








# BMJ Open Cohort profile: the South African HIV Cancer Match (SAM) Study, a national population-based cohort

Mazvita Muchengeti <sup>1,2,3</sup> Lina Bartels,<sup>4</sup> Victor Olago <sup>1</sup>,  
Tafadzwa Dhokotera,<sup>1,5</sup> Wenlong Carl Chen <sup>1,6</sup> Adrian Spoerri,<sup>4</sup>  
Eliane Rohner <sup>4</sup> Lukas Bütikofer,<sup>7</sup> Yann Ruffieux <sup>4</sup> Elvira Singh,<sup>1,2</sup>  
Matthias Egger <sup>4,8,9</sup> Julia Bohlius <sup>4,5,10</sup>

**To cite:** Muchengeti M, Bartels L, Olago V, *et al.* Cohort profile: the South African HIV Cancer Match (SAM) Study, a national population-based cohort. *BMJ Open* 2022;**12**:e053460. doi:10.1136/bmjopen-2021-053460

► Prepublication history and additional supplemental material for this paper are available online. To view these files, please visit the journal online (<http://dx.doi.org/10.1136/bmjopen-2021-053460>).

Dr Elvira Singh, the Head of Department of the South African National Cancer Registry, passed away on 26 February 2022. We dedicate this work to her memory, in honour of her contribution to cancer surveillance and policy development in South Africa.

Received 13 May 2021

Accepted 16 March 2022



© Author(s) (or their employer(s)) 2022. Re-use permitted under CC BY-NC. No commercial re-use. See rights and permissions. Published by BMJ.

For numbered affiliations see end of article.

## Correspondence to

Dr Mazvita Muchengeti; mazvitam@nicd.ac.za

## ABSTRACT

**Purpose** The South African HIV Cancer Match (SAM) Study is a national cohort of people living with HIV (PLWH). It was created using probabilistic record linkages of routine laboratory records of PLWH retrieved by National Health Laboratory Services (NHLS) and cancer data from the National Cancer Registry. The SAM Study aims to assess the spectrum and risk of cancer in PLWH in the context of the evolving South African HIV epidemic. The SAM Study's overarching goal is to inform cancer prevention and control programmes in PLWH in the era of antiretroviral treatment in South Africa.

**Participants** PLWH (both adults and children) who accessed HIV care in public sector facilities and had HIV diagnostic or monitoring laboratory tests from NHLS.

**Findings to date** The SAM cohort currently includes 5 248 648 PLWH for the period 2004 to 2014; 69% of these are women. The median age at cohort entry was 33.0 years (IQR: 26.2–40.9). The overall cancer incidence in males and females was 235.9 (95% CI: 231.5 to 240.5) and 183.7 (181.2–186.2) per 100 000 person-years, respectively.

Using data from the SAM Study, we examined national cancer incidence in PLWH and the association of different cancers with immunodeficiency. Cancers with the highest incidence rates were Kaposi sarcoma, cervix, breast, non-Hodgkin's lymphoma and eye cancer.

**Future plans** The SAM Study is a unique, evolving resource for research and surveillance of malignancies in PLWH. The SAM Study will be regularly updated. We plan to enrich the SAM Study through record linkages with other laboratory data within the NHLS (eg, tuberculosis, diabetes and lipid profile data), mortality data and socioeconomic data to facilitate comprehensive epidemiological research of comorbidities among PLWH.

## INTRODUCTION

The International Agency for Research on Cancer defined HIV-1 infection as a carcinogen in 1996.<sup>1</sup> There is, however, limited data on HIV-related malignancies in sub-Saharan Africa, where two-thirds of the world's HIV-infected population live. Most studies on the cancer risk in people living with HIV (PLWH)

## Strengths and limitations of this study

- A central strength of the South African HIV Cancer Match Study (SAM) is its large size, which allows analyses of rare malignancies in children, adolescents and adults in South Africa.
- The SAM cohort is representative of the South African public health system (which covers over 80% of the South African population) and reflects routine HIV care.
- The use of national routine laboratory data means that loss to follow-up and silent transfer are less of a problem than in longitudinal studies of HIV treatment programmes.
- All cancers were laboratory confirmed and classified by experienced coders, however, clinically diagnosed cancers are not captured and data on lifestyle factors, HIV treatment history and mortality are not included.
- Other weaknesses relate to the limitations inherent in the secondary use of laboratory and cancer registry data, including missing data and the lack of standardised follow-up visits.

were done in North America and Europe.<sup>2–5</sup> The HIV/AIDS epidemics in these regions differ from the epidemics in sub-Saharan Africa in terms of gender and ethnicity of the affected population, the route of transmission and the availability and type of antiretroviral therapy (ART).<sup>6–8</sup> There is an urgent need for large-scale studies to examine the effect of the evolving HIV epidemic on cancer in HIV-infected Africans in the ART era. As PLWH live longer on ART, cancer has become a critical comorbidity and cause of death.<sup>9 10</sup> The high cost of cancer care and the substantial burden of HIV in sub-Saharan Africa necessitates assessments of malignancies in PLWH for national resource planning and to inform cancer control and prevention strategies.

There are several challenges when studying cancer in PLWH in sub-Saharan Africa.



