IJC
International Journal of Cancer

Record linkage to correct under-ascertainment of cancers in HIV cohorts: The Sinikithemba HIV clinic linkage project

Mazvita Sengayi[1,2], Adrian Spoerri[3], Matthias Egger[3,4], Danuta Kielkowski[1], Tamaryn Crankshaw[5], Christie Cloete[6], Janet Giddy[6] and Julia Bohlius[3]

[1] National Cancer Registry, National Health Laboratory Service, Johannesburg, South Africa
[2] Graduate School for Cellular and Biomedical Sciences, University of Bern, Bern, Switzerland
[3] Institute of Social and Preventive Medicine (ISPM), University of Bern, Bern, Switzerland
[4] Centre for Infectious Disease Epidemiology and Research (CIDER), School of Public Health and Family Medicine, University of Cape Town, Cape Town, South Africa
[5] Health Economics HIV And AIDS Research Division (HEARD), University of KwaZulu-Natal, Durban, South Africa
[6] McCord Hospital, Overport, Durban, South Africa

The surveillance of HIV-related cancers in South Africa is hampered by the lack of systematic collection of cancer diagnoses in HIV cohorts and the absence of HIV status in cancer registries. To improve cancer ascertainment and estimate cancer incidence, we linked records of adults (aged ≥ 16 years) on antiretroviral treatment (ART) enrolled at Sinikithemba HIV clinic, McCord Hospital in KwaZulu-Natal (KZN) with the cancer records of public laboratories in KZN province using probabilistic record linkage (PRL) methods. We calculated incidence rates for all cancers, Kaposi sarcoma (KS), cervix, non-Hodgkin's lymphoma and non-AIDS defining cancers (NADCs) before and after inclusion of linkage-identified cancers with 95% confidence intervals (CIs). A total of 8,721 records of HIV-positive patients were linked with 35,536 cancer records. Between 2004 and 2010, we identified 448 cancers, 82% ($n = 367$) were recorded in the cancer registry only, 10% ($n = 43$) in the HIV cohort only and 8% ($n = 38$) both in the HIV cohort and the cancer registry. The overall cancer incidence rate in patients starting ART increased from 134 (95% CI 91–212) to 877 (95% CI 744–1,041) per 100,000 person-years after inclusion of linkage-identified cancers. Incidence rates were highest for KS (432, 95% CI 341–555), followed by cervix (259, 95% CI 179–390) and NADCs (294, 95% CI 223–395) per 100,000 person-years. Ascertainment of cancer in HIV cohorts is incomplete, PRL is both feasible and essential for cancer ascertainment.

Key words: record linkage, HIV, cancer, South Africa

Abbreviations: ADCs: AIDS defining cancers; ART: antiretroviral therapy; CI: confidence interval; HIV: human immunodeficiency virus; ICC: invasive cervical cancer; IeDEA: international epidemiological databases to evaluate AIDS; IQR: interquartile range; KS: Kaposi's sarcoma; KZN: KwaZulu-Natal; NADCs: non-AIDS defining cancers; NCR: national cancer registry of South Africa; NHL: non-Hodgkin's lymphoma; NHLS: national health laboratory service; PRL: probabilistic record linkage

Additional Supporting Information may be found in the online version of this article.

DOI: 10.1002/ijc.30154

History: Received 8 Sep 2015; Accepted 7 Apr 2016; Online 21 Apr 2016

Correspondence to: Dr. Mazvita Sengayi, National Cancer Registry, National Health Laboratory Service, 47 de Korte Street, Braamfontein, Johannesburg 2001, South Africa, Tel: +27-11-4899150; Fax: +27-11-4899152, E-mail: mazvita.sengayi@nhls.ac.za

The association between HIV and cancer is well-established,[1] but surveillance of HIV-related cancers is hampered by the lack of systematic collection of cancer diagnoses in HIV cohorts and the absence of HIV status in cancer registries. While probabilistic record linkage (PRL) has been widely used for cancer ascertainment in HIV in Europe and America,[2,3] it has not been fully exploited in resource-limited settings. The availability of electronic databases of HIV cohorts and a national cancer registry in South Africa, presents the opportunity of using data linkage to improve cancer diagnosis ascertainment in HIV cohorts, and to consequently be able to study prevalence and incidence of cancer in HIV-positive South Africans. The Uganda AIDS-Cancer Registry Match Study was the first HIV cancer linkage study conducted in sub-Saharan Africa before the era of antiretroviral treatment (ART) in Uganda.[4] Recently, the Nigerian AIDS-Cancer Match Study described cancer incidence in HIV-positive patients in the Nigerian ART era.[5] As ART coverage improves in Southern Africa, there is need for data on the spectrum and incidence of cancer in HIV-positive people.

