CANCER BURDEN AMONG HIV-INFECTED PERSONS ON ANTIRETROVIRAL THERAPY IN MALAWI: A RECORD LINKAGE STUDY

Marie-Josèphe Horner

A dissertation submitted to the faculty at the University of North Carolina at Chapel Hill in partial fulfillment of the requirements for the degree of Doctor of Philosophy in the Department of Epidemiology in the Gillings School of Global Public Health.

Chapel Hill
2018

Approved by:

Andrew Olshan

Satish Gopal

Jessie K. Edwards

Willam B. Miller

Steve Cole

ABSTRACT

Marie-Josèphe Horner: Cancer Burden Among HIV-Infected Persons on Antiretroviral Therapy in Malawi: A Record Linkage Study
(Under the direction of Andrew Olshan)

Sub-Saharan Africa represents 70% of the global number of people living with HIV. The regional HIV epidemic is reflected in the cancer burden, where AIDS-defining cancers are among the most common malignancies in the region. Early access and continued adherence to antiretroviral therapy (ART) may reduce the risk of certain cancers among the HIV population. Local epidemiological data are needed to characterize the cancer burden among African HIV populations during the ART era.

In the Malawi HIV-Cancer Match Study, we used algorithms to link cancer cases from the population-based cancer registry of Malawi with electronic medical records from two high volume HIV clinics. We constructed a clinical cohort of 29,000 people who initiated ART from 2000 to 2010 at Lighthouse Trust and Queen Elizabeth Central Hospital. We described implementation of a healthcare data linkage in a resource-constrained setting, common analytical barriers, and solutions. We used Poisson regression to estimate cancer incidence rates and describe the timing of new cancer diagnoses after starting ART.

Missing data and misreporting of patient identifiers resulted in a substantial proportion of potential cancer cases being discarded from analysis. Consequently, missing data on potential cancer cases may have diminished sensitivity of the linkage algorithms. Sensitivity analysis of incidence rates was used to address scenarios of uncertainty in the linkage process.

Two AIDS-defining malignancies, Kaposi sarcoma (KS) and cervical cancer, were the most common cancers in this young population of ART users who tend to present to care with severe immunosuppression. Most incident KS occurred within the first two years of starting ART, and elevated incidence rates persisted over the course of follow-up in spite of therapy. AIDS-associated non-Hodgkin lymphomas and a heterogeneous spectrum of NADC were also observed, but at low incidence rates.

Our study is a baseline against which to monitor the contemporary burden of cancer among people who are now starting ART at earlier stages of HIV, when therapy likely to have a substantial impact on

cancer incidence. Descriptive epidemiological data on people living with HIV is important for public health decision makers in Malawi to develop evidence-based cancer control plans targeting high-risk HIV populations.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

ADC	AIDS-defining cancer
AIDS	acquired immunodeficiency syndrome
ART	antiretroviral therapy
CI	confidence interval
EMR	electronic medical record
HAART	highly active anti-retroviral therapy
HBV	hepatitis B virus
HCV	hepatitis C virus
HIV	human immunodeficiency virus
HPV	human papilloma virus
IARC	International Agency for Research on Cancer, Lyon, France
IeDEA	International epidemiologic Databases to Evaluate AIDS
IR	incidence rate
IRR	incidence rate ratio
KS	Kaposi sarcoma
LT	Lighthouse Trust ART clinic, Lilongwe
NADC	non-AIDS-defining cancer
NHL	non-Hodgkin lymphoma
NHSRC	Malawi National Health Sciences Research and Ethics Committee
QECH	Queen Elizabeth Central Hospital ART clinic, Blantyre
SIR	standardized incidence ratio
SSA	sub-Saharan Africa
WHO	World Health Organization

CHAPTER I. SPECIFIC AIMS

The African HIV epidemic is reflected in a heavy burden of AIDS-defining cancers (ADC), with Kaposi sarcoma (KS), non-Hodgkin lymphoma (NHL), and cervical cancer being among the most common malignancies in sub-Saharan Africa (SSA).[1, 2] Eastern and Southern regions in SSA have implemented rapid scale-up of HIV treatment, and 10 countries have achieved greater than 80% ART coverage over the past decade.[3] Yet the extent to which contemporary ART availability is impacting the African cancer burden remains largely unknown. Since the introduction of combination ART regimens in the West in the mid-1990s and highly active antiretroviral therapy (HAART) in 1997[4] the burden and incidence rates of KS and NHL have declined substantially.[5-8] Incidence declines for NADC have been much more modest, and NADC burden has actually increased with growth and aging of HIV-infected populations, as well as declines in competing causes of death.[7, 9] It remains to be seen whether the cancer trends among persons living with HIV in developed countries[10] will be replicated in Africa, as epidemiological studies from the region are limited in quality and number.[11-13]

Given typically low-quality data sources for robust epidemiological studies in SSA, innovative approaches are required to overcome cancer surveillance obstacles in low- and middle-income countries. Electronic linkage of medical records is an efficient strategy for constructing observational patient cohorts, yet methods to implement this approach in low- and middle-income countries have not been well described. Our goal is to address the central knowledge gap of how to implement data linkage strategies in a resource-limited setting to construct a locally relevant data resource for evaluating cancer burden among people living with HIV in Africa.

Our proposal focuses on Malawi, where HIV prevalence is 11% and ART coverage has currently reached 67% among guideline-eligible HIV patients; one in 20 Malawian adults is currently on ART.[14, 15] Expanded eligibility criteria among adults include WHO stage 1 or 2 and CD4 counts ≤500 cells/mm^3, WHO stage 3 or 4 regardless of CD4 count, and universal ART for HIV-infected pregnant and breastfeeding women (Option B+).[16] With respect to malignancies, cervical cancer and KS feature as the

most common cancer types among women and men, respectively[17]. UNC-Project Malawi has long-standing collaborations with the Malawian government and in-country stakeholders for HIV and cancer care and research. These provide a strong foundation to implement this new initiative, which is the basis of this dissertation project: the Malawi HIV-Cancer Match Study. The study is innovative in that it will leverage a hybrid of probabilistic algorithms and extensive clerical review to link data from Malawi's national cancer registry with electronic medical systems created to support antiretroviral therapy delivery within large HIV cohorts. Record linkage using existing data is a cost-efficient strategy for resource-constrained environments and imposes no additional burden on local health care workers in Malawi. The hybrid method will be locally tailored to overcome real-world limitations of missing data and lack of unique identifiers in low- and middle-income countries. Further epidemiological methods will be applied to assess and account for potential outcome misclassification errors resulting in false positive matches (records that linked erroneously) and false negative matches (records that failed to link).[18, 19]

This large-scales study creates a new data resource to answer high impact public health questions regarding the effect of real-world ART delivery on cancer incidence and the clinical timing of cancer development among persons living with HIV in SSA. Characterization of contributing factors and patterns of cancer occurrence in HIV populations using high-quality data derived from within the region, rather than extrapolating from studies conducted in resource-rich settings, will be valuable for informing evidence-based national cancer control efforts. Our specific aims seek to characterize cancer incidence patterns among contemporary ART initiators in Malawi, while accounting for measurement error in a statistically rigorous way:

Specific Aim 1. Adapt, implement, and evaluate a hybrid approach of probabilistic-deterministic linkage methods suited to health systems in a resource-limited setting.

Subaim 1.1. Linkage methodology will be delineated in the context of locally relevant ethnographic considerations and missing data in low- and middle-income countries. Our study is a proof of concept of an innovative health systems approach tailored to cancer surveillance in SSA.

Subaim 1.2. Describe associations between missing data and data accuracy on linkage weights. Missing data and lack of resolution among patient identifiers may cause measurement error in HIV-cancer match status and may introduce bias in subsequent analyses. The use of probabilistic algorithms alone in

low- and middle-income settings may be overly conservative in assigning HIV-cancer matches due to underlying missing data in the local health systems.

Specific Aim 2. Characterize cancer incidence rates and clinical timing of cancer diagnosis relative to ART start.

The Malawi HIV-Cancer Match Study is uniquely positioned to link data from high quality population-based cancer registration and well-established ART cohorts participating in the International epidemiologic Databases to Evaluate AIDS (IeDEA) consortium. Our newly constructed resource incorporates patient-level clinical data enriched with active patient follow-up, WHO stage of HIV, ART regimens, and population-based cancer ascertainment meeting data quality standards for *Cancer in V Continents* (IARC, Lyon).[20]

CHAPTER 2. BACKGROUND

A. HIV infection, immunodeficiency, inflammation and risk of cancer

Hallmarks of HIV infection

People living with HIV have a greater risk of certain cancers compared to the general population and this association is largely due to profound immunodeficiency and co-infections with oncogenic viruses [5, 9, 21-28]. Three hallmarks of HIV infection are implicated in various pathways of cancer development: immune deficiency, chronic inflammation and immune system activation, and immune senesence[29, 30]. As HIV progresses, CD4+ T lymphocytes become increasingly depleted. Chronic immune activation and sustained inflammation are triggered. Lastly, the immune system of people living with HIV also shares a feature normally seen with old age: loss of regenerative capacity, or senescence[30]. Senescence drives the increased risk of cancer associated with aging[31, 32]. The effects of HIV- associated chronic inflammation and dysfunction of the immune system show similarities with cellular aging[33]. Inflammation and depleted CD4 cell counts often persist despite ART, and despite viral suppression, but at lower levels than untreated people[32, 34]. Together these findings have public health implications for managing the risk of cancer among persons living with HIV.

CD4 cell counts and cancer risk

Mechanistically, the question of how immunodeficiency contributes to cancer risk is more nuanced. Low CD4 cell count is inversely associated with risk of KS and NHL[35-37]. A remaining question is whether current CD4 cell count or the duration of immunodeficiency is what drives increased cancer risk, particularly for NADC (Table 2-1). The effect size of the association between CD4 levels on NADC versus ADC varies.