African HIV cohorts generally do not or only incompletely record cancer diagnoses. For example, we linked records of adults on ART enrolled at Sinikithemba HIV clinic in Durban, South Africa, with the cancer records of public laboratories in KwaZulu-Natal province to assess the degree of under ascertainment of cancers in the HIV cohort.<sup>11</sup> After the inclusion of linkage-identified malignancies, cancer incidence increased over sixfold, with similar under ascertainment of AIDS-defining and non-AIDS-defining malignancies.<sup>11</sup> In many countries in sub-Saharan Africa, the study of HIV-related and other cancers is hampered by the absence of high-quality cancer registries or the lack of information on HIV status in cancer registries. Finally, the studies of survival of patients with cancer are limited by high rates of loss to follow-up.<sup>12</sup>

South Africa has the largest number of PLWH in the world.<sup>13</sup> A national ART programme was initiated in 2004.<sup>14</sup> Since then, there have been progressive changes in the eligibility criteria for ART and the national ART coverage,<sup>14–19</sup> with a consequent shift in the spectrum and risk of cancer in PLWH.<sup>20,21</sup> The public health system provides nationwide access to CD4 cell count and HIV RNA viral load monitoring via a network of public laboratories with centralised data warehousing.<sup>22</sup> The South African National Health Laboratory Services (NHLS) is the main diagnostic pathology service responsible for supporting the national and provincial health departments in the delivery of healthcare to over 80% of the South African population through a nationwide network of laboratories. The South African National Cancer Registry (NCR) is a division of the NHLS and South Africa's primary cancer surveillance system.<sup>23</sup> It was established in 1986 as a voluntary, pathology-based cancer registry. In 2011, the government introduced legislation making the reporting of all confirmed cancer diagnoses obligatory.<sup>23</sup> The NCR is part of the NHLS (public laboratories) and had consistent cancer reporting from government laboratories even prior to obligatory cancer reporting.

By linking NHLS and NCR data, South Africa is uniquely positioned to create a large population-based HIV cohort with cancer outcomes. Here we describe the South African HIV Cancer Match (SAM) Study, a national cohort created from laboratory records from routine HIV care and cancer registry records, to study the spectrum and incidence of cancer in South Africa, including in subgroups by sex and age (children, adolescents and young adults, and the elderly).

## COHORT DESCRIPTION

### Study setting and data sources

The NHLS comprises specialised divisions, including the National Institute for Communicable Diseases, National Institute for Occupational Health, NCR and the Antivenom Unit. The NHLS Corporate Data Warehouse (CDW) is an electronic data repository for all public sector laboratory data.<sup>24</sup> Over 260 laboratories nationally, each with its own laboratory information system, feed into

the CDW. Unique patient identifiers or national IDs are mostly not available. A previous evaluation has shown that the data held in the CDW on CD4 cell counts and HIV RNA viral load are both comprehensive and accurate.<sup>24</sup>

### Eligibility criteria, data extraction and preparation

We retrieved cancer records from the NCR and HIV-related laboratory records from the NHLS CDW for the entire country of South Africa for the period 2004–2014. We included all individuals with HIV-related lab records at two or more different time points. We defined HIV-related laboratory records indicating HIV-positivity as follows: positive HIV ELISA test, positive HIV Western blot test, positive HIV rapid test, HIV RNA viral load or CD4 count/percentage. We also extracted the date of the test (specimen registration date).

For all HIV-related laboratory records from the NHLS CDW, we retrieved the corresponding patient identifying information which we used as linkage variables: episode number, given name(s), surname(s), sex, precise or estimated date of birth, the facility providing care and the corresponding province, district and subdistrict. The episode number is a unique number for all tests done at a particular visit and sample taken. We excluded records for HIV tests with negative, missing or inconclusive test results. From the NCR, we used records from both the private and public sector facilities. We retrieved the following variables for linkages: given name(s), surname, sex, the precise date of birth or, if not available, the year of birth or patient age. We retrieved the following patient and disease-related additional information for linked cases: ethnicity, the International Classification of Diseases 10th edition (ICD-10) code, the International Classification of Diseases for Oncology third edition (ICD-O-3) code, date of test (specimen registration date), the facility where the pathological examination was done with the province, district and subdistrict. We retrieved national IDs whenever available from NHLS and NCR records for the evaluation of the deduplication and linkage.

We used Python scripts to preprocess NHLS and NCR records and harmonised linkage variables in terms of format, structure and content. We created regular expression templates to standardise titles, prepositions, special characters, hyphens, blanks, single or multiple given names and surnames, and so on. We excluded records with missing or implausible surnames or first names, and we checked dates for impossible values. We excluded all records where neither date of birth nor age was available. We checked the South African National IDs for accuracy and considered ID values that did not match the 13-digit standard length as invalid. After preprocessing, we encrypted all names with a privacy-preserving probabilistic record linkage (P3RL) encryption tool.<sup>25</sup> We generated error-tolerant codes applying hash functions using the same keyword and the same length of the bit arrays for the two datasets' encryption. Data privacy and irreversibility of the encryption were achieved with P3RL. At the same time, estimating the similarity of strings within and

between databases remained possible. We used test words to validate the encryption, dropping the original national IDs and the unencrypted names from the dataset before deduplication and linkage.