Computerised record linkage techniques can be used to identify records belonging to the same individual from data sources collected at different times and places.[6,7] Deterministic record linkage involves exact matching of records requiring a

What's new?
The surveillance of HIV-related cancers is hampered by the lack of systematic collection of cancer diagnoses in HIV cohorts and the absence of HIV status in cancer registries. While probabilistic record linkage has been widely used in Europe and America, it has not been fully exploited in resource-limited settings. This study demonstrates the utility of record linkage in correcting under-ascertainment of cancers in HIV cohorts and a high cancer incidence among HIV-positive patients on antiretroviral therapy in South Africa. There is a need for a systematic approach to cancer surveillance in HIV-positive people in the South African antiretroviral therapy era.

full match on one or more linkage variables.[7] Routinely collected public health data are often prone to registration errors and unique identifiers (such as national identification number) are often not routinely or reliably collected, making the merging of databases with a unique identifier impractical. PRL is a matching method where for each potential pair of records a likelihood of representing the same identity (*e.g.*, person) is calculated, based on conditional probabilities.[6,8] It allows for meaningful interpretation of small differences in linkage variables, like typographical errors in names or small differences in dates. The accuracy of PRL is determined by several factors. The number of matching variables, accuracy of each variable with respect to missing data, typographical errors during data capture, unreliable coding, duplicate records, change in variables with time (*e.g.*, last name and marital status change with marriage) and discriminating power of variables.[9–11] Poor data quality increases the risk of false negatives (missed matches) and low discriminating power increases false positives (false matches) and can lead to misclassification bias.[12,13] Despite these shortfalls, PRL has been previously used to link public health databases with high precision and sufficient accuracy for epidemiological research.[9]

We conducted a PRL with a cohort of HIV-positive patients on antiretroviral therapy and the laboratory-based cancer registry in South Africa to ascertain cancer diagnoses and estimate cancer incidence.

Material and Methods

Data sources and settings

The pathology-based National Cancer Registry (NCR) was established in 1986 and is South Africa's national cancer surveillance system. The NCR is a division of the National Health Laboratory Services (NHLS), which comprises state-funded public laboratories serving over 80% of the South African population.[14] The NHLS is the primary source of data for the NCR. All malignancies diagnosed in public laboratories in KwaZulu-Natal (KZN) province from January 2004 to December 2010 were included in the linkage. Cancers diagnosed in private laboratories were not included in the linkage. The total number of cancer reports in this time period was 35,536 from 32,923 patients. Patient demographic data, tumour topography, morphology and date of diagnosis were extracted from the pathology reports. All cancers were coded according to the International Classification of Diseases for Oncology third edition (ICD-O-3).

International epidemiological Databases to Evaluate AIDS Southern Africa (IeDEA) is an international research consortium of HIV observational databases that includes four regional networks in sub-Saharan Africa.[15] The Southern African region (IeDEA-SA, www.iedea-sa.org) includes ART programs located in seven countries (Botswana, Lesotho, Malawi, Zambia, Mozambique, Zimbabwe and Republic of South Africa).[15] In the current study, we used data from only one cohort in South Africa, the Sinikithemba ART clinic. The Sinikithemba ART clinic, previously located at a state-subsidized hospital in Durban, KZN province provided HIV care to patients since 1999 up to its closure in June 2012. During the study period the clinic received PEPFAR (United States' President's Emergency Plan for AIDS Relief) funding. Patients paid a monthly copayment (ZAR180/USD18 at the time of closure).[16] Medications and laboratory testing processed by the NHLS were provided by the South African Department of Health from 2006. The clinic had full-time data managers and research staff who maintained the clinic electronic database.

Record linkage

Patient data of all HIV-positive adults aged ≥ 16 years at start of ART attending the Sinikithemba clinic between January 2004 and December 2011 ($n = 8{,}721$) were linked with the cancer records of public laboratories in KZN province for the time period January 2004 to December 2010 (32,923 patients and 35,536 cancer cases). Cancers were not systematically recorded in the Sinikithemba clinic database; however, 93 cancer cases were documented in the electronic patients' notes and known prior to record linkage. We linked records probabilistically using the record linkage software G-Link of Statistics Canada[17] based on name and surname, year of birth, gender, date of death and date of cancer diagnoses. The main steps in PRL are selection of linkage variables, data pre-processing, blocking, matching and post-processing.[7,13,18] Details of the record linkage procedure are provided in the Supporting Information. In brief, after pre-processing the data, we applied blocking criteria to identify potential pairs and to increase the efficiency of the linkage process. In the next step (matching), linkage rules were applied, probability weights estimated and upper and lower thresholds defined. The probability weights and the defined thresholds were

Table 1. Patient characteristics

	Patients with cancer	Patients not developing cancer	Total
N	431	6,920	7,351
Gender			
Male	158 (36.7)	2,743 (39.6)	2,901 (39.5)
Female	273 (63.3)	4,177 (60.4)	4,450 (60.5)
Median age at ART start in years (IQR)	36 (31–43)	35 (30–42)	35 (30–42)
16–29 years	70 (16.2)	1,503 (21.7)	1,573 (21.4)
30–34 years	107 (24.8)	1,653 (23.9)	1,760 (23.9)
35–39 years	107 (24.8)	1,530 (22.1)	1,637 (22.3)
40–54 years	125 (29.0)	1,910 (27.6)	2,035 (27.7)
≥55 years	22 (5.10)	324 (4.68)	346 (4.71)
Year start ART			
2004–2006	141 (32.7)	2,300 (33.2)	2,441 (33.2)
2007–2010	290 (67.3)	4,620 (66.8)	4,910 (66.8)

Abbreviations: ART: antiretroviral therapy; IQR: interquartile range.

used to classify the potential pairs as definite matches, possible matches and impossible matches, which were rejected. In the next step (post-processing), a thorough clerical review of the possible pairs was conducted and possible pairs were reclassified as definite or rejected pairs. For each definite match, we retrieved data on the cancer diagnosis, date of diagnosis, ICD-O-3 topography (site) and morphology (histology and behaviour) code. The additional information from the KZN cancer records was then incorporated into the IeDEA database and all patient identifying information was removed.