Table 2-1. Associations between CD4 cell count, timing of CD4 measurement, and duration of immunosuppression and cancer risk

Outcome	Association	Definition of CD4 exposure measurement	Incidence rate and distribution of cancer types
Grouped NADC	Inverse association, especially among virally-associated cancers, with lower CD4 count [38-41], even after adjusting for smoking, alcohol use, and co-infections with Hepatitis B virus or Hepatitis C Virus. Initial ART drug class was not associated with NADC [38]. Deaths from NADC was inversely associated with CD4 counts [42]. HIV viremia was not associated with increased risk of NADC [43].	Agence Nationale de Recherche sur le SIDA (ANRS) CO3 Aquitaine Cohort[40] used CD4<500 cut point AIDS Clinical Trials Group trials[38] used time-updated CD4 count; 25% had CD4<350 EuroSIDA cohort[41] used current CD4 D:A:D Study cohort [42] used most recent CD4 stratifying by counts <50 to >500 in Poisson modeling	In the Italian cohort [39] the incidence of ADC was 5.0 per 1000 person-years (95%CI 4.3, 5.8) and 2.4 per 1000 person-years (95%CI 1.9, 3.1) for NADC. In the EuroSIDA cohort, the incidence of NADC was 4.3 per 1000 person years (95%CI 3.8, 4.7). 48% of NADC were virus-related, 38% were non-virus-related epithelial cancers, and 14% were 'other'. In the D:A:D Study, the ADC mortality among those with a CD4 count <50 was 20 per 1000 person-years (95%CI: 14.4, 25.9) compared to 0.1 (95%CI 0.03, 0.3) among those with CD4 count >500; NADC mortality was 6.0 (95%CI 3.3, 10.1) and 0.6 (95%CI 0.4, 0.8), respectively.
HPV-related cancers: cervical cancer	Inverse association between with CD4 count and elevated HPV viral load, HPV persistence, precancerous lesions (reviewed in [44])		
HPV-related cancers: anal cancer	Inverse association with low CD4 at baseline[45], low CD4 at onset of AIDS[46], long duration of low CD4 count[43, 47], low nadir CD4[48]. 13 North American cohorts[6] , AIDS-cancer match[46] , U.S. Military Natural History Study[48]	French Hospital Database on HIV cohort[47] used current CD4 (most predictive among 72 models). Rate ratios using CD4>500 as referent and categories of CD4 counts 350-499 and <50 AIDS Therapy Evaluation in the Netherlands (ATHENA) cohort[43] used time-updated CD4 and HIV viral load and cumulative exposure time to CD4 counts of < 200, < 350, < 500 and viral load >50, >400, and >1000 copies/mL	Risk of HPV associated cancers increased with increasing levels of immunosuppresion[46]. SIRs compared HIV-infected and non-infected people

Hallmarks of cancer: evasion of immune destruction and tumor-promoting inflammation

The *hallmarks of cancer* represent a unifying view of processes that enable the development and proliferation of tumors [49, 50]. The classic etiologic framework recognizes tumorigenesis as a multistep process of cellular evolution in which normal and aberrant cells participate in complex interactions with the tumor microenvironment in which they grow. Many types of cells and features comprise the microenvironment: extracellular scaffolding, stromal compartments, cancer stem cells, endothelial cells, pericytes which synthesize vascular membranes, cytokine and chemokine cellular signaling chemicals, and immune inflammatory cells. Though different tumor types possess their own district set of etiologic mechanisms, it is recognized that there is a set of hallmark functional capabilities that are required for cancers to proliferate, survive, and metastasize. The six classic hallmark traits are self-sufficiency in growth signals, insensitivity to growth suppressors, evasion of apoptosis, limitless replicative potential, sustained angiogenesis, and tissue invasion and metastasis. Deregulation of cellular metabolism and *evasion of immune destruction* are now recognized as additional emerging hallmark capabilities[50]. Underlying these traits, genomic instability in tumor cells and a state of *tumor-promoting inflammation* are recognized as two critical *enabling characteristics* for acquiring hallmark cancer traits.

The enabling characteristic of *inflammation* plays a dualistic role in tumor- microenvironment interactions[51] . Paradoxically, immune inflammatory cells both antagonize and enhance the acquisition of hallmark capabilities among newly formed aberrant cells, a phenomenon that is not entirely counterintuitive considering the wound-healing properties of inflammation[51]. Cells from the innate arm of the immune system interact with the microenvironment by releasing growth factor signals, pro-angiogenic factors, signals to control programmed cell death, as well as enzymes that are used in remodeling extracellular scaffolding and promoting tissue proliferation -- all of which are also hallmark capabilities that enhance neoplastic initiation, proliferation, and invasion. But while wound-healing inflammation is generally transient, various degrees of chronic inflammation in the tumor microenvironment have been observed along the full spectrum of neoplastic transformation, from the early stages of pre-malignant tumor lesions through advanced stage cancer. In fact, observations of tumor infiltration by white blood cells led to the analogy of tumors as “wounds that never heal” [52]. Compounding the problem is that inflammatory cells also release reactive oxygen species into the tumor microenvironment. Reactive

oxygen species are mutagenic, and therefore add momentum in generating even further genomic instability among neoplastic cells.

The emerging hallmark of *evasion of immune destruction* by cancers has its roots in the early observations of immune system suppression of tumor growth by Elrich in 1909 and later theories of immunosurveillance by immunologists Burnet and Thomas in the 1950s[53-55]. They postulate a constant monitoring of cells and tissues by the immune system which recognize tumor cell-specific antigens and then attack nascent tumor cells. The early theory assumed that cancers arise through an acquired capability of evading detection by the immune system, though we now understand that immunosurveillance[56] is just one step in the larger scheme of immunoediting of tumors by the immune system[57-59].

Immunoediting has three 3 phases: elimination, equilibrium, and escape [57]. Direct and indirect mechanisms of the innate and adaptive arms of the immune system play important roles in cancer immunosurveillance and immunoediting (Table 2-2)[51]. The elimination phase refers to the role of immunosurveillance as an extrinsic tumor suppressor. The ability of the immune system to distinguish tumor cells predominantly comes from CD4+ and CD8+ T lymphocytes which recognize Major Histocompatibility Complex (MHC) class I and II proteins, though many other receptors and molecules are also involved in recruiting innate and adaptive immune effectors [57]. Tumor specific cell-surface antigens[a] presented on MHC proteins arise from the expression of mutated oncogenes, aberrant expression of self-proteins, and even certain viral proteins among cell types harboring these viruses[59-61]. Highly antigenic tumor cells are effectively pared down through components of both the innate and adaptive arms of the immune system[b], the process known as elimination. Equilibrium is the phase during which incomplete tumor destruction gives rise to weakly antigenic cellular clones that persist in tumor immune-mediated

[a] Tumor antigens can be broadly grouped according to 1) differentiation antigens, e.g., melanocyte differentiation antigens, Melan-A/MART-1, tyrosinase, gp-100; (2) mutational antigens, e.g., abnormal forms of p53; (3) overexpressed/amplified antigens, e.g.,HER-2/neu; (4) cancer-testis (CT) antigens, e.g., MAGE and NY-ESO-1; (5) viral antigens, e.g., EBV and HPV.

[b] The innate immune system is the “first line of defense” against pathogens such as bacteria. The innate arm is comprised of mast cells, granulocytes, natural killer cells, and tumor-infiltrating phagocytic macrophages and neutrophils (known as inflammatory cells). Inflammatory cells also play a role in directing the adaptive arm of the immune system. The adaptive arm is comprised of lymphocytes which undergo clonal selection for antigen-specific receptors. The adaptive arm is involved in immunologic memory.[23]

latency [57]. In the escape phase, tumors exit equilibrium and begin to proliferate and infiltrate by subverting normal innate and adaptive anti-tumor defenses.

Table 2-2. Summary of direct and indirect mechanisms of the innate and adaptive arms of immune system regulation during cancer immunosurveillance and immunoediting

Innate immunity	Adaptive immunity
Direct mechanisms	
Induction of DNA damage by reactive oxygen species and free radicals	Inhibition of tumor growth by antitumor cytotoxic-T-cell activity
Paracrine regulation of intracellular pathways via nuclear factor κB	Inhibition of tumor growth by cytokine-mediated lysis of tumor cells
Indirect mechanisms	
Promotion of angiogenesis and tissue remodeling by the production of growth factors, cytokines, chemokines and matrix metalloproteinases	Promotion of tumor growth by regulatory T cells that suppress antitumor T-cell responses
Cyclooxygenase-2 (COX-2) upregulation	Promotion of tumor development by humoral immune responses that increase chronic inflammation in the tumor microenvironment
Suppression of antitumor adaptive immune responses	

*adapted from Visser et al 2008

B. Epidemiology of cancer in the HIV/AIDS population

The overall distribution of cancers in the HIV population is skewed, with Kaposi sarcoma, cervical cancer and central nervous system lymphoma, three AIDS-defining cancers (ADC), featuring prominently. The HIV population is also at increased risk of non-AIDS defining cancers (NADC), particularly those associated with viral co-infections[22, 27, 28, 62-64]. Virally-associated cancers include Kaposi sarcoma and human herpes-virus 8 (HHV-8); subtypes of Hodgkin and non-Hodgkin lymphomas and Epstein Barr virus (EBV); cervical and subsets of anal, oropharyngeal cancers and human papilloma virus (HPV); liver cancer and hepatitis B and C viruses (HBV, HCV).

