2012, **41:**579–588.
19. Kahn K, Collinson MA, Gómez-Olivé FX, Mokoena O, Twine R, Mee P, Afolabi SA, Clark BD, Kabudula CW, Khosa A, Khoza S, Shabangu MG, Silaule B, Tibane JB, Wagner RG, Garenne ML, Clark SJ, Tollman SM: **Profile: Agincourt Health and Socio-demographic Surveillance System.** *Int J Epidemiol* 2012, **41:**988–1001.
20. Kahn K, Tollman SM, Collinson MA, Clark SJ, Twine R, Clark BD, Shabangu M, Gomez-Olive FX, Mokoena O, Garenne ML: **Research into health, population and social transitions in rural South Africa: data and methods of the Agincourt Health and Demographic Surveillance System.** *Scand J Public Health* 2007, **35:**8–20.
21. Kahn K, Tollman SM, Garenne M, Gear JSS: **Validation and application of verbal autopsies in a rural area of South Africa.** *Trop Med Int Health* 2000, **5:**824–831.
22. Division UNS: *Principles and Recommendations for a Vital Statistics System. Revision 3. Final Draft.* New York: United Nations; 2013.
23. Republic of South Africa: *Births and Deaths Registration Act, 1992 (No. 51 of 1992). In: Government Gazette No. 13953.* Cape Town: Government Printer; 1992.
24. Statistics South Africa: *Mortality and Causes of Death in South Africa, 2008: Findings from Death Notification. Statistical Release P0309.3.* Pretoria: Statistics South Africa; 2010.
25. **Deaths Certificates.** [http://www.home-affairs.gov.za/index.php/death-certificates1]
26. Li B, Quan H, Fong A, Lu M: **Assessing record linkage between health care and Vital Statistics databases using deterministic methods.** *BMC Health Serv Res* 2006, **6:**48.
27. Machado CJ: **A literature review of record linkage procedures focusing on infant health outcomes.** *Cadernos de Saúde Pública* 2004, **20:**362–371.
28. Maso LD, Braga C, Franceschi S: **Methodology used for software for automated linkage in Italy (SALI).** *Comp Biomed Research* 2001, **34:**395.
29. Victor TW, Mera RM: **Record linkage of health care insurance claims.** *J Am Med Inform Assoc* 2001, **8:**281–288.
30. Winkler WE: **String Comparator Metrics and Enhanced Decision Rules in the Fellegi-Sunter Model of Record Linkage.** In *Proceedings of the Section on Survey Research Methods.* Alexandria: American Statistical Association; 1990:354–359.
31. Durham E, Xue Y, Kantarcioglu M, Malin B: **Quantifying the correctness, computational complexity, and security of privacy-preserving string comparators for record linkage.** *Inform Fusion* 2012, **13:**245–259.
32. Jaro MA: **Advances in record-linkage methodology as applied to matching the 1985 census of Tampa, Florida.** *J Am Stat Assoc* 1989, **84:**414–420.
33. Sariyar M, Borg A, Pommerening K: **Evaluation of record linkage methods for iterative insertions.** *Methods Inf Med* 2009, **48:**429–437.
34. Howe GR: **Use of computerized record linkage in cohort studies.** *Epidemiol Rev* 1998, **20:**112–121.
35. Beauchamp A, Tonkin AM, Kelsall H, Sundararajan V, English DR, Sundaresan L, Wolfe R, Turrell G, Giles GG, Peeters A: **Validation of de-identified record linkage to ascertain hospital admissions in a cohort study.** *BMC Med Res Methodol* 2011, **11:**42.
36. Cook L, Olson L, Dean J: **Probabilistic record linkage: relationships between file sizes, identifiers, and match weights.** *Methods Inf Med* 2001, **40:**196–203.
37. Jaro MA: **Probabilistic linkage of large public health data files.** *Stat Med* 1995, **14:**491–498.
38. Nitsch D, Morton S, DeStavola BL, Clark H, Leon DA: **How good is probabilistic record linkage to reconstruct reproductive histories? Results from the Aberdeen children of the 1950 s study.** *BMC Med Res Methodol* 2006, **6:**15.
39. Grannis SJ, Overhage JM, Hui S, McDonald CJ: **Analysis of a probabilistic record linkage technique without human review.** *AMIA Annu Symp Proc* 2003, **2003:**259–263.
40. Herzog TN, Scheuren F, Winkler WE: *Data Quality and Record Linkage Techniques.* Heidelberg: Springer; 2007.
41. Méray N, Reitsma JB, Ravelli AC, Bonsel GJ: **Probabilistic record linkage is a valid and transparent tool to combine databases without a patient identification number.** *J Clin Epidemiol* 2007, **60:**883–e881.
42. Tromp M, Ravelli AC, Bonsel GJ, Hasman A, Reitsma JB: **Results from simulated data sets: probabilistic record linkage outperforms deterministic record linkage.** *J Clin Epidemiol* 2011, **64:**565–572.
43. Filmer D, Pritchett LH: **Estimating wealth effects without expenditure data or tears: an application to educational enrollments in states of India.** *Demography* 2001, **38:**115–132.
44. *SimMetrics*; [http://sourceforge.net/projects/simmetrics]
45. Roos NP, Black CD, Frohlich N, Decoster C, Cohen MM, Tataryn DJ, Mustard CA, Toll F, Carriere KC, Burchill CA, MacWilliam L, Bogdanovic B: **A population-based health information system.** *Med Care* 1995, **33:**DS13–DS20.
46. Karmel R, Rosman D: **Linkage of health and aged care service events: comparing linkage and event selection methods.** *BMC Health Serv Res* 2008, **8:**149.
47. Chandrasekar C, Deming WE: **On a method of estimating birth and death rates and the extent of registration.** *J Am Stat Assoc* 1949, **44:**101–115.
48. Hook EB, Regal RR: **Capture-recapture methods in epidemiology: methods and limitations.** *Epidemiol Rev* 1995, **17:**243–264.
49. Joubert J, Bradshaw D, Kabudula C, Rao C, Kahn K, Mee P, Tollman S, Lopez AD, Vos T: **Record-linkage comparison of verbal autopsy and routine civil registration death certification in rural north-east South Africa: 2006–09.** *Int J Epidemiol* In press.
50. Kabudula CW, Clark BD, Gómez-Olivé FX, Tollman S, Menken J, Reniers G: **The promise of record linkage for assessing the uptake of health services in resource constrained settings: a pilot study from South Africa.** *BMC Med Res Methodol* 2014, **14:**71.
51. Quantin C, Binquet C, Bourquard K, Pattisina R, Gouyon-Cornet B, Ferdynus C, Gouyon J-B, Allaert F-A: **Which are the best identifiers for record linkage?** *Inf Health and Social Care* 2004, **29:**221–227.
52. Tran TK, Eriksson B, Nguyen CT, Horby P, Bondjers G, Petzold M: **DodaLab: an urban health and demographic surveillance site, the first three years in Hanoi, Vietnam.** *Scand J Public Health* 2012, **40:**765–772.

doi:10.1186/s12963-014-0023-z
Cite this article as: Kabudula *et al.*: Evaluation of record linkage of mortality data between a health and demographic surveillance system and national civil registration system in South Africa. *Population Health Metrics* 2014 **12**:23.