Patient characteristics and spectrum of cancers

We used frequencies to describe patients characteristics and the spectrum of cancers diagnosed before (prevalent) and after (incident) starting ART plus 30 days in patients starting ART between January 2004 and December 2010. Infection-related cancers were defined as cancers that are typically associated with Group 1 infectious agents (classified as carcinogenic to humans by International Agency for Research on Cancer).[19,20] Age at starting ART was described using medians and interquartile ranges (IQRs).

Cancer incidence

For the analysis of cancer incidence we included all patients aged ≥16 years at ART initiation, who started ART at the Sinikithemba ART clinic between January 2004 and December 2010. We excluded patients diagnosed with cancer before ART initiation or within 30 days after (considered as prevalent cancer cases), and all patients with follow-up time of <30 days from ART initiation. Person-years (pys) at risk were calculated from 30 days after ART initiation until the date of first cancer diagnosis, death or 6 months after the last visit for patients who did not die, whichever occurred first. In sensitivity analyses, we calculated pys from 30 days after ART initiation to date of first cancer diagnosis, death, or for patients who did not die to *i*) date of last visit, *ii*) date of data base closure[21] and *iii*) censored 2 years after starting ART, see Supporting Information Table S2. We calculated age standardised incidence rates by dividing the cancer cases by the number of pys at risk weighted for the world standard population.[22] We present age standardised cancer incidence rates per 100,000 pys for specific cancers for men and women and in total. Because of few incident cancer cases we have grouped vulva, ano-rectal and vagina as “ano-genital cancers other than cervical cancer”. All other cancers with ≤5 incident cases were grouped as “other”. To assess changes over calendar period of starting ART we plotted the hazard function of cancer since starting ART by calendar period (early 2004–2006 versus late 2007–2010 ART period) and estimated the hazard ratio of late versus early ART period adjusted for age and gender. To make the follow-up length in the two periods comparable, we restricted these analyses to a maximum follow-up time of 3 years. In addition, we plotted Kaplan–Meier curves to show cumulative cancer incidence of KS, NHL, ICC and NADCs over time since starting ART. To assess the degree of under-ascertainment we calculated the incidence rate of all cancers, Kaposi sarcoma (KS), non-Hodgkin’s lymphoma (NHL), invasive cervical cancer (ICC) and non-AIDS defining cancers (NADCs) before and after inclusion of linkage-identified cancers and calculated the incidence rate difference.

All data including names were stored in encrypted format. After the linkage process all names were removed. Analyses were done with STATA version 13 (Stata Corporation, College Station, TX) using anonymised data. The study was approved by the McCord Research Ethics Committee in Durban, South Africa.

Table 2. Spectrum of cancers before and after starting ART

	Prevalent cancers, *N* (%)	Incident cancers, *N* (%)	Total, *N* (%)
Total cancers, *N*	250 (55.8)	198 (44.9)	448 (100)
Type of cancer			
ADCs	*162 (64.8)*	*126 (63.6)*	*288 (64.3)*
Kaposi sarcoma[1]	109 (43.6)	77 (38.9)	186 (41.5)
Cervix[1]	42 (16.8)	32 (16.2)	74 (16.5)
NHL[1]	11 (4.4)	17 (8.6)	28 (6.3)
NADCs	*88 (35.2)*	*72 (36.4)*	*160 (35.7)*
Conjunctiva, Eye[1]	19 (7.6)	8 (4.0)	27 (6.0)
Breast	10 (4.0)	12 (6.1)	22 (4.9)
Lip, oral cavity and pharynx[1]	5 (2.0)	8 (4.0)	13 (2.9)
Skin-excluding BCC	5 (2.0)	5 (2.5)	10 (2.2)
Liver[1]	7 (2.8)	1 (0.5)	8 (1.8)
Anorectal[1]	4 (1.6)	3 (1.5)	7 (1.6)
Nasal cavity and larynx[1]	5 (2.0)	2 (1.0)	7 (1.6)
Oesophagus	1 (0.4)	5 (2.5)	6 (1.3)
Vulva[1]	1 (0.4)	5 (2.5)	6 (1.3)
Colon	2 (0.8)	3 (1.5)	5 (1.1)
Uterine	3 (1.2)	2 (1.0)	5 (1.1)
Vagina[1]	1 (0.4)	4 (2.0)	5 (1.1)
Penis[1]	4 (1.6)	1 (0.5)	5 (1.1)
Bladder and urethra[1]	1 (0.4)	4 (2.0)	5 (1.1)
Hodgkin's lymphoma[1]	4 (1.6)	1 (0.5)	5 (1.1)
Lung	2 (0.8)	2 (1.0)	4 (0.9)
Stomach[1]	3 (1.2)	1 (0.5)	4 (0.9)
Basal cell carcinoma of skin	4 (1.6)	–	4 (0.9)
Prostate	2 (0.8)	1 (0.5)	3 (0.7)
Ill-defined sites	1 (0.4)	2 (1.0)	3 (0.7)
Ovary	2 (0.8)	–	2 (0.4)
Nerve	1 (0.4)	–	1 (0.2)
Kidney	1 (0.4)	–	1 (0.2)
Brain	–	1 (0.5)	1 (0.2)
Thyroid	–	1 (0.5)	1 (0.2)

[1] Cancers that are typically associated with group 1 infectious agents according to International Agency for Research on Cancer[19,20]: HPV: lip, oral cavity and pharynx, ano-rectal, vulva, vagina, cervix and penile cancer; HHV8: Kaposi's sarcoma; EBV: nasopharyngeal cancer, non-Hodgkin's and Hodgkin's lymphoma; *H. pylori*: stomach cancer; schistosomiasis: bladder cancer; hepatitis B/C: liver cancer; HIV only (sufficient evidence of HIV association but association with other infectious agents not confirmed): conjunctival cancer.