Advanced immune suppression, chronic inflammation and viral co-infection are important drivers of the increased risk of malignancies among people living with HIV [28, 64]. An important meta-analysis seeking to elucidate the role of immune deficiency compared the incidence of a broad range of cancer types in HIV patients and in organ transplant recipients to the general population[20]. These two patient groups share the common characteristic of immune deficiency, but presumably no other commonalities: organ

transplant recipients receive pharmacologically-induced immune suppressors to prevent transplant rejection while HIV patients ultimately succumb to profound immunodeficiency in the absence of antiretroviral therapy. In the 7 population-based studies of HIV patients, most were followed after onset of AIDS. Cancer incidence rates were elevated in both patient groups relative to the general population for the majority of the 28 cancers under investigation, suggesting that immune deficiency is driving the association. It is noteworthy that the 15 cancer types with infectious etiologies (Epstein-Barr virus, human herpesvirus 8, hepatitis B and C viruses, human papillomavirus, and *Helicobacter pylori*) all had elevated rates (Table 2-3). The rate of Kaposi sarcoma among HIV patients was 3640-times higher that of the general population, the rate of NHL was 77 times higher, and the rate of cervical cancer was 6 times higher. Rates were also 29 times higher for anal cancer, 11 times higher for Hodgkin lymphoma, and 6 times higher for liver cancer. Oncogenic viruses themselves have direct effects on cancer induction[63] and these are exacerbated by immunosuppression through persistence of infection, uncontrolled viral replication, and uncontrolled latent infections. Together, impaired ability of the immune system to control infection and a higher prevalence of HPV and HBV co-infections among people living with HIV drive the elevated occurrence of these infection-related cancers[65].

The HIV population does not exhibit a greater risk of non-infection related, common epithelial cancers compared to the general population [22, 23, 28, 64]. The rates of breast and colorectal cancers, both epithelial cancers, were not elevated compared to the general population, with standardized incidence ratios (SIR) generally close to one[28]. The rate of prostate cancer was 30% lower among the HIV population[28, 66]. Ovarian cancer had modest, significantly elevated rates. Though HIV patients experience chronic immunodeficiency arising primarily through deficits of T and B cells, they may still possess some functionality of the innate immune system (e.g. natural killer cells) to partially suppress cancer development, at least for non-infection related epithelial cancers. Rates of lung cancer were 3 times higher, and thought to be due to the higher prevalence of smoking in the HIV population[67-69]. Higher rates of other smoking-related cancers, kidney (SIR=1.7) and laryngeal cancers (SIR=1.5), were found in a subsequent meta-analysis[64]. However, the association between HIV and lung cancer persists independently of smoking[70-73] , suggesting that other factors could be important as well. Frequent

pulmonary infections and chronic inflammation may also be associated with the greater occurrence of lung cancer among the HIV population [74].

Table 2-3. Summary of meta-analysis standardized incidence ratios (SIR) and 95%CI among HIV/AIDS patients and organ transplant recipients relative to the general population

Infectious agent	Infection-related or possibly-related cancer site	HIV cohort SIR (95% CI)	Transplant cohort SIR (95%CI)
Epstein-Barr Virus	Hodgkin's lymphoma	11.03 (8.43-14.4)	3.89 (2.42-6.26)
	Non-Hodgkin lymphoma	76.67 (39.4-149)	9.07 (6.40-10.2)
Human Herpes Virus-8	Kaposi sarcoma	3640.0 (3326-3976)	208.0 (114-349)
Hepatitis B Virus Hepatitis C Virus	Liver cancer	5.22 (3.32-8.20)	2.13 (1.16-3.91)
Helicobacter pylori	Stomach cancer	1.09 (1.53-2.36)	2.04 (1.49-2.79)
Human papilloma viruses	Cervix uteri	5.82 (2.98–11.3)	2.13 (1.37–3.30)
	Vulva vagina	6.45 (4.07–10.2)	22.76 (15.8–32.7)
	Penis	22.76 (15.8–32.7)	15.79 (5.79–34.4)
	Anus	28.75 (21.6–38.3)	4.85 (1.36–17.3)
	Oral cavity and pharynx (a subset are possibly related)	2.32 (1.65–3.25)	3.23 (2.40–4.35)
	Non-melanoma skin (possibly related)	4.11 (1.08–16.6)	28.62 (9.39–87.2)
	Lip (possibly related)	2.80 (1.91–4.11)	30.00 (16.3–55.3)
	Esophagus (possibly related, though current evidence supporting this association is equivocal)	1.62 (1.20–2.19)	3.05 (1.87–4.98)
	Larynx (possibly related)	2.72 (2.29–3.22)	1.99 (1.23–3.23)
	Eye (possibly related)	1.98 (1.03–3.81)	6.94 (3.49–13.8)
Non-infection related epithelial cancers		HIV cohort SIR (95% CI)	Transplant cohort SIR (95%CI)
	Breast	1.03 (0.89-1.20)	1.15 (0.98-1.36)
	Prostate	0.70 (0.55-0.89)	0.97 (0.78-1.19)
	Ovary	1.63 (0.95-2.80)	1.55 (0.99-2.43)
	Colon, rectum	0.92 (0.78-1.08)	1.69 (2.34-2.43)
	Lung	2.72 (1.91-3.87)	2.18 (1.85-2.57)

* adapted from Grulich et al 2007

C. Evolution of the cancer burden during the HAART era

A major public health question is whether ART reduces the risk of HIV-associated cancers and NADC over time [75]. ART has improved immunocompetence and reduced the risk of AIDS, AIDS-related deaths and all-cause mortality[76-79]. In parallel, the cancer burden in the HIV population has also begun to evolve- combination ART has caused a dramatic shift in the number of new cancer cases and distribution of cancer types in the HIV/AIDS population, though the incidence rates of many cancers remain elevated relative to the general population[5, 9, 28, 64, 80]. Since the introduction of combination ART regimens in early 1990s and highly active antiretroviral therapy (HAART) in 1996, the burden and incidence rates of KS and NHL have declined substantially[5-8], while the burden of NADC has increased[7, 9].

In a study of cancer trends during the pre-HAART through late HAART period in the HIV/AIDS Cancer Match Study, the number of incident KS cases decreased by 82% and NHL by 53% over a 15 year period from 1991-2005[9]. Incidence rates for KS and NHL decreased sharply during 1991-1997, and more gradually through 2005. On the other hand, the number of cervical cancer cases increased by 62%, though the incidence rate had declined.

In the AIDS population, the overall incidence of NADC declined over time but trends were not uniform across cancer sites[9]. Incidence rates of Hodgkin lymphoma, liver, lung, and colorectal cancers declined or remained constant, while rates of anal and prostate cancer incidence increased. The increased incidence rate of prostate cancer occurred against a backdrop of declining incidence in the general population[81]. Overall, substantial increases in the absolute burden of NADC in the late HAART period were observed: anal cancer, lung cancer, prostate cancer, Hodgkin lymphoma and liver cancer, together accounted for 50% of the NADC burden, yet account these sites account for less than 20% in the general population [82]. The burden of prostate and colorectal cancers, which are considered to be non-HIV-related epithelial cancers, also increased significantly. The burden of uterine, vulva, and cervical cancer also increased.

Demographic changes are largely responsible for trends in the cancer burden: growth of the AIDS population has more than quadrupled in size in the US, and consequently increased number of people at risk[9]. Meanwhile, aging of the HIV/AIDS population has shifted at risk persons into a demographic where the risk of many NADC is greater[7, 9, 10]. Of note, the majority of prostate cancers occurred among those

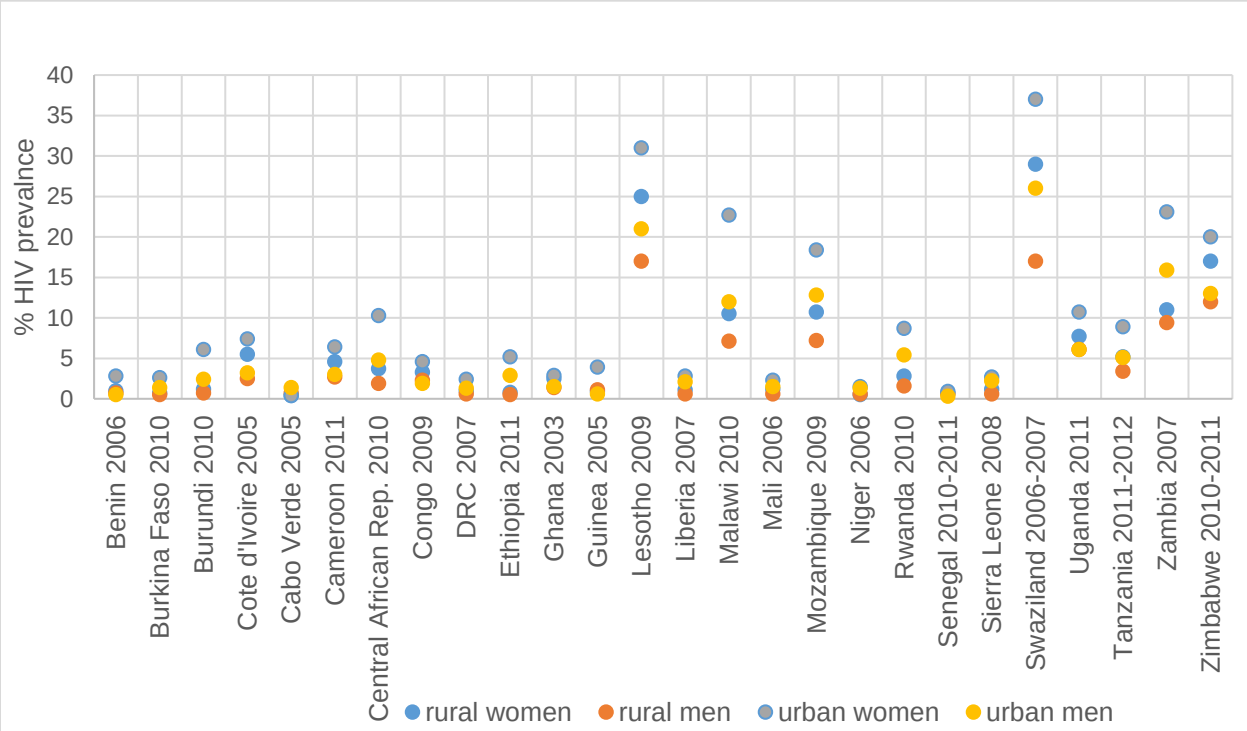
older than 50 years old in the HIV/AIDS Cancer Match Study. In the late HAART era, NADC represent 50% of malignancies in the HIV-only population, with lung cancer accounting for 20% of the total cancer burden. The number of NADC now exceed that of ADC among the AIDS population in the US, though KS continues to dominate as the most common specific cancer type[9].

