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

Mazvita Muchengeti, Lina Bartels, Victor Olago, Tafadzwa Dhokotera, Wenlong Carl Chen, Adrian Spoerri, Eliane Rohner, Lukas Bütkofer, Yann Ruffieux, Elvira Singh, Matthias Egger, Julia Bohlius

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 5248648 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 1996.1 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) were done in North America and Europe.2–5 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).6–9 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.9–10 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 Sipituphethu 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. After the inclusion of linkage-identified malignancies, cancer incidence increased over sixfold, with similar under ascertainment of AIDS-defining and non-AIDS-defining malignancies. 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.

South Africa has the largest number of PLWH in the world. A national ART programme was initiated in 2004. Since then, there have been progressive changes in the eligibility criteria for ART and the national ART coverage, with a consequent shift in the spectrum and risk of cancer in PLWH. 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. 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. 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. 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. 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.

**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 labs 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(s), 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. 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). 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:

1. \( P = \frac{a}{b + a} \), the proportion of record pairs classified as matches that are true matches,
2. \( R = \frac{a}{c + a} \), the proportion of true matching record pairs that are classified as matches, and
3. \( F = \frac{2PR}{P + R} = \frac{2a}{\frac{a}{b} + \frac{a}{c} + 2a} \), 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 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 130 686 064 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
muchengeti m, et al. bmj open 2022;12:e053460. doi:10.1136/bmjopen-2021-053460

chosen in accordance with previous studies. incident were not censored. the grace period of 180 days was another cancer type other than the cancer of interest earlier) for each cancer of interest. plwh who developed incident cancer by the number of person-years at risk (242 cells/μL vs 290 cells/μL). for patient characteristics by cancer diagnosed < or >90 days before or after first hiv-related laboratory record see online supplemental table 3.

cancer incidence

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 5248648 (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/μl (iqr: 153–465). the median first hiv rna viral load was 2.3 log10 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/μl vs 290 cells/μl). 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
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.

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% cts. 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. incident 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.
| 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 | 5248648 | 100% | 5187339 | 100% | 61309 | 100% |
| Sex | | | | | | |
| Male | 1635388 | 31% | 1613730 | 31% | 21658 | 35% |
| Female | 3606565 | 69% | 3566948 | 69% | 39617 | 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 | 252789 | 5% | 252322 | 5% | 467 | 1% |
| 10–19 | 228771 | 4% | 228160 | 4% | 611 | 1% |
| 20–29 | 1467896 | 28% | 1460582 | 28% | 7314 | 12% |
| 30–39 | 1693339 | 32% | 1675634 | 32% | 17705 | 29% |
| 40–49 | 913647 | 17% | 896313 | 17% | 17334 | 28% |
| ≥50 | 452301 | 9% | 434794 | 8% | 17507 | 29% |
| Missing | 239905 | 5% | 239534 | 5% | 371 | 1% |
| CD4 counts (cells/µL) | | | | | | |
| Total number of counts | 17317778 | 100% | 17081076 | 100% | 236703 | 100% |
| First count | | | | | | |
| Median (IQR) | 290 | (153–465) | 290 | (154–466) | 242 | (122–406) |
| <50 | 412050 | 8% | 406149 | 8% | 5901 | 10% |
| 50–99 | 404709 | 8% | 398425 | 8% | 6284 | 10% |
| 100–199 | 906551 | 17% | 893669 | 17% | 12882 | 21% |
| 200–349 | 1364287 | 26% | 1348476 | 26% | 15811 | 26% |
| 350–499 | 951113 | 18% | 941674 | 18% | 9439 | 15% |
| 500–699 | 638327 | 12% | 632567 | 12% | 5760 | 9% |
| ≥700 | 471652 | 9% | 467553 | 9% | 4099 | 7% |
| Missing | 99959 | 2% | 98826 | 2% | 1133 | 2% |
| HIV RNA viral load (log10 copies/mL) | | | | | | |
| Total number of measurements | 9875723 | 100% | 9744808 | 100% | 130915 | 100% |
| First measurement | | | | | | |
| Median (IQR) | 2.3 | (1–4.1) | 2.3 | (1–4.1) | 2.6 | (1–4.5) |
| <2.7 | 1096782 | 21% | 1085019 | 21% | 11763 | 19% |
| 2.7–3.9 | 384450 | 7% | 379863 | 7% | 4587 | 7% |
| 4.0–4.9 | 432794 | 8% | 426874 | 8% | 5920 | 10% |
| ≥5 | 461933 | 9% | 455496 | 9% | 6437 | 10% |
| Missing | 2872689 | 55% | 2840087 | 55% | 32602 | 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 | 2743486 | 52% | 2716878 | 52% | 26608 | 43% |
| Urban | 2471423 | 47% | 2437116 | 47% | 34307 | 56% |
| Missing | 33739 | 1% | 33345 | 1% | 394 | 1% |
| Province† | | | | | | |
| Gauteng | 1296791 | 25% | 1276014 | 25% | 20777 | 34% |