Abbreviations: ART: antiretrioviral therapy; ADCs: AIDS defining cancers; NADCs: non-AIDS defining cancers; NHL: non-Hodgkin's lymphoma; BCC: basal cell carcinoma of the skin.

Results

Linkage results and analyses data sets

A total of 246,897,369 potential record pairs were created after applying the blocking criteria, see Supporting Information Figure S1. These potential pairs were passed through the linkage rules (matching) resulting in 246,776,426 directly rejected pairs and 120,943 possible pairs. In an iterative process of refining linkage weights and applying thresholds, the possible pairs were reclassified as definite pairs (435, 0.4%), rejected pairs (116,927; 96.7%), and 3,581 (3%) remained possible pairs. In the post-processing step, we reviewed these 3,581 possible matches carefully and reclassified them as definite ($n = 571$) or rejected pairs ($n = 3{,}010$). The final set of definite pairs after post-processing had 1,006 definite record pairs. 460 of these were invasive cancers and 546 were cancer precursor lesions which were not considered in the analyses.

Cancer Epidemiology

Table 3. Cancer incidence rates in men and women after starting ART

	Men		Women		Total	
Type of cancer	*n* (%)	Incidence rate[1]	*n* (%)	Incidence rate[1]	*n* (%)	Incidence rate[1]
Total	48 (100)	752 (567–1,017)	94 (100)	949 (776–1,172)	142 (100)	877 (744–1,041)
ADCs	31 (64.6)	487 (345–712)	64 (68.1)	653 (512–845)	95 (66.9)	592 (485–731)
KS	29 (60.4)	457 (320–677)	39 (41.5)	417 (306–583)	68 (47.9)	432 (341–555)
ICC	–	–	27 (28.7)	259 (179–390)	27 (19.0)	259 (179–390)
NHL	6 (12.5)	88 (40–235)	9 (9.6)	83 (44–178)	15 (10.6)	85 (52–150)
NADCs	18 (37.5)	271 (172–454)	32 (34.0)	307 (219–445)	50 (35.2)	294 (223–395)
Breast	–	–	11 (11.7)	94 (53–185)	11 (7.7)	94 (53–185)
Ano-genital other than cervix	1 (2.1)	–	10 (10.6)	105 (57–214)	11 (7.7)	72 (42–146)
Conjunctiva	2 (4.2)	37 (8–411)	6 (6.4)	55 (25–146)	8 (5.6)	49 (25–110)
Lip, oral cavity and pharynx	5 (10.4)	80 (34–239)	2(2.1)	18 (4–185)	7 (4.9)	41 (20–99)
Other	14 (29.2)	193 (115–348)	13 (13.8)	118 (70–217)	27 (19.0)	146 (101–219)

[1] Age standardised incidence rates per 100,000 person-years (95% confidence interval). Other included cancers of the liver, penis, bladder and urethra, Hodgkin's lymphoma, skin, oesophagus, lung, prostate, colon, brain and ill-defined sites.

Abbreviations: ART: antiretroviral therapy; ADCs: AIDS defining cancers; KS: Kaposi's sarcoma; ICC: invasive cervical cancer; NHL: non-Hodgkin's lymphoma; NADCs: non-AIDS defining cancers.

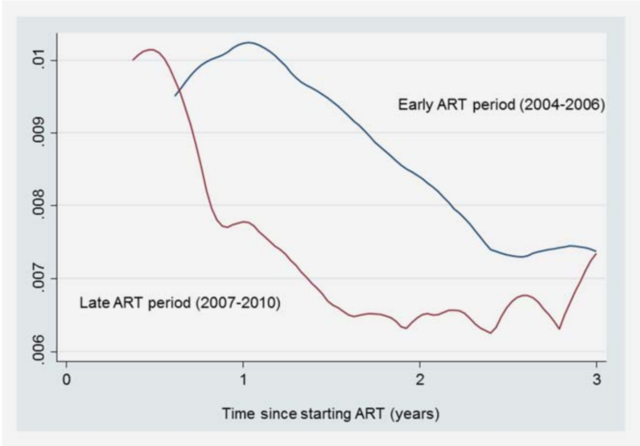

Figure 1. Hazard of developing cancer after starting ART: early (2004–2006) versus late (2007–2010) ART period. Abbreviations: ART: antiretroviral treatment.

Prior to record linkage, 93 cancers were documented in the HIV cohort, 43 of these were also identified through record linkage but the remaining 50, were only recorded in the HIV cohort. This brought the total number of cancer cases identified to 510 ($n = 460 + 50$). For analyses we removed patients without cancer date ($n = 6$) and patients starting ART before January 2004 or after December 2010 (1,312 patients without and 52 patients with cancer), see Supporting Information Figure S2. The resulting data set with 7,351 patients was used to describe the study population and the spectrum of cancers. For the analysis of incidence we further removed all patients diagnosed with cancer before starting ART ($n = 222$) and patients with follow-up time <30 days after starting ART (720 patients without and 39 patients with cancer). We censored cancer cases that were diagnosed >180 days after the last follow-up contact. The final data set for the analysis of cancer incidence included 6,370 patients. Of these 142 developed at least one incident cancer and 6,228 remained free from cancer during the study period, see Supporting Information Figure S2.