Despite many clinical gains through the introduction of HAART in 1996, prognosis after a cancer diagnosis remains poor. Compared to their HIV-uninfected counterparts, patients with HIV present with more advanced cancer stage at diagnosis[83] , experience worse survival[77], and have higher cancer-specific mortality[84-86]. Lung, breast, prostate, melanoma, and bladder cancers are more likely to be diagnosed at distal stages, highlighting potentially distinct tumorigenic mechanisms related to immune suppression that we still do not fully understand, and potential barriers in access to medical care among the HIV population[83]. Cancer deaths have declined due to HAART, but because of the decrease in overall mortality, malignancies now represent a growing proportion of total deaths among HIV/AIDS populations in developed countries[78, 87]. Cancer accounts for nearly one-third of deaths in the HIV/AIDS population[88]. Cancer-specific mortality remains elevated compared to the HIV-uninfected cancer patient population, independently of cancer stage at diagnosis and receipt of treatment[86].

D. The HIV epidemic and ART scale-up in Sub-Saharan Africa

Sub-Saharan Africa is at the center of the HIV epidemic with 25.6 million people affected in 2015, representing 70% of the global HIV burden for only 15% of the world's population[89, 90] . Two-thirds of new HIV infections occur in SSA. Within the region, there is geographic heterogeneity in the intensity of the epidemic: 9 countries in Southern Africa have the highest prevalence and together represent one-third of the global number of people living with HIV/AIDS. Swaziland, Botswana and Lesotho have the most severe epidemic with HIV prevalences between 23%-26% among adults; Malawi, Mozambique, Namibia, South Africa, Zambia and Zimbabwe have HIV prevalences exceeding 10%. Within-country variability also reveals sizeable disparities in the epidemic across gender and rural-urban areas (Figure 2-1).

Figure 2-1. HIV prevalence in sub-Saharan Africa, by sex and urban/rural residence



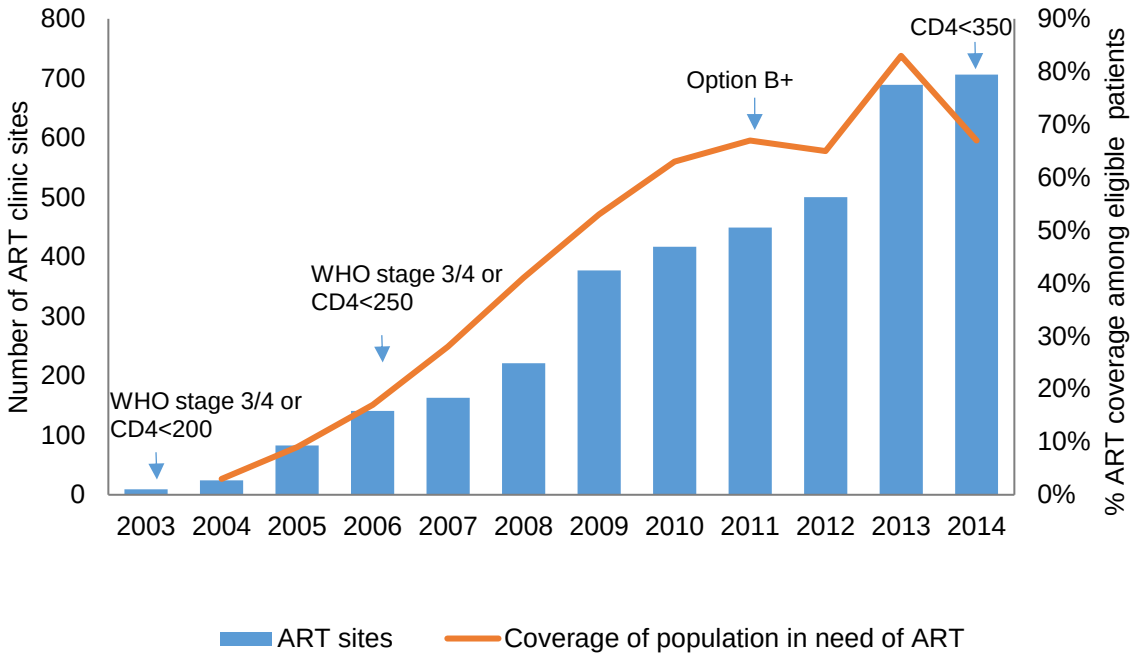
Source: World Health Organization. Global Health Observatory data repository. Accessible at: http://apps.who.int/gho/data/node.main.247?lang=en [Accessed on August 17, 2016]

Over the past decade, global partnerships with the WHO, donors, and local governments, combined with competitive bidding to reduce the cost of antiretroviral drugs, have led to a major expansion of access to ART in Africa. Though many obstacles and gaps still remain in the implementation of HIV prevention and treatment in SSA, there have also been notable successes. As of the 2010 WHO guidelines, more than 10 SSA countries have exceed 80% ART coverage[3] and this figure is expected to increase as a greater number of people are now eligible under the new WHO guidelines[91]. In the latest UNAIDS Gap Report, nearly 50% of persons living with HIV in SSA know their status. Among these, 87% are receiving ART, and 75% of those on treatment have achieved viral suppresion[92]. These initiatives have translated to 7.6 million people receiving ART as of 2012, and nearly 5 million lives saved in Africa[89]. Because ART also prevents transmission of HIV by up to 96%[93], the paradigm of treatment as prevention has in turn translated to a reduction in new HIV infections by more than 50% in Malawi, Botswana, Ethiopia, Namibia, Rwanda, Zambia and Zimbabwe, and reductions of more than 25% in Swaziland, South Africa, Kenya, and Mozambique[94].

HIV prevalence in Malawi is 10.6%, with the highest prevalence reaching 14.5% in southern districts and 22.5% among urban women[15]. Demographically, Malawi ranks 174 on the Human Development Index, the most recent estimate of average life expectancy at birth is 62.8 years, and gross national income per capita is less than one US dollar per day[95, 96]. Prior to the implementation of ART programs, life expectancy declined from 54 years to 39 years in 2000, and this has been directly attributed to the rampant, uncontrolled HIV epidemic[97].

In many ways, Malawi has been a leader in nationwide scale-up of ART to tackle the epidemic. In 2002, Malawi made a successful bid to the Global Fund for AIDS, Tuberculosis and Malaria, to receive funds in support of a national response to HIV, including the purchase of antiretroviral drugs. Prior to scale up, in only 1,220 patients were receiving ART at KCH and QECH, Malawi's two central hospitals in Lilongwe and Blantyre, respectively, and a rural district hospital in Chiradzulu[97]. Over the following years, the Ministry of Health and Population and the National AIDS Commission implemented nationwide scale-up of ART, beginning in the Northern districts and working towards the South[97]. Since Malawi began implementing free ART in 2004, ART delivery rose from 2-3% to 67% coverage among eligible HIV patients in 2014 [14, 98], representing more than a half million people who are now receiving therapy. Latest estimates show over 700 ART clinics in operation throughout the country (Figure 2-2). ART programs in Malawi now include life-long, universal access of ART for pregnant women *Option B+*, lay counselors for HIV testing, and simplified schemes for viral load monitoring.

Figure 2-2. Trends in ART coverage, number of ART clinics, and ART guideline criteria during pre-ART and post-ART national scale-up in Malawi



* Data on the number of ART clinics were approximated for year 2012. Regrettably, the exact number of clinics were not retrievable for 2012.
2002: Malawi bids to the Global Fund for AIDS, Tuberculosis and Malaria (http://www.theglobalfund.org/en/) The Ministry of Health and Population, with the assistance of the National AIDS Commission, prepares to scale-up ART.
2004: ART national scale-up is initialized
2003 to 2006: national ART eligibility: WHO stage 3/4 or CD4-lymphocytes<200/mm^3 or WHO stage 2 with total lymphocyte count (TLC) <1200 cells/µl (Reference: Treatment of Aids Guidelines for the Use of Antiretroviral Therapy in Malawi. First Edition: September 2003. National AIDS Commission of Malawi and Ministry of Health and Population, Malawi)
2006: ART eligibility raised to CD4<250 cell/mm^3
2008: ART eligibility for adults >15 years: HIV+ and WHO clinical stage 3/4 or CD4-lymphocyte count <250/mm^3 or WHO clinical stage 2 with TLC 1200/mm^3. ART eligibility for Children ≤14 years: HIV+ and WHO pediatric clinical stage3/4 or CD4-lymphocyte percent<threshold or WHO pediatric stage 2 with TLC< threshold (Reference: Guidelines for the Use of Antiretroviral Therapy in Malawi, Third Edition: April 2008. Ministry of Health, Malawi)
July 2011: ART eligibility for adults increased the threshold to CD4< 350 cell/mm^3. Option B+ made ART available to all HIV+ pregnant women for the rest of their lives, regardless of CD4 count or clinical staging (Reference: Clinical Management of HIV in Children and Adults. Malawi Integrated Guidelines For Providing HIV In Antenatal Care, Maternity Care, Under 5 Clinics, Family Planning Clinics, Exposed Infant/Pre-ART Clinics, ART Clinics. First Edition. July 2011. Ministry of Health, Malawi)
April 2014: new guidelines for HIV clinical management require ART for all HIV+ children < 5years; HIV+ children >5 years or adults with CD4<500 or those co-infected with Hepatitis B. Therefore the number of children and adults requiring ART increased from 681,000 to 798,000 (Reference: Malawi AIDS Response Progress Report 2015. April 2015)

E. Cancer burden among the HIV/AIDS population in sub-Saharan Africa

Cancer is a public health concern in Africa. Non-communicable diseases now feature among the leading causes of death. After heart disease, cancer is the 7th leading causes of mortality, accounting for approximately half a million deaths annually, or 4% of total mortality in SSA[99]. HIV/AIDS continues to be the leading cause of death in SSA, accounting for 1.7 million deaths (14.3%). The future cancer burden in Africa is expected to double to 1.3 million new cases and 970,000 deaths by 2030 [83], a trend that is strongly driven by aging of the population, population growth, and the adoption of Western risk factors including smoking, physical inactivity, obesity, and diet[1, 100]. Population projections expect the demographic over the age of 65 will account for 4.5% of the population structure in 2030 and 10% by 2050; simultaneously, unprecedented growth in the youth population is predicted[101, 102]. By 2030, the population structure in many African regions will resemble that of industrialized nations[102]. Against this backdrop, it is increasingly relevant to adopt cancer control programs that are tailored to countries with low resources[103].