Continued
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 15674538 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. 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

| Cancer type                          | Entire HIV cohort | Patients without cancer | Patients with cancer |
|--------------------------------------|-------------------|------------------------|----------------------|
| Kwazulu-Natal                        | 1165213           | 1157314                | 7899                 |
| Eastern Cape                         | 631619            | 626675                 | 4944                 |
| Mpumalanga                           | 535185            | 529006                 | 6179                 |
| North West                           | 415603            | 411088                 | 4515                 |
| Western Cape                         | 393587            | 387797                 | 5790                 |
| Limpopo                              | 382674            | 378184                 | 4490                 |
| Free State                           | 322519            | 317966                 | 4553                 |
| Northern Cape                        | 94983             | 92947                  | 2036                 |
| Missing                              | 10474             | (1%)                   | 10348                |

*Location of first laboratory facility providing HIV services.
†Provincial location of first laboratory facility providing HIV services.

Table 2: Top 10 cancers stratified by sex

| Cancer type                          | Incident cancer cases | Cancer incidence rate per 100 000 person-years (95% CI)* |
|--------------------------------------|-----------------------|----------------------------------------------------------|
| Cervical cancer                      | 7433                  | 66.1 (64.6 to 67.6)                                       |
| Kaposi sarcoma                       | 6373                  | 72.7 (70.2 to 75.2)                                       |
| Breast cancer                        | 2745                  | 24.0 (23.1 to 25.0)                                       |
| Non-Hodgkin’s lymphoma               | 2588                  | 12.3 (11.7 to 13.0)                                       |
| Eye cancer                           | 1324                  | 10.4 (9.5 to 11.4)                                        |
| BCC                                  | 1097                  | 14.4 (13.3 to 15.5)                                       |
| SCC of skin                          | 886                   | 10.1 (9.2 to 11.1)                                        |
| Lung cancer                          | 732                   | 11.7 (10.7 to 12.7)                                       |
| Prostate cancer                      | 660                   | 14.7 (13.6 to 15.8)                                       |
| Colorectal cancer                    | 654                   | 6.5 (5.7 to 7.2)                                          |
| Primary site unknown                 | 1422                  | 15.4 (14.3 to 16.6)                                       |
| Ill defined                          | 11                    | 0.2 (0.1 to 0.3)                                          |
| Total†                               | 31112                 | 235.9 (231.5 to 240.5)                                    |
| Total excluding BCC and SCC of skin  | 29400                 | 214.7 (210.5 to 219.1)                                    |

*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.21

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.20 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/μL: 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.29