### Deduplication and linkage

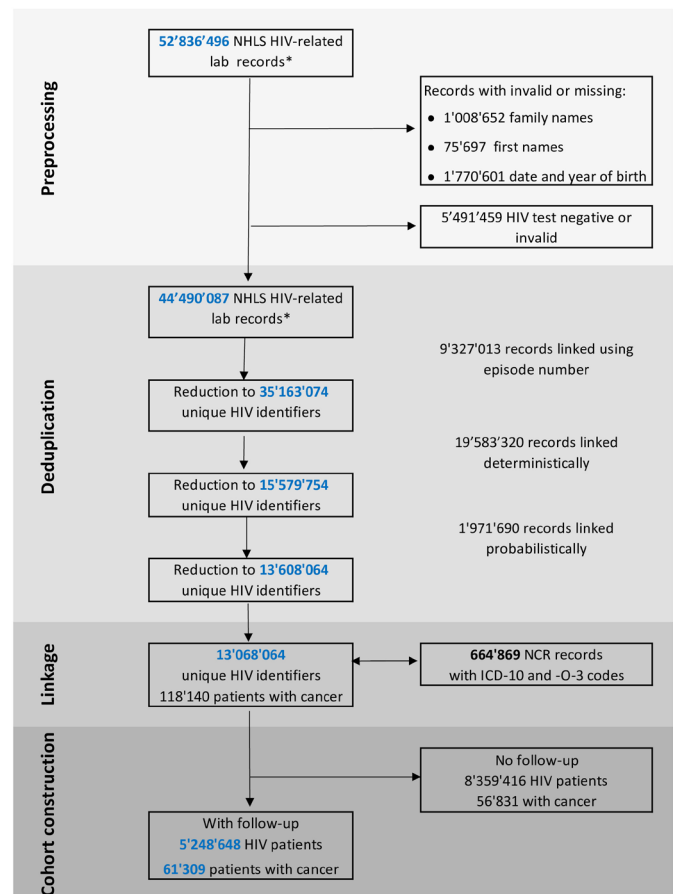
Deduplication refers to the identification of records belonging to a given patient using records from one data file. Linkage refers to identifying records pertaining to a given patient using two separate data files as the source of information. We deduplicated NHLS HIV records and linked deduplicated NHLS HIV and NCR records using G-Link (G-Link V.3.3 Rel V.5.2).<sup>26</sup> In a first step, we deduplicated the preprocessed NHLS HIV records deterministically. First, we used the episode number and then the encrypted first names and surnames, sex, date and year of birth. Second, we probabilistically deduplicated records using the encrypted first names and surnames, sex, the exact date of birth or year of birth, facility providing care and geographic location. Third, we probabilistically linked the deduplicated NHLS HIV data with the NCR dataset using the following linkage variables: encrypted first names, encrypted surnames, sex, and exact date of birth or year of birth.

### Linkage thresholds and linkage evaluation

We determined optimal linkage thresholds using the NHLS HIV and NCR records that had national IDs from the province with the largest population (Gauteng). First, we used deterministic record linkages using recoded national IDs to identify the same patient's records in the two datasets. Next, we probabilistically linked records as described above using the same set of linkage variables without the recoded national IDs. We derived precision ( $P$ ) and recall ( $R$ ) and the  $F$  measure as follows<sup>27</sup>:

1.  $P = a / (b + a)$ , the proportion of record pairs classified as matches that are true matches,
2.  $R = a / (c + a)$ , the proportion of true matching record pairs that are classified as matches, and
3.  $F = 2PR / (P + R) = 2a / (c + b + 2a)$ , the harmonic mean of  $P$  and  $R$ ; where  $a$  is a true match,  $b$  is a false match,  $c$  is a false non-match and  $d$  is a true non-match.

We assessed linkage quality at 16 different cut-offs for the total weight for the deduplication and linkage between NHLS HIV records from Gauteng province and NCR records. A total of 99 250 NCR records and 138 365 NHLS HIV records from Gauteng province had a national ID. There were 863 unique IDs between both datasets, which were used to calculate the true match pairs. For the linkages between NHLS HIV and NCR records, we set the threshold for the total weight to 0. At that threshold  $P$  was 0.98,  $R$  was 0.96 and  $F$  0.97. For deduplication of NHLS HIV records,  $P$ ,  $R$  and  $F$  were highest (0.97 each) at a threshold of 20. However, to avoid large clusters of records potentially belonging to the same person, we set the threshold to 200. At that threshold,  $P$  was 0.99,  $R$  0.89 and  $F$  0.94. We subsequently used the threshold of 200 to deduplicate the records from all provinces. We reassessed

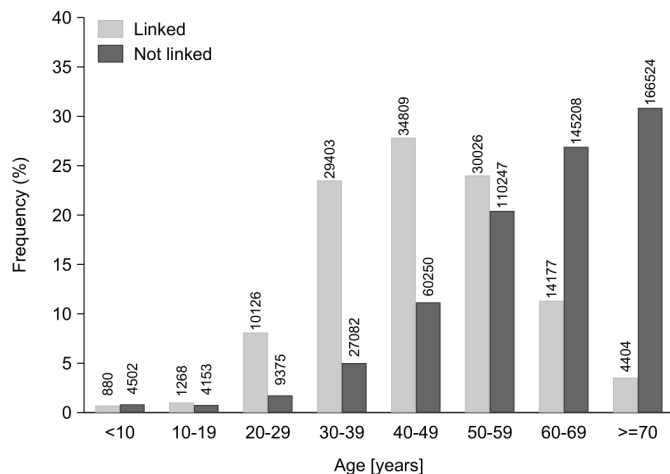


**Figure 1** Creation of the South African HIV Cancer Match Study cohort (2004–2014). \*Positive HIV tests, any CD4 cell and HIV RNA measurements. NCR, National Cancer Registry; NHLS, National Health Laboratory Services.

the linkage quality for NHLS HIV records with national IDs from all provinces ( $n=1\ 712\ 697$ ). At a threshold of 20,  $P$  was 0.93,  $R$  0.92 and  $F$  0.93. At a threshold of 200, which was used for the deduplication, the corresponding numbers were 0.99, 0.84 and 0.91.