The current cancer burden in SSA is remarkably different from that of industrialized countries in terms of the distribution of cancer types, stage at diagnosis, incidence, mortality, and patient survival[2]. HIV-associated cancers represent a high burden in Eastern and Southern SSA as a direct reflection of the rampant HIV epidemic (Figure 2-3, Figure 2-4, Figure 2-5). Three AIDS-defining malignancies, KS, cervical cancer, and NHL are among the top 10 types of cancers presenting in the region[1, 83]. In Malawi, KS and cervical cancer are the most common cancers among men and women, respectively[104]. It is recognized that the HIV epidemic and cancer burden are interconnected in SSA, yet African data from the field of population sciences is sorely needed to guide cancer prevention measures targeting persons living with HIV/AIDS.

Population-based cancer data required for evidence-based cancer control programs in SSA is direly lacking[105-107]. Only 1% of Africa's population is covered by 5 population-based cancer registries that meet IARC's data quality standards of completeness, validity, and timeliness in reporting[20]. Routine cancer surveillance is crucial for informing evidence-based cancer control, guiding health policy, and planning the allocation of clinical care. Pathology services are considered the backbone of cancer surveillance, yet population coverage by pathology services in Sub-Saharan Africa is only one-tenth that in resource-

replete settings[108]. Limited health infrastructure and cancer diagnostic services further hinder the process of formally and completely ascertaining cancer incidence. Lastly, because HIV is not a reportable disease in many African countries and due to the local poor health infrastructure, HIV status is not routinely collected by cancer registries, or other public health registries. At the same time, cancer outcomes are not routinely collected by ART clinics, with the exception of KS and cervical cancer, which are, by definition, clinical indications of stage 4 HIV/AIDS.

Figure 2-3. Incidence of Kaposi Sarcoma in Africa, by country and sex. IARC GLOBOCAN 2012

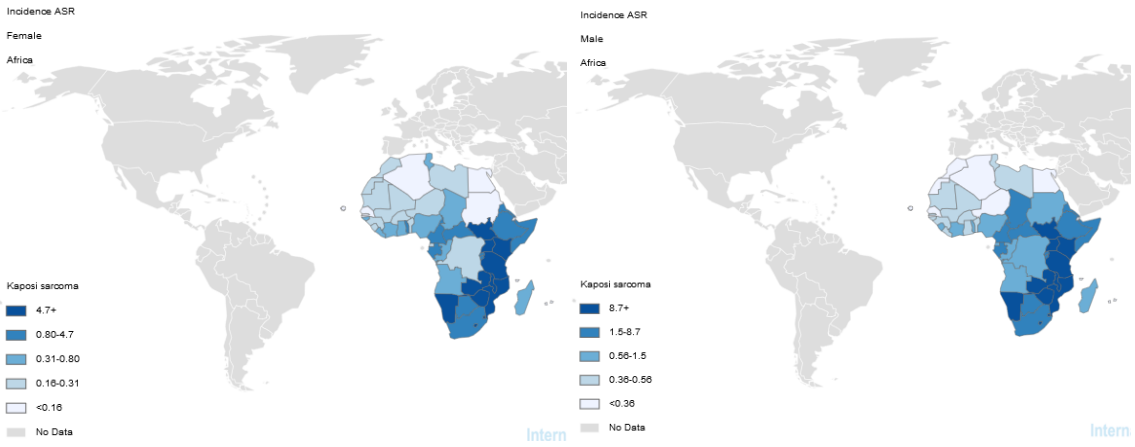


Figure 2-4. Incidence of cervical cancer in Africa, by country. IARC GLOBOCAN 2012

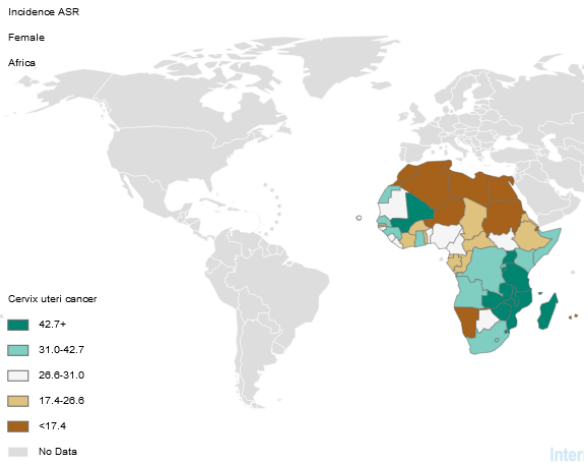
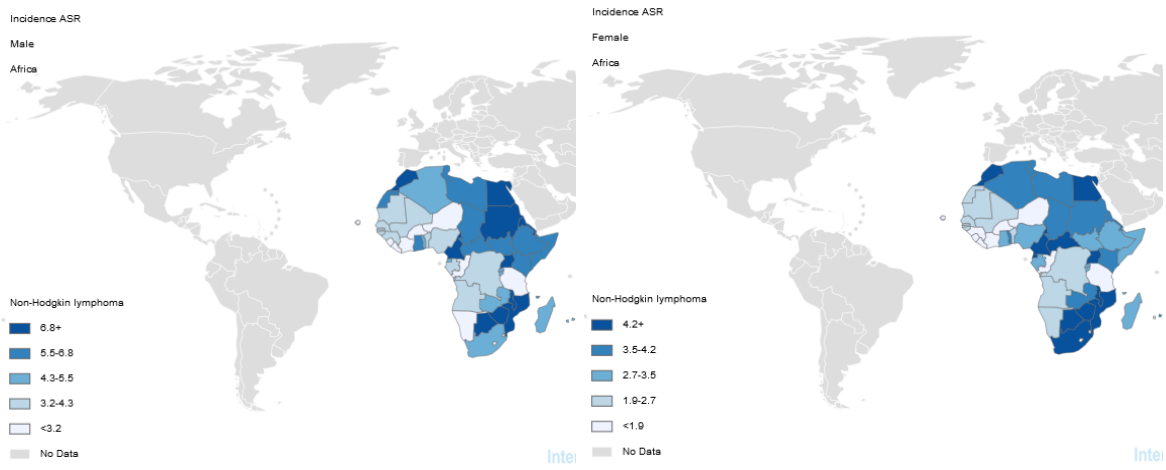


Figure 2-5. Incidence of non-Hodgkin lymphoma in Africa, by country and sex. IARC GLOBOCAN 2012



To overcome these logistical setbacks, ecologic studies, HIV-cancer record linkages, and registry-based descriptive studies have been the main epidemiological approaches to assess the impact of ART scale-up on recent cancer incidence and to assess the cancer burden among persons living with HIV/AIDS in select SSA countries (Table 2-4). Two recent ecologic studies in Uganda[13] and Botswana[11, 109] sought to examine the impact of recent ART scale-up on cancer incidence rates and to estimate the proportion of cancer cases living with HIV. By using a subsample of the registry with known HIV status, the Botswana study used inverse probability weighting to assign HIV status among all cancer cases in the registry. The Uganda study used cancer outcomes classified by ADC and NADC since HIV status was unknown among cancer cases. In Botswana, 61% of the total cancer cases in the registry occurred among HIV-positive people. As ART coverage increased from 7% to 82% over a five year period, the incidence rate of all cancers among HIV-positive cases declined by 8% per year. The incidence of KS among HIV-positive cases declined by an estimated 5% per year while NHL and HPV-associated cancers increased by 11.5% and 4% per year, respectively. Concurrently, the overall cancer incidence among HIV-negative cases increased by 7.5% per year. Using the population attributable fraction of HIV, approximately 45% of cancer among men and 36% of cancer among women was attributable to HIV in Botswana. Similarly, Uganda found only a modest decline in KS incidence during the 10 year period of ART scale-up: for every 10% increase in ART coverage, KS incidence in the overall general population fell by 5% per year. Among the other ADC, cervical cancer incidence did not significantly change and, after excluding predominantly pediatric cases of Burkitt's lymphoma, which is not HIV/AIDS related, the

incidence of NHL did not change. Stomach cancer, an infection-related cancer caused by *H. pylori*, declined by 12%. Among virally-associated cancers, liver cancer incidence increased by 12%. Prostate cancers, which are two epithelial, non-virally-associated NADC, increased by 5% per year. Together, the estimated impact of ART on cancer incidence in Uganda and Botswana was not as robust compared to the United States[10, 23]. For perspective, cancer incidence among the HIV population in the United States declined sharply by 29% per year for KS and 16% for NHL during the early HAART era, and subsequently by 7.8% and 5.5% those cancers[10]. Cervical cancer and Hodgkin lymphoma declined consistently by 11% and 4% per year. Of note, breast cancer incidence among the HIV population did not change during that time in the US[10].