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 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 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.31 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 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
1National Cancer Registry, National Health Laboratory Service, Johannesburg, South Africa
2School of Public Health, Faculty of Health Sciences, University of the Witwatersrand, Johannesburg, South Africa
3South African DSI-NRF Centre of Excellence in Epidemiological Modelling and Analysis, Stellenbosch University, Stellenbosch, South Africa
4Institute of Social and Preventive Medicine, University of Bern, Bern, Switzerland
5Swiss Tropical and Public Health Institute, Allschwil, Switzerland
6Sydney Brenner Institute for Molecular Bioscience, Faculty of Health Sciences, University of the Witwatersrand, Johannesburg, South Africa
7CTU Bern, University of Bern, Bern, Switzerland
8Centre for Infectious Disease Epidemiology and Research (CIDER), School of Public Health and Family Medicine, University of Cape Town, Cape Town, South Africa
9Population Health Sciences, Bristol Medical School, University of Bristol, Bristol, UK
10University of Basel, Basel, Switzerland

Twitter Victor Olago @VOlago and Matthias Egger @eggersnsf

Contributors ME, ES and JB are co-PIs of the study and were involved in the study’s conception, supervision and obtaining funding. LinaB, VD, 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_DA31-16-62705-1, 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, 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 recorded 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 fulfill 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 Oлага 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

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, 1996. http://monographs.iarc.fr/ENG/Monographs/vol67/book67.pdf
2. Hieyhel 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.
3. 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.
4. Shills 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.
5. Clifford GM, Polese 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.
6. 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.
7. 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.
8. Cornell M, Technau K, Fairall L, et al. Monitoring the South African national antiretroviral treatment programme, 2003-2007: the leDSEA southern Africa collaboration. S Afr Med J 2009;99:653–60.
9. 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.
10. 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.
11. Sengayi M, Spoerri A, Egger M, et al. Record linkage to correct under-ascertaining of cancers in HIV cohorts: the SInkithemba HIV clinical linkage project. J Clin Epidemiol 2014;67:1035–41.
12. 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.
13. Statistics South Africa. Cancer in South Africa, 2015-2020. Pretoria. Available: http://www.statssa.gov.za/publications/P0302/P03022019.pdf
14. Johnson LF. Access to antiretroviral treatment in South Africa, 2004 - 2011. South Afr J HIV Med 2012;13:22–7.
15. Adam MA, Johnson LF, et al. Estimation of adult antiretroviral treatment coverage in South Africa. S Afr Med J 2009;99:661–7.
16. 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.
17. National Department of Health. National antiretroviral treatment guidelines, 2004. Pretoria. Available: http://www.hst.org.za/sites/default/files/SA_ART_Guidelines1.pdf
18. Plazzy M, Dabis F, Naidu K, et al. Change of treatment guidelines and evolution ofart initiation in rural South Africa: data of a large HIV care and treatment programme. BMC Infect Dis 2015;15:452.
19. National department of health. The South African antiretroviral treatment guidelines, 2013. Pretoria. Available: http://www.kznhealth.gov.za/medicine/2013_art_guidelines.pdf
20. Stein L, Urban M, 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.
21. 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.
22. 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 J Med 2014;104:235–8.
23. Singh E, Ruf P, Babb C, et al. Establishment of a cancer surveillance programme: the South African experience. Lancet Oncol 2015;16:e41–21.
24. 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.
25. Schimidin 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.
26. Chevreau A, G-LINK : A Probabilistic Record Linkage System, 2011. Available: http://www.norc.org/PDFs/May 2011 Personal Validation and Entity Resolution Conference/G-Link_Probabilistic Record Linkage paper_PVERConf_May2011.pdf
27. Christen P. Data Matching: Concepts and techniques for record linkage, entity resolution and duplicate detection. Berlin, Heidelberg: Springer Berlin Heidelberg, 2012.
28. Hoover DR. Using events from dropouts in nonparametric survival function estimation with application to incubation of AIDS. J Am Stat Assoc 1993;88:37–46.
29. 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 Taillé 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.
## Supplement Table 1: Characteristics of unique PLHW included in the SAM study, linked or unlinked to NCR cancer records