### The creation of the SAM cohort

A total of 52 836 496 HIV-related laboratory records were extracted from the NHLS CDW from 2004 to 2014. After exclusion of records with negative or invalid HIV tests and missing or invalid names and birth dates, we retained 44 490 087 records. Deduplication resulted in 13 068 064 unique HIV identifiers (figure 1). A total of 664 869 unique cancer cases were recorded in the NCR in the same period (figure 1). From the HIV-related laboratory records, 118 140 unique HIV identifiers were matched to at least one cancer record. A link to a cancer record was more likely in males, older age and care provided in urban facilities than for females, younger age and care provided in rural areas (online supplemental table 1). Unique HIV identifiers that had HIV-related laboratory records at two or more different time points were considered as patients, whereas HIV unique identifiers with one record only or two records occurring on the same day only may represent patients who died soon after HIV



**Figure 2** Age distribution in cancer records linked and not linked to HIV records.

diagnosis, left the country or unlinked records. Out of the cancer records reported to the NCR, 125 093 (18.8%) records were linked to HIV records. Among cancers linked to HIV records, 61% occurred in females, and 73% in Black Africans, compared with 49% females and 31% Black Africans among records not linked (online supplemental table 2). Age at cancer diagnosis in records linked to an HIV record was lower (median age: 45.9 years; IQR: 36.8–55.1) than for cancer records not linked (63.1 years; IQR: 52.9–72.3).

The age distribution in cancer records linked to HIV records followed the national HIV prevalence age structure. Among cancer records not linked to HIV records, the distribution followed the age structure of incident cancer cases (figure 2). Seventy-seven percent of cancer records linked to HIV records were diagnosed in the public sector compared with 43% of cancer records not linked to an HIV record. Online supplemental table 2 shows proportions of specific cancers by aetiology in patients with cancer with and without a link to HIV records. Common cancers with links to HIV records were cervix cancer, Kaposi sarcoma and breast cancer. Common cancers in people without a link to HIV records were basal cell carcinoma, breast and prostate cancers.

### Cancer incidence analysis

We included all patients who had HIV-related laboratory records at two or more different time points (one diagnostic and monitoring test, or two monitoring tests) for the cancer incidence analysis. We calculated cancer incidence rates by dividing the number of patients who developed incident cancer by the number of person-years at risk with exact Poisson 95% CIs. Person-years at risk were measured from the first HIV-related laboratory record until the last HIV-related laboratory record plus 180 days or the date of the incident cancer diagnosis (if earlier) for each cancer of interest. PLWH who developed another cancer type other than the cancer of interest were not censored. The grace period of 180 days was chosen in accordance with previous studies.<sup>11 28</sup> Incident

cancer cases were defined as cancers diagnosed after the first HIV-related laboratory record. Incident cancer cases occurring after the last laboratory record plus 180 days were not considered in the incidence analysis.

### Patient and public involvement

The SAM Study is based on routine laboratory data and cancer registry data. No patients were involved in developing the research questions, outcome measures and the cohort's overall design. Due to the data's anonymous nature, we cannot disseminate the results of the data analyses directly to study participants. The NCR is directly involved in the development of cancer control policies in South Africa, with a representation on the Ministerial Advisory Committee on cancer. Findings from this study will be shared with policymakers.

### Findings to date

#### Characteristics of the cohort

Patients who had HIV-related lab records at two or more different time points (one diagnostic and monitoring test, or two monitoring tests) were 5 248 648 (figure 1; table 1). Women made up about two-thirds of the cohort (n=3 606 565; 69%) and the median age at the first HIV-related laboratory record was 33.0 years (IQR: 26.2–40.9). Ninety-eight percent of all PLWH in the cohort had at least one CD4 cell count measurement. The median first CD4 cell count was 290 cells/ $\mu$ L (IQR: 153–465). The median first HIV RNA viral load was 2.3 log<sub>10</sub> copies/mL (IQR: 1–4.1). The total follow-up time from first to last HIV-related laboratory record was 13 198 218 person-years; for a median follow-up time of 2.0 years (IQR: 0.6–3.8), not including the 180 days grace period. The median time between any two HIV-related laboratory records was 8.3 months (IQR: 5.1–13.3). Patients were receiving HIV care at a total of 5563 distinct facilities in 263 subdistricts and 58 districts across all nine South African provinces. About half of the PLWH (52%) were receiving care at a rural facility. Gauteng, Kwazulu-Natal, and Eastern Cape provinces had the highest numbers of PLWH, while Northern Cape had the lowest.

A total of 61 309 PLWH included in the cohort had at least one cancer (table 1), of whom 21 185 were prevalent cancers (diagnosed at or before the earliest HIV-related laboratory test), and 40 124 were incident cancers (diagnosed at any time after the earliest HIV-related laboratory test). PLWH with prevalent or incident cancers were older at entry into the cohort (42.4 years; IQR: 34.2–51.1) compared with those without cancer (32.9 years; IQR: 26.1–40.7) and tended to have lower first CD4 cell counts (242 cells/ $\mu$ L vs 290 cells/ $\mu$ L). For patient characteristics by cancer diagnosed < or >90 days before or after first HIV-related laboratory test see online supplemental table 3.

### Cancer incidence

We included 5 248 648 patients in the incidence analysis; 31 112 developed an incident cancer within the follow-up