There are several possible explanations for the difference in magnitude between African and Western cancer trends in the HIV/AIDS population during the ART era. The first is an issue of the denominator used in rate calculations. Analytically, the African studies used the overall population as the denominator rather than the HIV population specifically which may have resulted in an attenuation of estimated incidence trends compared to the US. Differences in the population used for age and sex standardization also prohibit direct comparison of rates between the studies. Second, the threshold of CD4 cell counts was lower (<200 cell/μl) in African settings, resulting in ART initiation at advanced stage of disease[110]. Third, modest declines in the incidence of KS among total cases in Uganda may be partially attributable to the high background rate of endemic, HIV-negative KS in the overall population, which accounted for up to 10% of total KS prior to the HIV epidemic[111, 112]. Endemic KS has been described geographically as a “KS belt” extending from Cameroon through the Democratic Republic of Congo, and through the Rift Valley including Uganda, Tanzania, Malawi and Zambia[113]. Fourth, at least in the Ugandan setting, the lower ART coverage translates to a substantial proportion of the HIV population still at high risk for KS; additionally the less frequent use of protease inhibitors during early ART scale-up may have resulted in less effective HIV viral control. Fifth, the latency period for cervical cancer is long and it remains somewhat uncertain at which points restored immunity from ART is impacting the process of HPV-associated carcinogenesis. Low CD4 cell counts are associated with the persistence of HPV, HPV viral load and precancerous cervical lesions[44] , however partially restored immunity conferred by ART may not be sufficient to stop tumor development once a malignant tumor is initiated[114]. Furthermore, in

the Uganda study, high background rates of HIV-negative cervical cancer in the region due to limited screening may dilute the true impact of ART scale-up when using incidence rates from the general population of HIV-positive and HIV-negative people.

Record linkage studies have the advantages of characterizing rates using patient level data, and as such circumvent ecologic fallacy, overcoming limitations of unknown HIV status among cancer registry cases, and explicitly enumerating the denominator of HIV patients at risk for cancer in incidence rate calculations. The Uganda AIDS-Cancer Registry Match Study is a population-based study conducted prior to the introduction of ART[12]. More recently, South Africa[115, 116] and Nigeria[117] have completed HIV-AIDS cancer match studies during the era of contemporary ART delivery (Table 2-4). Findings affirm high ADC incidence rates among the HIV population for KS (Uganda: 240, South Africa: 307 per 100,000 person-years), cervical cancer (Uganda: 70, South Africa: 447 per 100,000 person-years), and NHL (Uganda: 19 per 100,000 person-years, South Africa: not reported). However corresponding excess risk of these cancers relative to the general population, presumably uninfected with HIV, was lower compared to those observed in Western countries. Incidence rate ratios in these two studies may be attenuated towards the null in part due to under ascertainment of cancer cases, and more importantly, due to high HIV prevalence in the general population, thereby attenuating results towards the null despite analytic corrections. There was no excess risk for liver or penile cancer, virally-associated cancers, despite their prevalence in Uganda. High rates of squamous cell carcinoma of the conjunctiva among the HIV population were reported in South Africa (men: 58, women: 60 per 100,000), as well as excess risk in Uganda (standardized incidence ratio 4.0); together the findings are concordant with early case-control reports on this unique association with HIV within the region [118, 119], and later confirmed among persons with HIV/AIDS in western countries[120] . While still preliminary, Nigerian results show incidence rates and excess risk for ADC among the HIV/AIDS population that were lower by orders of magnitude, particularly for KS (4.9 per 100,000 person-years), which may possibly be due to low sensitivity of the record linkage itself rather than distinct etiologic circumstances from the rest of the region. Further results have yet to be released.

Table 2-4. Population-level studies of the cancer burden among the HIV/AIDS population and trends during ART expansion in sub-Saharan Africa

Study/Country Time Period/ % HIV Country Prevalence	Research Objective Context	Study Population/ Methods	Results	Strengths/ Limitations
Ecologic studies				
Dryden-Peterson et al. 2015 Botswana 2003-2008 25%	Objective: To assess impact of ART scale-up on cancer incidence trends during 5-year period Context: ART eligibility was CD4 <200 cells/μl or WHO stage 3/4. ART increased from 7.3% to 82.3% over the study period	Study population: Botswana Cancer Registry, with inverse probability weighting for HIV status Methods: ecologic study. Trends analysis of cancer incidence using ART coverage (or calendar year) as ecologic variable in Poisson regression	▪61% of cancer occurred among HIV+ ▪45% of cancer in men were attributable to HIV; 36% in women ▪among HIV+, overall cancer incidence rate ↓ 8.3% per year (95%CI -14.1, -2.1) ▪ KS incidence rate ↓4.6% per year (95%CI -6.9, -2.2) ▪NHL ↑11.5%(95%CI 6.3, 17.0) ▪HPV-associated cancers ↑3.9% (95%CI 1.4, 6.5) ▪among HIV-, overall cancer incidence rate ↑7.5% per year (95%CI 1.4, 15.2) ▪annual number of total cases did not significantly increase	Strengths: ▪Cancer registry captures >85% cancer cases ▪Use of IPTW to account for the large proportion of missing HIV status in the registry, including a sensitivity analysis of IPTW estimates Limitations: ▪Cases not receiving oncology diagnosis and care at referral facilities may be under-represented in the registry, particularly KS treated with ART alone ▪Consistent with other SSA registries routine linkage with HIV registries is not in place ▪patient level ART status, follow-up not available
Mutyaba et al. 2015 Uganda 1999-2008 6.7% in 2004-2005	Objective: To assess impact of ART scale-up on cancer incidence trends during 10-year study period Context: ART coverage increased from 0% to 43% over study period	Study population: Kampala Cancer Registry Methods: Ecologic study. Trends analysis of cancer incidence using ART coverage as ecologic variable in Poisson and negative binomial models	For every 10%↑ in ART coverage: ▪KS incidence ↓5% (IRR 0.96, 95%CI 0.91, 0.99) ▪stomach cancer incidence ↓ 13% (IRR 0.80, 95%CI 0.80, 0.95) ▪liver cancer incidence ↑ 12% (IRR 1.12, 95%CI 1.01, 1.21)	Strengths: ▪Historic, well-established registry covering 90% of population in Kyadondo county; registry is included in IARC Globocan. ▪Conducted an analysis restricted to pathologically confirmed cancer cases (the results were similar) Limitations: ▪possible under-reporting of Hodgkin lymphoma in the

Study/Country Time Period/ % HIV Country Prevalence	Research Objective Context	Study Population/ Methods	Results	Strengths/ Limitations
			▪prostate cancer incidence↑ 5% (IRR 1.05, 95%CI 1.0, 1.10) ▪breast cancer incidence↑ 5% (IRR 1.05, 95%CI 1.0, 1.11) ▪no change in incidence of NHL ▪no change in Hodgkin's, cervical, lung or colon cancer	cancer registry may be associated with null trend ▪time period may be too short to observe a change in the incidence of cervical cancer, which has a long latency period
Record linkage studies				
Sengayi et al. 2016 (CROI) South Africa 2004-2011 17%	Objective: To estimate cancer incidence rates, risk factors for cancer	Study population: Two ART cohorts in KwaZulu-Natal, Themba Lethu (N=23,120) Methods: Probabilistic record linkage between ART cohorts and cancer registry. Cancer incidence and risk factors were estimated from Cox regression models, using sex, age, CD4 counts and hemoglobin levels at ART initiation	▪overall cancer incidence was 1,315 per 100,000 person-years (95%CI 1225, 1410) ▪cervical cancer incidence was the highest: 447 per 100,000 py (95%CI 413, 551) ▪KS incidence was 307 (95%CI 206, 355) ▪breast cancer incidence was 159 (95%CI 124, 204) ▪CD4<100 was associated with higher infection-related cancer incidence compared to CD4>350 (HR=0.24) ▪ Later age at ART initiation was associated with non-infection related cancer incidence (HR=2.63 comparing >56 versus 16-25)	Strengths: ▪use of strong methods to link cancer outcomes to ART cohorts ▪large contemporary ART cohorts with baseline laboratory data Limitations: ▪accuracy of the record linkage is not known ▪time-varying CD4 may not have been available from the cohort's lab data

Study/Country Time Period/ % HIV Country Prevalence	Research Objective Context	Study Population/ Methods	Results	Strengths/ Limitations
Mbulaiteye et al. 2006 Uganda 1999-2002 8%	Objective: To estimate cancer incidence rates among a sample of the HIV/AIDS population. Estimate cancer risk relative to the general, HIV-uninfected population. Compare the timing of cancer incidence across early and late incident cancers Context: pre-ART era; 10%-15% HIV prevalence in Kyandondo county	Study population: HIV/AIDS patients across all stages of disease registered in referral center for AIDS support organization living near Kampala (n>15,000). Methods: Probabilistic record linkage used to match cancer outcomes recorded by Kampala cancer Registry with HIV patients. Cancer incidence and incidence rate ratios were estimated from Poisson regression, using sex, age, WHO stage, category of follow-up time	▪70% of cancers were ADC ▪ Within 2 years post-enrollment in the HIV/AIDS cohort, risk of cancer was increased relative to the general population for KS (SIR 6.4), NHL (SIR 6.7), and cervical cancer (SIR 2.4). ▪Within the 5 years post enrollment, risk was increased for Hodgkin's (SIR 5.7), conjunctiva (SIR 4.0), kidney (SIR 16), uterus (SIR 5.5), thyroid (SIR 5.7)	Strengths: ▪AIDS support organization covers 7 districts, large sample size ▪Historic, well-established registry covering >90% of population in Kyadondo county; registry is included in IARC Globocan. ▪used KS rates prior to the HIV epidemic as a comparison in SIRs to prevent underestimation of SIRs caused by high prevalence in general population Limitations: ▪possible under-ascertainment of KS ▪does not have follow-up data, therefore assumed 60 months follow-up after registration, which may result in an overestimation of time at risk for cancer
Akarolo-Anthony et al. 2014 Nigeria 2009-2012 4.2% in 2010 3.3% in 2011	Objective: To estimate cancer incidence rates among a sample of the HIV/AIDS population. Estimate cancer risk relative to the general, HIV-uninfected population. Context: 24%ART coverage; 8.6% HIV prevalence in Abuja	Study population: cohort of HIV/AIDS patients (n=17,826) Methods: Probabilistic record linkage used to match cancer outcomes recorded by Abuja cancer registry (n=2,029) with HIV patients. Observed cancer incidence rates were calculated	▪low rates of KS (4.9 per 100,000 py) and cervical cancer (7.8 per 100,000 py) ▪low incidence rates of NADC: liver (1.8), breast (1.6) ovary (3.6) eye (1.5), anus (0.3), non-epithelial skin cancers (1.8) ▪Relatively low increased risk of KS (SIR 5.0 95%CI 4.1, 7.2) ▪Risk of cervical cancer (SIR 2.0, 95%CI 0.4, 3.5) was not significantly increased relative to the general	Strengths: ▪record linkage is a robust and efficient approach to incorporate cancer outcomes in the HIV cohort Limitations: ▪incomplete ascertainment of cases by the cancer registry ▪NHL and NADC may be difficult to detect due to small sample size ▪Incidence rate ratios approximating risk, may be attenuated towards the null due to back ground rate of HIV prevalence in the general population