|                              | Unique PLWH | Not linked to NCR cancer records | Linked to NCR cancer records |
|------------------------------|-------------|----------------------------------|-----------------------------|
| **Total number of PLWH**     | 13,608,064  | 13,489,924                       | 118,140                     |
| **Sex**                      |             |                                  |                             |
| Male                         | 4,404,996 (32%) | 4,360,522 (32%)                  | 44,474 (38%)               |
| Female                       | 9,034,098 (66%) | 8,960,975 (66%)                  | 73,123 (62%)               |
| Missing                      | 168,970 (1%)  | 168,427 (1%)                     | 543 (0%)                   |
| **Age at first test [years]**|             |                                  |                             |
| Median [IQR]                 | 33.3 [26.3 - 41.6] | 33.2 [26.2 - 41.4]              | 44.9 [35.8 - 54.3]         |
| <10                          | 631,588 (5%)  | 630,757 (5%)                     | 831 (1%)                   |
| 10-19                        | 571,413 (4%)  | 570,284 (4%)                     | 1,129 (1%)                 |
| 20-29                        | 3,441,071 (25%) | 3,429,400 (25%)                 | 11,671 (10%)               |
| 30-39                        | 4,003,243 (29%) | 3,974,206 (29%)                 | 29,037 (25%)               |
| 40-49                        | 2,233,976 (16%) | 2,202,125 (16%)                 | 31,851 (27%)               |
| 50-59                        | 953,800 (7%)  | 927,252 (7%)                     | 26,548 (22%)               |
| 60-69                        | 247,504 (2%)  | 235,673 (2%)                     | 11,831 (10%)               |
| >=70                         | 70,925 (1%)   | 67,179 (0%)                      | 3,746 (3%)                 |
| Missing                      | 1,454,544 (11%) | 1,453,048 (11%)                 | 1,496 (1%)                 |
| **Follow-up**                |             |                                  |                             |
| Link with NHLS HIV-related lab records | 7,508,132 (55%) | 7,389,992 (55%)                  | 118,140 (100%)            |
| Any follow-up                | 5,248,648 (39%) | 5,187,339 (38%)                  | 61,309 (52%)               |
| **Urbanity level**           |             |                                  |                             |
| Rural                        | 7,301,863 (54%) | 7,249,683 (54%)                  | 52,180 (44%)               |
| Urban                        | 6,219,411 (46%) | 6,154,205 (46%)                  | 65,206 (55%)               |
| Missing                      | 86,790 (1%)   | 86,036 (1%)                      | 754 (1%)                   |
| **Province**                 |             |                                  |                             |
| Gauteng                      | 3,573,178 (26%) | 3,531,814 (26%)                  | 41,364 (35%)               |
| KwaZulu-Natal                | 2,687,135 (20%) | 2,672,662 (20%)                  | 14,473 (12%)               |
| Eastern Cape                 | 1,849,763 (14%) | 1,840,192 (14%)                  | 9,571 (8%)                 |
| Province     | NCHC 2017 | NCHC 2016 | Diff 2017-16 |
|--------------|-----------|-----------|--------------|
| Mpumalanga   | 1,426,683 (10%) | 1,414,867 (10%) | 11,816 (10%) |
| Limpopo      | 1,156,355 (8%)    | 1,145,857 (8%)    | 10,498 (9%)  |
| North West   | 1,074,749 (8%)    | 1,066,501 (8%)    | 8,248 (7%)   |
| Free State   | 858,306 (6%)      | 848,687 (6%)      | 9,619 (8%)   |
| Western Cape | 688,504 (5%)      | 680,223 (5%)      | 8,281 (7%)   |
| Northern Cape| 238,293 (2%)      | 234,462 (2%)      | 3,831 (3%)   |
| Missing      | 55,098 (0%)       | 54,659 (0%)       | 439 (0%)     |