**Table 1** Characteristics of people living with HIV included the South African HIV Cancer Match Study cohort

	Entire HIV cohort		Patients without cancer		Patients with cancer	
Total number of patients	5 248 648	100%	5 187 339	100%	61 309	100%
Sex						
Male	1 635 388	31%	1 613 730	31%	21 658	35%
Female	3 606 565	69%	3 566 948	69%	39 617	65%
Missing	6695	(<1%)	6661	(<1%)	34	(<1%)
Age at first test (years)						
Median (IQR)	33	(26.2–40.9)	32.9	(26.1–40.7)	42.4	(34.2–51.5)
<10	252 789	5%	252 322	5%	467	1%
10–19	228 771	4%	228 160	4%	611	1%
20–29	1 467 896	28%	1 460 582	28%	7314	12%
30–39	1 693 339	32%	1 675 634	32%	17 705	29%
40–49	913 647	17%	896 313	17%	17 334	28%
≥50	452 301	9%	434 794	8%	17 507	29%
Missing	239 905	5%	239 534	5%	371	1%
CD4 counts (cells/μL)						
Total number of counts	17 317 778		17 081 076		236 703	
First count						
Median (IQR)	290	(153–465)	290	(154–466)	242	(122–406)
<50	412 050	8%	406 149	8%	5901	10%
50–99	404 709	8%	398 425	8%	6284	10%
100–199	906 551	17%	893 669	17%	12 882	21%
200–349	1 364 287	26%	1 348 476	26%	15 811	26%
350–499	951 113	18%	941 674	18%	9439	15%
500–699	638 327	12%	632 567	12%	5760	9%
≥700	471 652	9%	467 553	9%	4099	7%
Missing	99 959	2%	98 826	2%	1133	2%
HIV RNA viral load (log <sub>10</sub> copies/mL)						
Total number of measurements	9 875 723		9 744 808		130 915	
First measurement						
Median (IQR)	2.3	(1–4.1)	2.3	(1–4.1)	2.6	(1–4.5)
<2.7	1 096 782	21%	1 085 019	21%	11 763	19%
2.7–3.9	384 450	7%	379 863	7%	4587	7%
4.0–4.9	432 794	8%	426 874	8%	5920	10%
≥5	461 933	9%	455 496	9%	6437	10%
Missing	2 872 689	55%	2 840 087	55%	32 602	53%
Follow-up time (years)						
Median (IQR)	2	(0.6–3.8)	2	(0.6–3.8)	2	(0.5–4.2)
Median time (IQR) between labs (months)	8.3	(5.1–13.3)	8.3	(5.2–13.3)	7.2	(3.5–12.0)
Urbanity level*						
Rural	2 743 486	52%	2 716 878	52%	26 608	43%
Urban	2 471 423	47%	2 437 116	47%	34 307	56%
Missing	33 739	1%	33 345	1%	394	1%
Province†						
Gauteng	1 296 791	25%	1 276 014	25%	20 777	34%

Continued



Table 1 Continued

	Entire HIV cohort		Patients without cancer		Patients with cancer	
Kwazulu-Natal	1 165 213	22%	1 157 314	22%	7899	13%
Eastern Cape	631 619	12%	626 675	12%	4944	8%
Mpumalanga	535 185	10%	529 006	10%	6179	10%
North West	415 603	8%	411 088	8%	4515	7%
Western Cape	393 587	7%	387 797	7%	5790	9%
Limpopo	382 674	7%	378 184	7%	4490	7%
Free State	322 519	6%	317 966	6%	4553	7%
Northern Cape	94 983	2%	92 947	2%	2036	3%
Missing	10 474	(<1%)	10 348	(<1%)	126	(<1%)

\*Location of first laboratory facility providing HIV services.

†Provincial location of first laboratory facility providing HIV services.

period, that is, after the first HIV laboratory measurement and before the last HIV-related laboratory record plus 180 days (table 2). The overall incidence to develop any cancer was 198.5 (196.3–200.7); 235.9 (231.5–240.5) in males and 183.7 (181.2–186.2) in females; total follow-up time from the first HIV-related laboratory test to 180 days after the last test was 15 674 538 person-years with a median follow-up time of 2.44 (IQR: 1.13–4.26) years. Table 2 shows the incidence rates for the 10 most frequent cancers in males and females. The most frequently diagnosed cancer in females was cervical cancer, followed by Kaposi sarcoma and breast cancer. The most commonly diagnosed cancer in males was Kaposi sarcoma, followed

by non-Hodgkin's lymphoma and prostate cancer (table 2).

#### The spectrum of cancers in patients with and without HIV

Using data on patients with cancer from the SAM cohort and a control group of HIV-negative patients with cancer from the NCR, we assessed excess cancer risk in PLWH in South Africa from 2004 to 2014.<sup>21</sup> Among patients with cancer, PLWH had higher odds of AIDS-defining cancers, namely, Kaposi sarcoma, non-Hodgkin's lymphoma and cervical cancer, than HIV-negative individuals. PLWH also had higher odds of conjunctival cancer and human papillomavirus (HPV)-related cancers, including penile, anal

Table 2 Top 10 cancers stratified by sex

Cancer type	Incident cancer cases			Cancer incidence rate per 100 000 person-years (95% CI)*		
	Males	Females	Total	Males	Females	Total
Cervical cancer	–	7433	7433	–	66.1 (64.6 to 67.6)	
Kaposi sarcoma	3268	3105	6373	72.7 (70.2 to 75.2)	27.6 (26.6 to 28.6)	40.4 (39.4 to 41.4)
Breast cancer	39	2706	2745	0.9 (0.6 to 1.2)	24.0 (23.1 to 25.0)	17.4 (16.8 to 18.1)
Non-Hodgkin's lymphoma	1202	1386	2588	26.7 (25.2 to 28.2)	12.3 (11.7 to 13.0)	16.4 (15.8 to 17.0)
Eye cancer	470	854	1324	10.4 (9.5 to 11.4)	7.6 (7.1 to 8.1)	8.4 (7.9 to 8.9)
BCC	647	450	1097	14.4 (13.3 to 15.5)	4.0 (3.6 to 4.4)	7.0 (6.5 to 7.4)
SCC of skin	454	432	886	10.1 (9.2 to 11.1)	3.8 (3.5 to 4.2)	5.6 (5.2 to 6.0)
Lung cancer	526	206	732	11.7 (10.7 to 12.7)	1.8 (1.6 to 2.1)	4.6 (4.3 to 5.0)
Prostate cancer	660	–	660	14.7 (13.6 to 15.8)	–	
Colorectal cancer	291	363	654	6.5 (5.7 to 7.2)	3.2 (2.9 to 3.6)	4.1 (3.8 to 4.5)
Primary site unknown	694	728	1422	15.4 (14.3 to 16.6)	6.5 (6.0 to 6.9)	9.0 (8.5 to 9.5)
Ill defined	7	4	11	0.2 (0.1 to 0.3)	0.0 (0.0 to 0.1)	0.1 (0.0 to 0.1)
Total†	10 542	20 570	31 112	235.9 (231.5 to 240.5)	183.7 (181.2 to 186.2)	198.5 (196.3 to 200.7)
Total excluding BCC and SCC of skin	9604	19 796	29 400	214.7 (210.5 to 219.1)	176.7 (174.3 to 179.2)	187.5 (185.3 to 189.6)