Patient characteristics

Table 1 shows the characteristics of the 7,351 patients included in the analysis. 61% of the total population was female, the median age at starting ART was 35 years (IQR 30–42 years). A total of 431 patients were diagnosed with at least one prevalent or incident cancer. Patients developing cancer were similar to those not developing cancer. Sixty-three percent of cancer patients were women ($n = 273$) and the median age at cancer diagnosis was 36 years (IQR 31–43 years).

Spectrum of cancers—Prevalent and incident cancer cases

Table 2 shows the spectrum of all cancers identified before (prevalent) and after (incident) starting ART plus 30 days. Eighty-two percent ($n = 367$) of all cancer cases ($n = 448$) were recorded in the cancer registry only, 10% ($n = 43$) were recorded in the HIV cohort only and 8% ($n = 38$) were recorded in both the HIV cohort and the cancer registry. The 448 prevalent and incident cancers cases occurred in 431 patients; 15 patients had two cancers and 1 patient had three cancers. For a complete listing of secondary cancers see Supporting Information Table S1. About half (56%) of all cancers were prevalent occurring prior to ART initiation or within 30 days after ART initiation. Sixty-four percent of all cancers were AIDS defining (KS, NHL and ICC) and 83% were infection-related. Cancers associated with human herpes virus 8 (*i.e.*, KS) accounted for 42% of all cancers. HPV-related cancers (*i.e.*, cervix, ano-rectal, vulva, vagina, penile, lip, oral

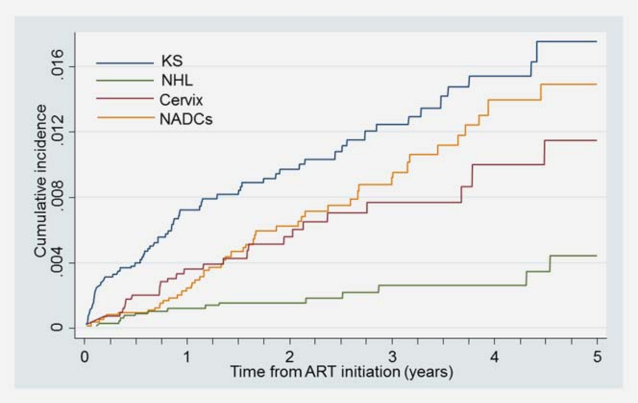

Figure 2. Cumulative incidence of KS, ICC, NHL and NADCs after starting ART. Abbreviations: KS: Kaposi's sarcoma; ICC: invasive cervical cancer; NHL: non-Hodgkin's lymphoma; NADCs: non AIDS-defining cancers; ART: antiretroviral treatment.

cavity and pharynx cancer) accounted for 25% and cancers typically associated with Epstein Barr virus (*i.e.*, NHL and Hodgkin's lymphoma, nasopharyngeal cancer), represented 7% of all cancers. The distribution of cancer types was similar between prevalent and incidence cancer cases.

Cancer incidence rate

6,370 patients were included in the incidence analysis, contributing a total of 15,976 pys. The median follow-up time was 2.0 years (IQR 1.2–3.6 years), the maximum follow-up time was 7.4 years. The overall age-standardised cancer incidence rate in patients starting ART was 877 [95% confidence interval (CI) 744–1,041] per 100,000 pys (Table 3). Cancer incidence rates in the main analyses and in sensitivity analyses using different definitions for the end of observation time were comparable (see Supporting Information Table S2). Total incidence rates were highest for ADCs (592; 95% CI 485–731) and infection-related cancers (753, 95% CI 630–907) per 100,000 pys. The overall cancer incidence rate was higher in women (949; 95% CI 776–1,172) than in men (752; 95% CI 567–1,017) per 100,000 pys. The most frequent cancer both in men and women was KS with an incidence rate of 457 (95% CI 320–677) in men and 417 (95% CI 306–583) in women per 100,000 pys. Other frequent cancers were ICC (259, 95% CI 179–390) and the group of NADCs (294, 95% CI 223–395) per 100,000 pys.

Figure 1 shows the hazard of developing cancer by time since starting ART stratified by calendar period of ART start year (2004–2006 versus 2007–2010). The hazard of developing cancer was lower in the more recent calendar period, however, the hazard ratio (HR) comparing both periods was not statistically significant (HR 0.90, 95% CI 0.63–1.30, adjusted for age and gender). Figure 2 shows the cumulative incidence of KS, ICC, NHL and NADCs. Of note is the steady rise of NADCs gaining prominence over the cervix and NHL by the second year on ART.

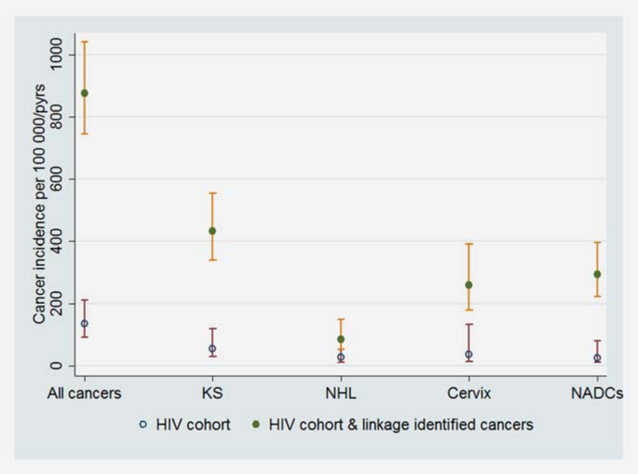

Figure 3. Cancer incidence rate per 100,000 person-years in patients after starting ART: before and after inclusion of linkage identified cancers. Abbreviations: KS: Kaposi's sarcoma; NHL: non-Hodgkin's lymphoma; NADCs: non AIDS-defining cancers; ART: antiretroviral treatment.