* Location of first laboratory facility providing HIV services, ** provincial location of first laboratory facility providing HIV services; PLWH people living with HIV, SAM South African HIV Cancer Match Study, NCR National Cancer Registry, IQR interquartile range, NHLS National Health Laboratory Services,
|                  | Total NCR cancer records | NCR cancer records not linked to HIV records | NCR cancer records linked to HIV records |
|------------------|--------------------------|---------------------------------------------|------------------------------------------|
| **Total**        | 664,869                  | 539,776                                     | 125,093                                  |
| **Sex**          |                          |                                             |                                          |
| Male             | 324,391 (49%)            | 276,106 (51%)                               | 48,285 (39%)                             |
| Female           | 339,757 (51%)            | 262,992 (49%)                               | 76,765 (61%)                             |
| Missing          | 721 (0%)                 | 678 (0%)                                    | 43 (0%)                                  |
| **Ethnicity**    |                          |                                             |                                          |
| Black            | 259,711 (39%)            | 168,991 (31%)                               | 90,720 (73%)                             |
| Coloured (mixed race) | 73,467 (11%)   | 63,590 (12%)                               | 9,877 (8%)                               |
| White            | 286,425 (43%)            | 266,305 (49%)                               | 20,120 (16%)                             |
| Asian            | 17,246 (3%)              | 16,316 (3%)                                 | 930 (1%)                                 |
| Unknown          | 28,020 (4%)              | 24,574 (5%)                                 | 3,446 (3%)                               |
| **Age at cancer diagnosis [years]** |                     |                                             |                                          |
| Median [IQR]     | 60.0 [48.0-70.2]         | 63.1 [52.9-72.3]                            | 45.9 [36.8-55.1]                         |
| <10              | 5,382 (1%)               | 4,502 (1%)                                  | 880 (1%)                                 |
| 10-19            | 5,421 (1%)               | 4,153 (1%)                                  | 1,268 (1%)                               |
| 20-29            | 19,501 (3%)              | 9,375 (2%)                                  | 10,126 (8%)                              |
| 30-39            | 56,485 (8%)              | 27,082 (5%)                                 | 29,403 (24%)                             |
| 40-49            | 95,059 (14%)             | 60,250 (11%)                                | 34,809 (28%)                             |
| 50-59            | 140,273 (21%)            | 110,247 (20%)                               | 30,026 (24%)                             |
| 60-69            | 159,385 (24%)            | 145,208 (27%)                               | 14,177 (11%)                             |
| >=70             | 170,928 (26%)            | 166,524 (31%)                               | 4,404 (4%)                               |