\*Including incident cancer cases that occurred up to 180 days after last HIV-related laboratory record.

†Including all cancers.

BCC, basal cell carcinoma; SCC, squamous cell carcinoma.

and vulvar cancer. Squamous cell carcinoma of the skin was also confirmed to be HIV associated.<sup>21</sup>

### Immunodeficiency and cancer in PLWH

We examined CD4 trajectories and cancer risk in 3.5 million SAM Study participants with at least two CD4 counts and 1 year of follow-up.<sup>29</sup> When assuming a linear relationship between time-updated CD4 cell counts and the log-hazard, the association between low CD4 cell count and higher rates of cancer was most substantial in conjunctival cancer (adjusted HR (aHR) per decrease of 100 CD4 cells/ $\mu$ L: 1.46; 95% CI: 1.38–1.54), followed by Kaposi sarcoma (aHR: 1.23, 95% CI: 1.20 to 1.26) and non-Hodgkin's lymphoma (aHR: 1.18; 95% CI: 1.14 to 1.22). Among the infection-unrelated cancers, we found low CD4 cell count to be associated with higher rates of oesophageal cancer (aHR: 1.06; 95% CI 1.00 to 1.11), but not with higher rates of lung, breast or prostate cancer.<sup>29</sup>

### Strengths and limitations

The SAM Study has over 5 million participants and is one of the largest cohorts of PLWH worldwide. This cohort allows for analyses of AIDS-defining and many non-AIDS-defining cancers, including rare malignancies in children, adolescents and adults in South Africa. All cancers were laboratory confirmed and classified according to the ICD-O-3<sup>30</sup> by experienced coders.

The SAM cohort is representative of the South African public health system (which covers over 80% of the South African population) and reflects routine HIV care. The use of national routine laboratory data means that loss to follow-up and silent transfers (when a patient switches clinics without informing the clinic where they were accessing care) are less of a problem than in longitudinal studies of HIV treatment programmes. HIV-related tests done at any public sector clinic are recorded by the NHLS CDW and can be linked to the patient concerned.

The South African NCR is a pathology-based cancer surveillance system, and clinically diagnosed cancers are not captured. This means that particularly for cancers with low biopsy rates (such as cancers of the liver, oesophagus and pancreas), cancer incidence will be underestimated. Some cancer cases in PLWH may not have been linked to the HIV cohort due to data entry errors leading to further underestimation of cancer incidence. Furthermore, data on lifestyle factors, HIV treatment history and mortality are not included. Other weaknesses relate to the limitations inherent in the secondary use of laboratory and cancer registry data, including missing data and the lack of standardised follow-up visits.

### Future plans

The SAM Study will be regularly updated. Additional data on screening tests for precancerous cervical lesions, tuberculosis, ART resistance, diabetes, lipid profiles and other comorbidities will facilitate research on comorbidities. Data from external sources, such as socioeconomic status by geographic area, will further enrich the cohort.<sup>31</sup>

At present, incident cancer cases in the SAM Study are cancers occurring after the first HIV laboratory record. We will evaluate whether the date of HIV infection based on the first CD4 count recorded<sup>32</sup> or the last negative and first positive HIV test can be imputed. The privacy-preserving methods we have used for the construction of this cohort allow linkages without compromising patient privacy. Ultimately, this will enable linkages with the National Death Registry in South Africa to obtain vital status to assess cancer-related mortality. The cohort's size and richness of the data will allow us to examine cancer incidence and mortality and associated risk factors for a wide spectrum of different cancers stratified by variables of interest, including comorbidities or area-based socioeconomic position. This national population-based cohort will monitor cancer incidence and mortality in PLWH in South Africa, thus contributing to public health surveillance.

### CONCLUSIONS

The SAM Study is one of the largest HIV and cancer cohorts worldwide, allowing surveillance and research of malignancies in PLWH as the South African and global HIV epidemics evolve.

#### Author affiliations

<sup>1</sup>National Cancer Registry, National Health Laboratory Service, Johannesburg, South Africa

<sup>2</sup>School of Public Health, Faculty of Health Sciences, University of the Witwatersrand, Johannesburg, South Africa

<sup>3</sup>South African DSI-NRF Centre of Excellence in Epidemiological Modelling and Analysis, Stellenbosch University, Stellenbosch, South Africa

<sup>4</sup>Institute of Social and Preventive Medicine, University of Bern, Bern, Switzerland

<sup>5</sup>Swiss Tropical and Public Health Institute, Allschwil, Switzerland

<sup>6</sup>Sydney Brenner Institute for Molecular Bioscience, Faculty of Health Sciences, University of the Witwatersrand, Johannesburg, South Africa

<sup>7</sup>CTU Bern, University of Bern, Bern, Switzerland

<sup>8</sup>Centre for Infectious Disease Epidemiology and Research (CIDER), School of Public Health and Family Medicine, University of Cape Town, Cape Town, South Africa

<sup>9</sup>Population Health Sciences, Bristol Medical School, University of Bristol, Bristol, UK

<sup>10</sup>University of Basel, Basel, Switzerland

**Twitter** Victor Olago @VOlago and Matthias Egger @eggernsnf

**Contributors** ME, ES and JB are co-PIs of the study and were involved in the study's conception, supervision and obtaining funding. LinaB, VO, WCC and AS were involved in data cleaning and conducted record linkages. LinaB, LukasB and YR performed data analyses. LinaB, MM and JB wrote the first draft of the manuscript. TD prepared ethics application documents. ER and all authors contributed to revising or commenting of the paper and approved the final version. MM is the guarantor of this article.