Cancer under-ascertainment

Figure 3 shows the incidence rate of all cancers, KS, NHL, ICC and NADCs before and after inclusion of linkage identified cancers. For all types of cancers analysed the incidence rates increased after inclusion of linkage identified cancers. The incidence rate of all cancers increased from 134 (95% CI 91–212) to 877 (95% CI 744–1,041) per 100,000 pys, for an incidence rate difference of 745 (95% CI 590–904). Incidence rates differences per 100,000 pys were 368 (95% CI 262–474) for KS, 229 (95% CI 121–338) for ICC, 279 (95% CI 189–369) for NADCs and 62 (95% CI 8–116) for NHL.

Discussion

In this South African cohort of HIV-positive adults starting ART, the majority of cancer cases identified were not recorded in the HIV cohort and only identified through linkage to the cancer registry (82%, 367). Despite starting ART overall cancer incidence rates were high, with 949/100,000 pys in women and 752/100,000 pys in men. Incidence rates were highest for KS (432, 95% CI 341–555), followed by ICC (259, 95% CI 179–390) and NADCs (294, 95% CI 223–395) per 100,000 pys. Patients were young at the time of cancer diagnosis and most patients were female.

This is the first study to use PRL to identify cancers in HIV-positive adults in the era of ART in South Africa. Our study had several strengths. We had histological confirmation of all cancers identified through linkage with the cancer registry. The use of PRL allowed matching HIV and cancer data based on the similarity of the data and accepting minor differences and errors (*e.g.*, typographical errors). PRL is a well-recognised method of record linkage[9] and allows for optimum use of the available data given the data quality limitations of routinely collected data. We used cancer data and HIV data from the same province and this potentially increased the prevalence of true pairs. Clerical review was done for uncertain pairs, allowing for exclusion of any bad pairs. Most of the common cancers identified are known HIV-related cancers and this gives some confidence that most of them were true matches. For cancer incidence analyses, we used different censoring strategies for calculation of follow-up time which yielded comparable estimates suggesting that our results were robust (see Supporting Information Table S2). Despite these strengths, our study had some limitations. There were a limited number of linkage variables common to both datasets. We used only names, gender and date of birth and for the majority of cancer records (61%) the year of birth was estimated from stated age at cancer diagnosis. Data quality limitations might have resulted in some missed matches particularly due to the lack of reliability of the date of birth variable and the high frequency of common names and surnames. KZN has a high proportion of people of Zulu origin and Indian origin and regional clustering of similar names in this area could have reduced the discriminating power of names and increased risk of matches by chance (false positives). Linkages in other countries have used unique personal identification numbers[3] and postal address codes,[23] resulting in more precise linkages, but these variables were not available for the current study. Cancers diagnosed in private laboratories in KZN were not included in the linkage because up to 2011 private laboratories from all over South Africa sent pathology reports of confirmed malignancies to the NCR on a voluntary basis only. However, this was unlikely to affect our results. The private sector covers 15% of the South African population that has medical insurance and people who access HIV care in public sector are unlikely to access cancer care in private sector because of the cost.[14] Since 2011, Regulation 380 of the National Health Act requires mandatory reporting of all confirmed cancers to the NCR including private laboratories.[24] We therefore expect that future registry linkages would be even more complete than the current study. Cancers diagnosed in public laboratories in other provinces were potentially missed. We used the data of one HIV clinic only, which may not be representative for South Africa. The clinic was closed in June 2012 because of restructuring of funding of HIV care.[16] However, this occurred after the end of our study period and did not affect the patient population included in our study. The number of incident cancer cases identified per cancer type was with the exception of KS too small to allow cancer specific risk factors analyses. We had no access to data of HIV-negative controls or populations and could therefore not provide comparisons with HIV-negative persons.

Our study demonstrated high incidence rates of cancer in HIV-positive people on ART with, for example, an age standardised incidence rate of 432 (95% CI 341–555) for KS and 259 (95% CI 179–390) for ICC per 100,000 pys. These incidence rates are much higher compared to the age standardised incidence rates of cancer in the Nigerian AIDS match study where the incidence rates of KS and ICC were 4.9 and 7.8 per 100,000 pys, respectively.[5] These differences may be due to poor access to diagnostic cancer services in Nigeria compared to South Africa.[25] Infection-related cancers accounted for over 80% of cancers identified in our study as compared to 67% in a cohort of HIV-positive adults in California.[26] The higher proportion of infection-related cancers in our study is expected because of the higher prevalence of communicable diseases in South Africa compared to the US.[27] It also demonstrates that ADCs are still an important cause of cancer morbidity even in the era of ART in South Africa. The cancer burden shift from ADCs to NADCs in the ART era that was seen in the US[3] and Europe,[28] is not yet evident in South Africa. However, we demonstrated a steady rise of the incidence of NADCs gaining prominence over cervix and NHL by the second year on ART (Fig. 2), suggesting that this shift may occur with time. Conjunctival cancer was common in this cohort with an incidence rate of 49 (95% CI 25–110) per 100,000 pys. This is consistent with previous studies which have shown an association between HIV and conjunctival cancer in African populations.[29–31] Over two-thirds (63%) of HIV patients with cancer were female. This is comparable with the Nigerian Cancer Match Study where 74% of cancer patients were female[5] and is expected since the HIV population in South Africa accessing ART is predominantly female.[32] In contrast, in cancer match studies from Europe and the US $\leq$25% of included patients were female.[33,34]

Most of the cancers identified in our study were not recorded in the HIV cohort patient notes. While Sinikithemba ART clinic provided HIV care, referral to another institution for either HIV or cancer care would mean gaps in patient records due to poor feedback systems and lack of a centralised database. The under-ascertainment of cancer in HIV cohorts and the high incidence of cancer that was shown in our study, highlight the need for a systematic approach to the surveillance of cancers in HIV-positive people in South Africa. Record linkage is one way of ascertaining cancers in HIV-positive people. Other potential strategies which may be explored for creating robust cancer data in the HIV population include routine collection of cancer data in HIV cohorts,[35] strengthening of HIV testing in cancer patients and HIV clinician and patient education on current knowledge on HIV-related cancers.