Study/Country Time Period/ % HIV Country Prevalence	Research Objective Context	Study Population/ Methods	Results	Strengths/ Limitations
			population and cervical cancer among HIV/AIDS patients compared to the general population ▪non ▪NHL not reported in this cohort	
Cancer registry studies (selected for direct relevance and quality)				
Msyamboza et al. 2012 Malawi 2000-2010 10.6%	Objective: To estimate cancer incidence rates and distribution of common types among the total population covered by the registry Context: 2007-2010 expansion to nationwide coverage; 2000-2007 surveillance covered Blantyre only, in the south	Study population: cancer patients ascertained by population-based registration Methods: population-based, cross-sectional survey of private, district-level (secondary) and central hospitals (tertiary) across all districts. Observed cancer prevalence, incidence rates, linear incidence trends were calculated	▪KS, cervical cancer, and esophageal cancer are the most common sites ▪Incidence rates of KS (31.1 per 100,000) and cervical cancer (25.4 per 100,000) continued to increase during the period of ART scale-up	Strengths: ▪near complete population coverage of hospitals only Limitations: ▪mostly clinical diagnosis and therefore limited pathologic confirmation: 18% laboratory verified diagnoses Concerning under-reporting of cancer cases: ▪Registry does not collect ADC from ART clinics, therefore, it is likely that incidence rates of KS and cervical cancer are underestimated, but the extent is hard to predict ▪ liver, bladder cancers, lymphomas are likely underreported due to lack of diagnostic capacity

Study/Country Time Period/ % HIV Country Prevalence	Research Objective Context	Study Population/ Methods	Results	Strengths/ Limitations
Parkin et al. 2010 Uganda 1991-2006	Objective: To estimate cancer incidence rates among the total population covered by the registry Context: the time period covers pre-ART and post ART roll-out eras. 40% coverage HAART	Study population: cancer patients ascertained by population-based registration Methods: population-based surveillance, compared incidence rates across 1991-1995, 1996-2001, 2002-2006; average annual percent change over the entire time period	▪4.5% per year ↑ in incidence of breast and prostate cancers; esophagus is constant ▪prostate cancer is now most common cancer in men ▪↑ 3.0% per year (95%CI 0.9, 5.1) incidence rate of cervical cancer ▪↓ 2.8% per year incidence of KS in men, but non-significant ↑1.4% (95%CI -0.5, 3.2) among women with later age at diagnosis compared to earlier period ▪↓ ~30% incidence pediatric KS ▪↓ squamous cell ca. conjunctiva since mid-1990s	Strengths: ▪near complete population coverage Limitations: ▪mostly clinical diagnosis and therefore limited pathologic confirmation: 18% laboratory verified diagnoses Concerning under-reporting of cancer cases: ▪Registry does not collect ADC from ART clinics, therefore, it is likely that incidence rates of KS and cervical cancer are underestimated, but the extent is hard to predict liver, bladder cancers, lymphomas are likely underreported due to lack of diagnostic capacity

In prior descriptive reports by the Malawi Cancer Registry[17] and Uganda Cancer Registry[121] KS incidence rates remained high since the introduction of HAART. Though treatment coverage and time periods analyzed by the registries do differ, 2007-2010 in Malawi and 2002-2006 in Uganda, the findings underscore a residual high burden of ADC in the overall population despite ART roll-out initiatives. The incidence rate of KS among men in Malawi was 25.4 per 100,000 and 27.9 in Uganda; KS among women in Malawi was 11.9 and 20.1 in Uganda. The rate of cervical cancer also remained high: 33.6 in Malawi and 38.6 in Uganda, rates that are 3.5 to 4 times higher than in the Unites States[82]. In Malawi, KS incidence increased greater than 2-fold among both men and women between 2000-2003 (pre-ART roll-out) and 2007-2010, while cervical cancer increased 3.5 fold. The incidence of NHL in Malawi was 2.5, one- eighth the rate in the United States, and likely represents an underestimate of the true burden due to constraints in laboratory services for this specific cancer diagnosis.

It is worth mentioning that the methods used in the registry reports were not designed to directly assess the impact of ART. An ecologic study using ART coverage over calendar time and analytic methods to account for *non-linear* trends[10, 122] would be better suited to address this question specifically in Malawi. Furthermore, the increase in incidence rates reported in the Malawi study may be an inadvertent artifact of analytically combining multiple geographic areas covered by new surveys during 2007-2010. A last limitation to interpreting the temporal increase in incidence of ADC in Malawi is that the majority of KS is diagnosed in ART clinics, where it is used as a criteria for WHO clinical staging. ART clinics are not covered by the Malawi registry's cancer surveys, only hospitals are, therefore estimates of national KS incidence are likely to be underestimated. The 2000-2003 period of pre-ART scale-up is probably less subject to under ascertainment of KS since cancer surveillance covered only Blanytyre, which housed one of three centers nationwide where ART was available at the time. With these limitations, many questions remain why ADC appear to have increased in Malawi as ART was introduced, but due to the aforementioned constraints in healthcare systems, partitioning cancer incidence trends among HIV-positive cases has not been possible to date. With this in mind, we propose the Malawi HIV-Cancer Match Study to more definitively characterize the cancer burden among patients receiving ART.

F. Overview of the Malawi HIV-Cancer Match Study

The Malawi HIV-Cancer Record Linkage Study aims to develop an epidemiological resource for studying HIV-associated cancer in SSA. The study involves research collaborators and data from the Malawi cancer registry and International epidemiologic Databases to Evaluate AIDS (IeDEA) network. We will use probabilistic record linkage of ART cohorts and cancer outcomes. Record linkage is versatile tool for efficiently merging information across healthcare databases[123-125] where it is commonly used to supplement routinely collected information on health outcomes, laboratory findings, and risk exposures. Our study will be similar to the HIV-AIDS Cancer Match Study[22] of the National Cancer Institute in the United States, but trends in HIV-associated cancer incidence and patient outcomes in Malawi are likely to differ substantially from Western countries.

Our study will use existing secondary data from the cancer registry and from two ART cohorts in Malawi's largest cities. All patients enrolled at the Queen Elizabeth Central Hospital's ART clinic in Blantyre and the Lighthouse Trust HIV clinic in Lilongwe will be included. The Malawi Cancer Registry is a national population-based registry, a founding member of the African Cancer Registry Network, and one of only a few cancer registries from sub-Saharan Africa currently included in the WHO *Cancer Incidence on Five Continents* monograph[17, 20]. Our design incorporates active longitudinal follow-up of HIV-patients that will be used to assess risk factors associated with developing cancer following ART enrollment.

G. Summary and public health significance

The African cancer burden is not static. Over the next 15 years, demographic transitions and the adoption of western risk factors will contribute to a doubling of the cancer burden that many countries in the region are not adequately prepared to handle. At present, many countries in SSA continue to shoulder a high burden of ADC as a result of the HIV epidemic.

Early initiation of ART and adherence to therapy may afford the opportunity for cancer prevention in African countries with high HIV prevalence. Resource rich countries have seen dramatic declines in KS since the early days of ART, even in the era prior to HAART. Further epidemiological studies may address the question whether such trends could be achieved in SSA where KS continues to be among the most common cancers, if not the most common cancer in many countries. Addressing the burden of

NADC is also a priority, especially those caused by chronic viral infections, since risk remains elevated among the HIV/AIDS population in spite of immune reconstitution conferred by ART.

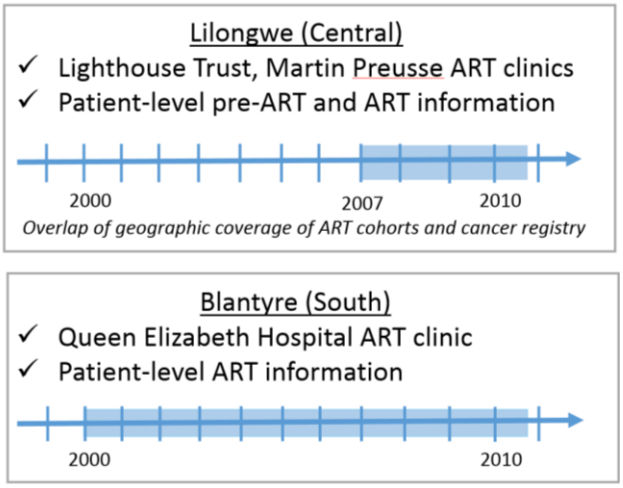
Our study will contribute to a more thorough understanding of cancer incidence and incidence trends among ART initiators in Malawi, while overcoming limitations from previous studies in the region. Our study circumvents limitations from previous ecologic study designs by using patient-level characteristics, including clinical factors, drug regimens, and duration of ART. The HIV cohorts are representative of the African context where ART delivery typically occurs at low CD4 thresholds and losses to follow-up are non-trivial. The record linkage approach is robust and cost-effective; it is a sustainable health systems approach that may be adaptable to other low- and middle-income countries. We will assess the performance of this methodology in a resource constrained setting taking into account constraints due to missing data and ethnographical considerations specific to Africa.