| Missing          | 12,435 (2%)              | 12,435 (2%)                                 | 0 (0%)                                   |
| **Calendar year of cancer diagnosis** |                     |                                             |                                          |
| 2004 – 2006      | 156,784 (24%)            | 132,775 (25%)                               | 24,009 (19%)                             |
| 2007 – 2010      | 219,123 (33%)            | 173,736 (32%)                               | 45,387 (36%)                             |
| 2011 - 2014      | 288,962 (43%)            | 233,265 (43%)                               | 55,697 (45%)                             |
| **Facility diagnosing cancer** |               |                                             |                                          |
| Public sector    | 329,280 (50%)            | 232,333 (43%)                               | 96,947 (77%)                             |
| Private sector | 335,589 (50%) | 307,443 (57%) | 28,146 (23%) |
| --- | --- | --- | --- |
| **Cancer type** |  |  |  |
| **HPV associated** |  |  |  |
| Cervix cancer | 57,161 (9%) | 32,393 (6%) | 24,768 (20%) |
| Oral cancer | 12,007 (2%) | 9,834 (2%) | 2,173 (2%) |
| Anal cancer | 1,906 (0%) | 1,305 (0%) | 601 (0%) |
| Penile cancer | 1,409 (0%) | 860 (0%) | 549 (0%) |
| Vulva cancer | 2,596 (0%) | 1,420 (0%) | 1,176 (1%) |
| Vaginal cancer | 1,587 (0%) | 1,058 (0%) | 529 (0%) |
| **HHV-8 associated** |  |  |  |
| Kaposi sarcoma | 23,046 (3%) | 6,961 (1%) | 16,085 (13%) |
| **EBV associated** |  |  |  |
| Non-Hodgkin lymphoma | 16,788 (3%) | 10,272 (2%) | 6,516 (5%) |
| Hodgkin lymphoma | 3,909 (1%) | 2,332 (0%) | 1,577 (1%) |
| Nasopharyngeal cancer | 4,335 (1%) | 3,368 (1%) | 967 (1%) |
| Burkitt lymphoma | 1,352 (0%) | 486 (0%) | 866 (1%) |
| **HBV/HCV associated** |  |  |  |
| Liver cancer | 4,707 (1%) | 3,724 (1%) | 983 (1%) |
| **Other HIV associated** |  |  |  |
| Eye cancer | 6,958 (1%) | 2,893 (1%) | 4,065 (3%) |
| **H. pylori associated** |  |  |  |
| Stomach cancer | 11,817 (2%) | 10,109 (2%) | 1,708 (1%) |
| **Schistosomiasis associated** |  |  |  |
| Bladder cancer | 12,098 (2%) | 10,971 (2%) | 1,127 (1%) |
| **Other aetiology** |  |  |  |
| Breast cancer | 73,609 (11%) | 60,006 (11%) | 13,603 (11%) |
| BCC | 123,185 (19%) | 116,036 (21%) | 7,149 (6%) |
| Prostate cancer | 57,970 (9%) | 53,703 (10%) | 4,267 (3%) |
| SCC of skin | 52,469 (8%) | 48,408 (9%) | 4,061 (3%) |
| Lung cancer | 24,762 (4%) | 21,011 (4%) | 3,751 (3%) |
| Colorectal cancer | 28,923 (4%) | 25,459 (5%) | 3,464 (3%) |
| Oesophageal cancer | 18,480 (3%) | 15,327 (3%) | 3,153 (3%) |
| Uterus cancer | 11,535 (2%) | 9,836 (2%) | 1,699 (1%) |
| Diagnosis                 | 2019 (Number, %) | 2020 (Number, %) | 2021 (Number, %) |
|---------------------------|------------------|------------------|------------------|
| Leukaemia                 | 6,859 (1%)       | 5,378 (1%)       | 1,481 (1%)       |
| Melanoma                  | 14,826 (2%)      | 13,364 (2%)      | 1,462 (1%)       |
| Any other cancer*         | 55,495 (8%)      | 44,973 (8%)      | 10,522 (8%)      |
| Primary site unknown      | 34,473 (5%)      | 27,766 (5%)      | 6,707 (5%)       |
| Ill defined               | 607 (0%)         | 523 (0%)         | 84 (0%)          |