There are questions that remain unanswered on the epidemiology of HIV-related cancers in the South African ART era. We reported cancer incidence in patients on ART, but the additional risk of cancer in HIV-positive people on ART compared to uninfected people in South Africa is still to be determined. Further HIV-cancer linkage studies are planned in South Africa. The National Health Laboratory Services (NHLS) public laboratories serve over 80% of the South African population[14] and provide the unique opportunity to link longitudinal laboratory HIV data from the NHLS to cancer data from the NCR and thus to assess cancer risk in HIV-positive South Africans in the ART era, at national level. The large sample size and statistical power will allow analyses of uncommon NADCs, taking into account the CD4 count and HIV RNA viral load.

In conclusion, ascertainment of cancer in HIV cohorts is incomplete, and PRL is a feasible and essential tool for cancer ascertainment in HIV cohorts. There is need for a systematic approach to the surveillance of cancers in HIV-positive people in South Africa.

Acknowledgements

The authors would like to acknowledge colleagues at the Sinikithemba ART Clinic who maintain the electronic database used, the National Health Laboratory Service and National Cancer Registry. This work was funded by The International Epidemiological Databases to Evaluate AIDS in Southern Africa (IeDEA SA), Grant Number U01AI069924 from the NIH (NIAD, NICHD, NCI) (PI: Egger and Davies) and the Swiss National Science Foundation (Ambizione-PROSPER fellowship PZ00P3_160407 for J.B). Data from this study was presented at the Ninth African Organisation for Research and Training in Cancer (AORTIC) conference in November 2013.

IeDEA-SA Steering Group: Frank Tanser, Africa Centre for Health and Population Studies, University of Kwazulu-Natal, Somkhele, South Africa; Christopher Hoffmann, Aurum Institute for Health Research, Johannesburg, South Africa; Michael Vinikoor, Centre for Infectious Disease Research in Zambia, Lusaka, Zambia; Eusebio Macete, Centro de Investigação em Saúde de Manhiça, Manhiça, Mozambique; Robin Wood, Desmond Tutu HIV Centre (Gugulethu clinic), Cape Town, South Africa; Kathryn Stinson, Khayelitsha ART Programme and Médecins Sans Frontières, Cape Town, South Africa; Geoffrey Fatti, Kheth'Impilo Programme, South Africa; Sam Phiri, Lighthouse Trust Clinic, Lilongwe, Malawi; Janet Giddy, McCord Hospital, Durban, South Africa; Cleophas Chimbetete, Newlands Clinic, Harare, Zimbabwe; Kennedy Malisita, Queen Elizabeth Hospital, Blantyre, Malawi; Brian Eley, Red Cross War Memorial Children's Hospital and Department of Paediatrics and Child Health, University of Cape Town, Cape Town, South Africa; Ola Faturiyele, SolidarMed SMART Programme Lesotho; Michael Hobbins, SolidarMed SMART Programme, Pemba Region, Mozambique; Kamelia Kamenova, SolidarMed SMART Programme, Masvingo, Zimbabwe; Matthew Fox, Themba Lethu Clinic, Johannesburg, South Africa; Hans Prozesky, Tygerberg Academic Hospital, Cape Town, South Africa; Karl Technau, Empilweni Clinic, Rahima Moosa Mother and Child Hospital, Johannesburg, South Africa and Shobna Sawry, Harriet Shezi Children's Clinic, Chris Hani Baragwanath Academic Hospital, Soweto, South Africa.