**Funding** The research reported in this publication was supported by the National Cancer Institute (NCI) and by the National Institute of Allergy and Infectious Diseases of the National Institutes of Health under Award Number U01AI069924. The SAM study was supported by a National Institutes of Health (NIH) administrative supplement to Existing NIH Grants and Cooperative Agreements (Parent Admin Supp) (grant U01AI069924-09, to ME and JB), a PEPFAR supplement (to ME), the Swiss National Science Foundation (grant 320030\_169967, to JB, ME) and the U.S. CRDF Global (HIV\_DAA3-16-62705-1, to MM). ME was supported by special project funding (grant 189498) from the Swiss National Science Foundation. The contents are solely the authors' responsibility and do not necessarily reflect the funding bodies' views.

**Competing interests** None declared.

**Patient and public involvement** Patients and/or the public were not involved in the design, or conduct, or reporting or dissemination plans of this research.

**Patient consent for publication** Not applicable.

**Ethics approval** The study received ethical approval from the Human Research Ethics Committee of the University of the Witwatersrand, Johannesburg (ID M190594), and the Cantonal Ethics committee in Bern (ID 2016--00589). Data pre-processing with full access to patient names was done by NCR staff at the premises of the NCR in Johannesburg. Encrypted data were sent to the Institute of Social and Preventive Medicine (ISPM), University of Bern. ISPM staff used privacy-preserving probabilistic record linkage methods to deduplicate and link records. Encrypted names and recoded IDs were removed before analyses.

**Provenance and peer review** Not commissioned; externally peer reviewed.

**Data availability statement** Data are available upon reasonable request. The SAM data centre is housed at the NHLS in Johannesburg, South Africa. To encourage open scientific inquiry, the advancement of science and efficient use of this resource, this database is open to internal and external investigators wishing to explore new analyses subject to the following conditions. (1) Investigators are required to submit a concept sheet describing details of the proposed analyses, which is subject to approval by the SAM Study scientific committee comprising representatives from the NCR, the Swiss Tropical and Public Health Institute, Basel, Switzerland, and the University of Bern, Switzerland. (2) All data provided will be deidentified. (3) All investigators are required to sign a data-sharing agreement committing to (i) using the data only for research purposes; (ii) to secure the data using appropriate technology; (iii) to destroy the data after analyses are completed; (iv) not distributing data to third parties. Publications arising from these data analyses will be subject to final approval by the SAM Study scientific committee. (4) No papers or presentations using the provided data may be published or made without the SAM scientific committee's written consent. SAM Study investigators or data centre staff should be coauthors on publications arising from these data provided they fulfil the authorship criteria defined by the International Committee of Medical Journal Editors.

**Supplemental material** This content has been supplied by the author(s). It has not been vetted by BMJ Publishing Group Limited (BMJ) and may not have been peer-reviewed. Any opinions or recommendations discussed are solely those of the author(s) and are not endorsed by BMJ. BMJ disclaims all liability and responsibility arising from any reliance placed on the content. Where the content includes any translated material, BMJ does not warrant the accuracy and reliability of the translations (including but not limited to local regulations, clinical guidelines, terminology, drug names and drug dosages), and is not responsible for any error and/or omissions arising from translation and adaptation or otherwise.

**Open access** This is an open access article distributed in accordance with the Creative Commons Attribution Non Commercial (CC BY-NC 4.0) license, which permits others to distribute, remix, adapt, build upon this work non-commercially, and license their derivative works on different terms, provided the original work is properly cited, appropriate credit is given, any changes made indicated, and the use is non-commercial. See: <http://creativecommons.org/licenses/by-nc/4.0/>.