*Any other cancer include Bone, Brain, Central Nervous System, Connective tissue, Endocrine, Gum, Haematology other, Kidney, Larynx, Mesothelioma, Myeloma, Other specified, Ovary, Pancreas, Placenta, Salivary gland, Skin other, Small intestine, Testis, Thyroid; NCR National Cancer Registry, IQR interquartile range, HPV Human Papilloma Virus, HHV-8 Human Herpesvirus 8, EBV Epstein Barr Virus, HBV Hepatitis B Virus, HCV Hepatitis C Virus, BCC basal cell carcinoma, SCC squamous cell carcinoma
Supplementary Table 3: Characteristics of people living with HIV with cancer included the South African HIV Cancer Match Study cohort

| Patients with prevalent cancer | Patients with prevalent cancer diagnosed more than 90 days before first HIV related record | Patients with prevalent cancer diagnosed up to 90 days before first HIV related record | Patients with incident cancer diagnosed 0 up to 90 days after first HIV record | Patients with incident cancer diagnosed > 90 days after first HIV record | Patients with cancer |
|--------------------------------|-------------------------------------------------------------|-------------------------------------------------------------|-------------------------------------------------------------|-------------------------------------------------------------|---------------------|
| Total number of patients       | 16963                                                       | 4222                                                        | 7549                                                        | 32575                                                      | 61309               |
| Sex                            |                                                             |                                                             |                                                             |                                                             |                     |
| Male                           | 6262 (37%)                                                 | 1264 (30%)                                                 | 2992 (40%)                                                 | 11140 (34%)                                                | 21658 (35%)         |
| Female                         | 10678 (63%)                                                | 2957 (70%)                                                 | 4555 (60%)                                                 | 21427 (66%)                                                | 39617 (65%)         |
| Missing                        | 23 (0%)                                                     | 1 (0%)                                                      | 2 (0%)                                                      | 8 (0%)                                                     | 34 (0%)             |
| Age at first test (years)      |                                                             |                                                             |                                                             |                                                             |                     |
| Median [IQR]                   | 47.0 [38.0 - 55.6]                                          | 42.7 [35.2 - 51.9]                                          | 39.8 [32.4 - 48.4]                                          | 40.7 [33.0 - 49.6]                                          | 42.4 [34.2 - 51.5]  |
| <10                            | 101 (1%)                                                    | 26 (1%)                                                     | 106 (1%)                                                   | 234 (1%)                                                   | 467 (1%)            |
| 10-19                          | 145 (1%)                                                    | 51 (1%)                                                     | 93 (1%)                                                     | 322 (1%)                                                   | 611 (1%)            |
| 20-29                          | 1115 (7%)                                                   | 409 (10%)                                                   | 1064 (14%)                                                 | 4726 (15%)                                                 | 7314 (12%)          |
| 30-39                          | 3751 (22%)                                                  | 1226 (29%)                                                  | 2539 (34%)                                                 | 10189 (31%)                                                | 17705 (29%)         |
| 40-49                          | 4849 (29%)                                                  | 1234 (29%)                                                  | 2066 (27%)                                                 | 9185 (28%)                                                 | 17334 (28%)         |
| >= 50                          | 6875 (41%)                                                  | 1233 (29%)                                                  | 1618 (21%)                                                 | 7781 (24%)                                                 | 17507 (29%)         |
| Missing                        | 127 (1%)                                                    | 43 (1%)                                                     | 63 (1%)                                                     | 138 (0%)                                                   | 371 (1%)            |
| CD4 counts (cells/µl)          |                                                             |                                                             |                                                             |                                                             |                     |
| Total No. of counts            | 54035                                                       | 13180                                                       | 22378                                                       | 147110                                                     | 236703              |
| First count | Median [IQR] | 263 [137 - 432] | 273 [132 - 466] | 214 [106 - 383] | 235 [117 - 391] | 242 [122 - 406] |
|-------------|-------------|----------------|----------------|----------------|----------------|----------------|
| <50         |             | 1373 (8%)      | 367 (9%)       | 837 (11%)      | 3324 (10%)     | 5901 (10%)     |
| 50-99       |             | 1519 (9%)      | 388 (9%)       | 885 (12%)      | 3492 (11%)     | 6284 (10%)     |
| 100-199     |             | 3411 (20%)     | 763 (18%)      | 1744 (23%)     | 6964 (21%)     | 12882 (21%)    |
| 200-349     |             | 4446 (26%)     | 1011 (24%)     | 1757 (23%)     | 8597 (26%)     | 15811 (26%)    |
| 350-499     |             | 2862 (17%)     | 639 (15%)      | 983 (13%)      | 4955 (15%)     | 9439 (15%)     |
| 500-699     |             | 1844 (11%)     | 431 (10%)      | 585 (8%)       | 2900 (9%)      | 5760 (9%)      |
| ≥ 700       |             | 1259 (7%)      | 456 (11%)      | 522 (7%)       | 1862 (6%)      | 4099 (7%)      |
| Missing     |             | 249 (1%)       | 167 (4%)       | 236 (3%)       | 481 (1%)       | 1133 (2%)      |