References

1. IARC Working Group on the Evaluation of Carcinogenic Risks to Humans. IARC monographs on the evaluation of carcinogenic risks to humans: Volume 67 Human immunodeficiency viruses and human T-cell lymphotropic viruses. Lyon: International Agency for Research on Cancer, 1996. 183p.
2. Polesel J, Franceschi S, Suligoi B, et al. Cancer incidence in people with AIDS in Italy. *Int J Cancer* 2010; 127:1437–45.
3. Shiels MS, Pfeiffer RM, Gail MH, et al. Cancer burden in the HIV-infected population in the United States. *J Natl Cancer Inst* 2011; 103:753–62.
4. Mbulaiteye SM, Katabira ET, Wabinga H, et al. Spectrum of cancers among HIV-infected persons in Africa: the Uganda AIDS-Cancer Registry Match Study. *Int J Cancer* 2006; 118:985–90.
5. Akarolo-Anthony SN, Maso LD, Igbinoba F, et al. Cancer burden among HIV-positive persons in Nigeria: preliminary findings from the Nigerian AIDS-cancer match study. *Infect Agent Cancer* 2014; 9:1
6. Fellegi I, Sunter AB. A theory for record linkage. *J Am Stat Assoc* 1969; 64:1183–210.
7. Gill L, Goldacre M, Simmons H, et al. Computerised linking of medical records: methodological guidelines. *J Epidemiol Commun Health* 1993; 47: 316–9.
8. Clark DE, Hahn DR. Comparison of probabilistic and deterministic record linkage in the development of a statewide trauma registry. *Proc Annu Symp Comput Appl Med Care* 1995; 397–401.
9. Silveira DP, Da Artmann E. Accuracy of probabilistic record linkage applied to health databases: systematic review. *Rev Saude Publica* 2009; 43: 875–82.
10. Quantin C, Binquet C, Bourquard K, et al. Which are the best identifiers for record linkage? *Med Inform Internet Med* 2004; 29:221–7.
11. Mason CA, Tu S. Data linkage using probabilistic decision rules: a primer. *Birth Defects Res a Clin Mol Teratol* 2008; 82:812–21.
12. Blakely T, Salmond C. Probabilistic record linkage and a method to calculate the positive predictive value. *Int J Epidemiol* 2002; 31:1246–52.
13. Clarke C, Glaser SL. Population-based surveillance of HIV-associated cancers: utility of cancer registry data. *J Acquir Immune Defic Syndr* 2004; 36:1083–91.
14. Blecher M, Kollipara A, Dejager P, et al. Health financing. In: Padarath A, English R, eds. South African health review 2011. Health Systems Trust, 2011. 29–48.
15. Egger M, Ekouevi DK, Williams C, et al. Cohort Profile: the international epidemiological databases to evaluate AIDS (IeDEA) in sub-Saharan Africa. *Int J Epidemiol* 2012; 41:1256–64.
16. Cloete C, Regan S, Giddy J, et al. The linkage outcomes of a large-scale, rapid transfer of HIV-infected patients from hospital-based to community-based clinics in South Africa. *Open Forum Infect Dis* 2014; 1:ofu058–ofu058.
17. Chevrette A. G-LINK: a probabilistic record linkage system, NORC conference proceedings, May 2011. Available at: http://www.norc.org/PDFs/May 2011 Personal Validation and Entity Resolution Conference/G-Link_Probabilistic Record Linkage paper_PVERConf_May2011.pdf.
18. Statistics Canada. G-Link version 2.4 user guide, Ottawa, Statistics Canada, 2014.
19. IARC Working Group on the Evaluation of Carcinogenic Risks to Humans (2009:Lyon France). A review of human carcinogens. Part B: biological agents (IARC monographs on the evaluation of carcinogenic risks to humans, Volume 100B). Lyon: International Agency for Research on Cancer, 2012.
20. Bouvard V, Baan R, Straif K, et al. A review of human carcinogens—part B: biological agents. *Lancet Oncol* 2009; 10:321–2.
21. Hoover DR, Munoz A, Carey V, et al. Using events from dropouts in nonparametric survival function estimation with application to incubation of AIDS. *J Am Stat Assoc* 1993; 88:37–43.
22. Segi M. Cancer mortality for selected sites in 24 countries (1950–57). Department of Public Health, Tohoku University of Medicine, Sendai, 1960.
23. Victor TW, Mera RM. Record linkage of health care insurance claims. *J Am Med Inform Assoc* 2001; 8:281–8.
24. Department of Health. Regulations relating to cancer registration. National Health Act, 2003. Pretoria, South Africa, 2011.
25. Adesina A, Chumba D, Nelson AM, et al. Improvement of pathology in sub-Saharan Africa. *Lancet Oncol* 2013; 14:e152–7.
26. Silverberg MJ, Chao C, Leyden W, et al. HIV infection and the risk of cancers with and without a known infectious cause. *Aids* 2009; 23: 2337–45.
27. de Martel C, Ferlay J, Franceschi S, et al. Global burden of cancers attributable to infections in 2008: a review and synthetic analysis. *Lancet Oncol* 2012; 13:607–15.
28. Franceschi S, Lise M, Clifford GM, et al. Changing patterns of cancer incidence in the early- and late-HAART periods: the Swiss HIV Cohort Study. *Br J Cancer* 2010; 103:416–22.
29. Nagaiah G, Stotler C, Orem J, et al. Ocular surface squamous neoplasia in patients with HIV infection in sub-Saharan Africa. *Curr Opin Oncol* 2010; 22:437–42.
30. Gichuhi S, Sagoo MS, Weiss H, et al. Epidemiology of ocular surface squamous neoplasia in Africa. *Trop Med Int Health* 2013; 18:1424–43.
31. Rogena E, Simbiri KO, De Falco G, et al. A review of the pattern of AIDS defining, HIV associated neoplasms and premalignant lesions diagnosed from 2000–2011 at Kenyatta National Hospital, Kenya. *Infect Agent Cancer* 2015; 10:28
32. Cornell M, Technau K, Fairall L, et al. Monitoring the South African National Antiretroviral Treatment Programme, 2003-2007: the IeDEA Southern Africa collaboration. *South African Med J* 2009; 99:653–60.
33. Dal Maso L, Polesel J, Serraino D, et al. Pattern of cancer risk in persons with AIDS in Italy in the HAART era. *Br J Cancer* 2009; 100:840–7.
34. Shiels MS, Pfeiffer RM, Engels EA. Age at cancer diagnosis among persons with AIDS in the United States. *Ann Intern Med* 2010; 153: 452–60.
35. Semeere A, Wenger M, Busakhala N, et al. *A prospective ascertainment of cancer incidence in Saharan Africa: The Case of Kaposi sarcoma, Cancer Medicine* 2016; 1–15.