#### ORCID iDs

Mazvita Muchengeti <http://orcid.org/0000-0002-1955-923X>

Victor Olago <http://orcid.org/0000-0002-0154-0688>

Wenlong Carl Chen <http://orcid.org/0000-0002-3248-4906>

Eliane Rohner <http://orcid.org/0000-0002-0554-2875>

Yann Ruffieux <http://orcid.org/0000-0002-0891-2448>

Matthias Egger <http://orcid.org/0000-0001-7462-5132>

Julia Bohlius <http://orcid.org/0000-0003-1955-1585>

#### REFERENCES

- 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, 1996. <http://monographs.iarc.fr/ENG/Monographs/vol67/mono67.pdf>
- Hleyhel M, Belot A, Bouvier AM, *et al*. Risk of AIDS-defining cancers among HIV-1-infected patients in France between 1992 and 2009: results from the FHDH-ANRS CO4 cohort. *Clin Infect Dis* 2013;57:1638–47.
- Engels EA, Biggar RJ, Hall HI, *et al*. Cancer risk in people infected with human immunodeficiency virus in the United States. *Int J Cancer* 2008;123:187–94.
- 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.
- Clifford GM, Polesel J, Rickenbach M, *et al*. Cancer risk in the Swiss HIV cohort study: associations with immunodeficiency, smoking, and highly active antiretroviral therapy. *J Natl Cancer Inst* 2005;97:425–32.
- 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.
- Mary-Krause M, Grabar S, Lièvre L, *et al*. Cohort profile: French Hospital database on HIV (FHDH-ANRS CO4). *Int J Epidemiol* 2014;43:1425–36.
- Cornell M, Technau K, Fairall L, *et al*. Monitoring the South African national antiretroviral treatment programme, 2003–2007: the IeDEA southern Africa collaboration. *S Afr Med J* 2009;99:653–60.
- Vandenhende M-A, Roussillon C, Henard S, *et al*. Cancer-Related causes of death among HIV-infected patients in France in 2010: evolution since 2000. *PLoS One* 2015;10:e0129550.
- Simard EP, Engels EA, *et al*. Cancer as a cause of death among people with AIDS in the United States. *Clin Infect Dis* 2010;51:957–62.
- Sengayi M, Spoerri A, Egger M, *et al*. Record linkage to correct under-ascertainment of cancers in HIV cohorts: the Sinikithemba HIV clinic linkage project. *Int. J. Cancer* 2016;139:1209–16.
- Freeman E, Semeere A, Wenger M, *et al*. Pitfalls of practicing cancer epidemiology in resource-limited settings: the case of survival and loss to follow-up after a diagnosis of Kaposi's sarcoma in five countries across sub-Saharan Africa. *BMC Cancer* 2016;16:1–7.
- Statistics South Africa. Mid-year population estimates 2019, 2020. Pretoria. Available: <http://www.statssa.gov.za/publications/P0302/P03022019.pdf>
- Johnson LF. Access to antiretroviral treatment in South Africa, 2004–2011. *South Afr J HIV Med* 2012;13:22–7.
- Adam MA, Johnson LF. Estimation of adult antiretroviral treatment coverage in South Africa. *S Afr Med J* 2009;99:661–7.
- Johnson LF, Dorrington RE, Moolla H. Progress towards the 2020 targets for HIV diagnosis and antiretroviral treatment in South Africa. *South Afr J HIV Med* 2017;18:1–8.
- National Department of Health. National antiretroviral treatment guidelines, 2004. Pretoria. Available: [http://www.hst.org.za/sites/default/files/sa\\_ART\\_Guidelines1.pdf](http://www.hst.org.za/sites/default/files/sa_ART_Guidelines1.pdf)
- Plazy M, Dabis F, Naidu K, *et al*. Change of treatment guidelines and evolution of art initiation in rural South Africa: data of a large HIV care and treatment programme. *BMC Infect Dis* 2015;15:452.
- National department of health. The South African antiretroviral treatment guidelines, 2013. Pretoria. Available: [http://www.kznhealth.gov.za/medicine/2013\\_art\\_guidelines.pdf](http://www.kznhealth.gov.za/medicine/2013_art_guidelines.pdf)
- Stein L, Urban MI, O'Connell D, *et al*. The spectrum of human immunodeficiency virus-associated cancers in a South African black population: results from a case-control study, 1995–2004. *Int J Cancer* 2008;122:2260–5.
- Dhokotera T, Bohlius J, Spoerri A, *et al*. The burden of cancers associated with HIV in the South African public health sector, 2004–2014: a record linkage study. *Infect Agent Cancer* 2019;14:12.
- Sherman GG, Lilian RR, Bhardwaj S, *et al*. Laboratory information system data demonstrate successful implementation of the prevention of mother-to-child transmission programme in South Africa. *S Afr Med J* 2014;104:235–8.
- Singh E, Ruff P, Babb C, *et al*. Establishment of a cancer surveillance programme: the South African experience. *Lancet Oncol* 2015;16:e414–21.
- Bassett IV, Huang M, Cloete C, *et al*. Assessing the completeness and accuracy of South African national laboratory CD4 and viral load data: a cross-sectional study. *BMJ Open* 2018;8:e021506–7.
- Schmidlin K, Clough-Gorr KM, Spoerri A, *et al*. Privacy preserving probabilistic record linkage (P3RL): a novel method for linking existing health-related data and maintaining participant confidentiality. *BMC Med Res Methodol* 2015;15:46.
- Chevrette A. G-LINK : A Probabilistic Record Linkage System, 2011. Available: [http://www.norc.org/PDFs/May2011PersonalValidationandEntityResolutionConference/G-Link\\_ProbabilisticRecordLinkagepaper\\_PVERConf\\_May2011.pdf](http://www.norc.org/PDFs/May2011PersonalValidationandEntityResolutionConference/G-Link_ProbabilisticRecordLinkagepaper_PVERConf_May2011.pdf)
- Christen P. *Data Matching : Concepts and techniques for record linkage, entity resolution and duplicate detection*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2012.
- Hoover DR. Using events from dropouts in nonparametric survival function estimation with application to incubation of AIDS. *J Am Stat Assoc* 1993;88:37–43.
- Ruffieux Y, Muchengeti M, Egger M, *et al*. Immunodeficiency and cancer in 3.5 million people living with human immunodeficiency



- virus (HIV): the South African HIV cancer match study. *Clin Infect Dis* 2021;73:e735–44 <https://pubmed.ncbi.nlm.nih.gov/33530095/>
- 30 Fritz A, Percy C, Jack A, eds. *International Classification of Diseases for Oncology*. 3 ed. Geneva: World Health Organization, 2000.
- 31 Noble M, Wright G. Using indicators of multiple deprivation to demonstrate the spatial legacy of apartheid in South Africa. *Soc Indic Res* 2013;112:187–201.
- 32 Taffé P, May M, Swiss HIV Cohort Study. A joint back calculation model for the imputation of the date of HIV infection in a prevalent cohort. *Stat Med* 2008;27:4835–53.