| HIV RNA viral load (log10 copies/ml) |
|--------------------------------------|
| Total No. of measurements            | 32654 | 6022 | 10253 | 81986 | 130915 |

| First measurement | Median [IQR] | 2.0 [-0.3 - 3.5] | 2.8 [1.4 - 4.7] | 3.3 [1.7 - 4.9] | 2.6 [1 - 4.6] | 2.6 [1 - 4.5] |
|-------------------|-------------|----------------|----------------|----------------|----------------|----------------|
| <2.7              |             | 4112 (24%)     | 588 (14%)      | 972 (13%)      | 6091 (19%)     | 11763 (19%)    |
| 2.7-3.9           |             | 1253 (7%)      | 288 (7%)       | 515 (7%)       | 2531 (8%)      | 4587 (7%)      |
| 4.0-4.9           |             | 1216 (7%)      | 397 (9%)       | 710 (9%)       | 3597 (11%)     | 5920 (10%)     |
| ≥ 5               |             | 1265 (7%)      | 379 (9%)       | 871 (12%)      | 3922 (12%)     | 6437 (10%)     |
| Missing           |             | 9117 (54%)     | 2570 (61%)     | 4481 (59%)     | 16434 (50%)    | 32602 (53%)    |
Table 3 cont.

| Follow-up time (years) | Median [IQR] | Median time [IQR] between labs (months) | Median time [IQR] between labs (months) |
|------------------------|--------------|----------------------------------------|----------------------------------------|
| Follow-up time (years) | 1.7 [0.6 - 3.3] | 0.5 [0.1 - 2.1] | 0.3 [0.1 - 1.5] | 3.0 [1.1 - 5.4] | 2.0 [0.5 - 4.2] |
| Follow-up time (years) | 7.7 [4.6 - 12.0] | 3.0 [0.6 - 8.0] | 1.9 [0.5 - 6.0] | 8.2 [5.2 - 13.4] | 7.2 [3.5 - 12.0] |
| Urbanity level* | Rural | 8191 (48%) | 1521 (36%) | 2515 (33%) | 14381 (44%) | 26608 (43%) |
| Urbanity level* | Urban | 8667 (51%) | 2681 (64%) | 4986 (66%) | 17973 (55%) | 34307 (56%) |
| Urbanity level* | Missing | 105 (1%) | 20 (0%) | 48 (1%) | 221 (1%) | 394 (1%) |
| Province** | Gauteng | 5126 (30%) | 1594 (38%) | 3129 (41%) | 10928 (34%) | 20777 (34%) |
| Province** | Kwazulu-Natal | 4192 (25%) | 282 (7%) | 437 (6%) | 2988 (9%) | 7899 (13%) |
| Province** | Eastern Cape | 1183 (7%) | 318 (8%) | 587 (8%) | 2856 (9%) | 4944 (8%) |
| Province** | Mpumalanga | 1543 (9%) | 287 (7%) | 632 (8%) | 3717 (11%) | 6179 (10%) |
| Province** | North West | 1001 (6%) | 225 (5%) | 620 (8%) | 2669 (8%) | 4515 (7%) |
| Province** | Western Cape | 868 (5%) | 419 (10%) | 927 (12%) | 3576 (11%) | 5790 (9%) |
| Province** | Limpopo | 1433 (8%) | 451 (11%) | 404 (5%) | 2202 (7%) | 4490 (7%) |
| Province** | Free State | 1048 (6%) | 533 (13%) | 622 (8%) | 2350 (7%) | 4553 (7%) |
| Province** | Northern Cape | 529 (3%) | 106 (3%) | 181 (2%) | 1220 (4%) | 2036 (3%) |
| Province** | Missing | 40 (0%) | 7 (0%) | 10 (0%) | 69 (0%) | 126 (0%) |

* Location of first laboratory facility providing HIV services, ** provincial location of first laboratory facility providing HIV services; IQR interquartile range.