CHAPTER 3. RESEARCH DESIGN AND METHODS

A. Study design

The Malawi HIV-Cancer Match Study is comprised of 2 observational cohorts of HIV patients receiving ART in the central city of Lilongwe and southern city of Blantyre. Cancer outcomes were ascertained two ways: as part of the clinical work up for WHO staging during baseline enrollment in the ART cohorts, and cross-sectionally through linkage of clinic records with the Malawi Cancer Registry over the periods of geographic coverage by the registry (Figure 3-1).

Figure 3-1. Geographic and temporal coverage of cancer by the cancer registry in relation to the ART cohorts



B. Study populations

ART cohorts

In Malawi, ART is primarily delivered through public sector hospitals and clinics. Since the 2004 implementation of free ART to all eligible HIV-infected persons in Malawi, nationwide coverage has grown from 3,000 HIV-infected persons receiving ART to greater than a half million in 2014. Currently, 67% of the HIV population meeting ART-guideline eligibility are receiving ART.[14]

Data for this record linkage is derived from two large, well-established ART facilities in the central and southern regions of Malawi (Figure 3-1). In the central region, Lighthouse Trust clinic (LT), located adjacent to at Kamuzu Central Hospital, Lilongwe is the largest public ART provider. By the end of 2010, LT delivered ART to 5,335 HIV-infected persons, of which 3,959 were still alive and on treatment.[126] Queen Elizabeth Central Hospital (QECH) is a tertiary hospital in Blantyre and the largest referral center in the Southern Region; by the end of 2010 QECH delivered ART to 7,100 HIV-infected persons (Table 3-1). QECH began as a fee-for service clinic in 2000 and since 2004 has provided free ART under the Ministry of Health ART scale-up plan.[127]

Table 3-1. Coverage of ART initiators by city and region

ART clinic	City, Region	Number of patients receiving ART in 2010	2015: Number of new ART initiators, coverage of ART initiators
Lighthouse Trust (LT)	Lilongwe, Central	5,335	N=5,360 36% of new initiators in Lilongwe 18% of all initiators in Central region
Queen Elizabeth Central Hospital (QECH)	Blantyre, South	7,100	N=1,160 10% of new initiators in Blantyre, 2% of al initiators in Southern region

Source: Malawi Ministry of Health Q1 2015 report

LT and QECH use an electronic data system to routinely collect demographics, baseline and follow-up clinical data (WHO stage, HIV viral load, CD4 count), and ART information since January 2005 and 2008, respectively.[126, 127] Historical records of ART initiators were retroactively validated and added to the electronic data system.

Active tracing is used for patient follow-up and ascertainment of vital status. Confirmatory HIV diagnosis and WHO clinical staging are provided at time of entry into care. WHO stages, categorized as 1 through 4, are defined by specific clinical conditions or symptoms. The staging scheme is practical for managing patients in low-income settings with limited laboratory capacity; studies in Africa show strong correlation between WHO staging and CD4 count and total lymphocyte count.[128-131] In Malawi, historically, CD4 counts were restricted to clinical stages 1 and 2 patients who were not clinically eligible for ART, and among these patients, timing of laboratory assays was not uniformly proximal to start of therapy. For QECH prior to 2011, CD4 counts were not captured in the electronic monitoring system and therefore

were not available for analysis. Beginning in 2011, new guidelines were instituted for routinely monitoring HIV RNA in all patients at 6 months, 1 year, and 2 year intervals. Extensive data for CD4 counts and HIV RNA were therefore not available during the time period of our study reflecting practice within the Malawi national HIV program.

Malawi cancer registry

The Malawi cancer registry is a founding member of the African Cancer Registry Network (http://afcrn.org/index.php), which provides training in population-based cancer registration, technical support, and coordination of international research projects. The cancer registry began as a pathology-based cancer registry at QECH in Blantyre in 1985, and subsequently expanded to a national cancer registry in 1989.[17, 132] In 1993, population-based activities for the Blantyre District were expanded. Active case-finding is conducted through cross-sectional surveys of secondary (district-level), tertiary (central-level), and major departments of other public and private hospitals. Sources of information include outpatient departments, inpatient wards, clinics, laboratory, pharmacy, surgery, and mortuaries. Population-based cancer registration was expanded over a catchment area of urban and rural Blantyre during 2001, 2003, and 2005 surveys. Select neighboring districts were also surveyed though case-ascertainment was less complete than in Blantyre. In 2010, a nation-wide population-based survey covering 2007-2010 diagnoses was conducted at 81 out of 84 hospitals. Three international facilities declined participation in the 2010 survey; their population catchment areas are not known.

Cancer diagnoses are collected using the IARC standard procedure manual for cancer registration in SSA. Data are coded using the International Classification of Disease for Oncology (ICD-O).[133] Information on diagnosis and treatment are processed and stored in WHO cancer registry software CanReg4 which checks for duplicates, reviews implausible combinations of codes, and calculates age-standardized incidence rates.[134] Stage of disease is generally limited in the registry as a direct result of limited diagnostic capacity in many hospitals. Systematic vital statistics, including death certificates, are not available and therefore not routinely used as a source of information in the cancer registry. Incidence rate calculations rely on population projections covering 1999-2012 from the National Statistical Office of Malawi (http://www.nsomalawi.mw/). Registry operations are approved by the Malawi National Health Sciences Research and Ethics Committee (NHSRC).

Inclusion and exclusion criteria for record linkage

Additional years of data for QECH and cancer registry datasets were made available to the analytic team. Incorporating the additional years of data does not statistically harm the linkage, therefore all data was deemed eligible for linkage under the assumption that the final analytic set would be trimmed to reflect temporal and geographical overlap between coverage of cancer registry diagnoses and ART cohorts. No restriction on age or cancer type was placed on the record linkage.

All patients enrolled at QECH from January 1, 2000 – October, 1 2015 were eligible regardless of prior cancer diagnosis or length of patient follow-up (N=23,743). Similarly, all patients enrolled at LT from January 1, 2007 – October 1, 2010 were eligible (N=26,977). The entire dataset for the cancer registry from 1985-2010 was used for the record linkage (N=62,944). Among cases with missing district of residence, the district hospital where the case was recorded was used as a proxy for residence, except for cases collected from Kamuzu Central Hospital (Lilongwe), Zomba Central Hospital, or Queen Elizabeth Central Hospital (Blantyre) which are large oncology referral centers with catchment areas spanning the most of the central and southern regions, respectively.

C. Background on probabilistic record linkage

Overview

Our goal is to merge cancer outcomes to two observational cohorts of patients receiving ART in Malawi. Because the datasets are large, conducting a manual linkage of records was unfeasible. We considered two main approaches: deterministic and probabilistic methods for linking the datasets together. Deterministic linkage groups records using unique or non-unique identifiers, such as names and birth dates, and relies on the *exact match* of one or more identifiers. Probabilistic linkage also groups records using non-unique identifiers, but instead uses the *probability* of matching across a set of variables for a given record pair.

Probabilistic methods overcome the setbacks of strict deterministic linkage. In the real world, data entry errors and changes in patient demographics may arise in patient records: changes in civil status and last name, change in residence over time, the use of nicknames and aliases, and data entry errors from handwritten notes. Such errors and changes lead to the failure of deterministic methods which rely on exact matches. In contrast, probabilistic methods handle minor discrepancies across the linkage

variables. A match probability and weight is calculated for each potential record pair across each linkage variable. The match weight is the likelihood that two records truly match given agreement on a set of patient identifiers. A high weight indicates a high degree of similarity between a record pair, and a low weight indicates dissimilarity. Record pairs are then classified by their weights into one of 3 categories using a process known as highest-weight classification, resulting in three outcomes: definite match, possible match requiring review, and non-match. Records with the highest weights above a predetermined weight threshold are categorized as definite matches, while potential matches require review.

Linkage variables

Ideally, a combination of unique identifiers (e.g. Malawi health passport numbers, cell phone numbers) and non-unique identifiers (names, birth dates) are typically used for linking a master dataset with another file. In our study, unique identifiers such as Malawi health passport numbers were not available in the study datasets, therefore we used a set of 5 demographic variables for linking the datasets together: first name, last name, year of birth, sex, and district of residence. The chosen variables are shared across data sources.

Linkage methods overview

The objective is to match pairs of records across datasets through a process of assessing patterns of agreement across the set of linkage variables. The end goal is to produce a final summary score for each record pair separately by considering agreements across the set of 5 linkage variables.

Under the simplest scenario, a pair of records may have exact agreement or disagreement on a given linkage variable. For example, a binary variable sex coded as 0 for male and 1 for female will have only an exact agreement or disagreement. Under a more complex scenario, we also allow for partial agreement across a text field such as name, or numeric field such as year of birth. Partial agreement allows for small discrepancies in text fields. Missing values on a given linkage variable are not scored, and therefore record pairs with one or both missing values do not receive a weight on that linkage variable.

Definitions of *m* and *u* conditional probabilities

Probabilistic record linkage relies on Bayes' theorem.[135, 136] Two conditional probabilities are the basis for constructing a summary weight for each record pair. In the literature of record linkage these probabilities are commonly referred to as ***m probability***, also known as *match probability*, and ***u*** probability, also known as *unmatch probability*.

The *m* probability is the conditional probability of a match on a linkage variable given that two records truly do belong to the same person:

$$m = \text{p (match on link variable | True Record Pair)}$$

Using the linkage variable first name as an example, an m probability of 0.9 is interpreted as the probability that any given pair of records truly belonging to the same person will agree on first name 90% of the time. The 0.1 discrepancy is attributed to data quality issues such as misspelling (MarieJo versus MaryJoe), data entry error (MarieJo versus MarieJane), missing data (.), or a possible actual name change (MarieJo versus Jo).

The *u* probability is the probability of a match on a linkage variable given that two records do not belong to the same person or, randomly matching by chance:

$$u = \text{p (match on link variable | True non-Pair)}$$

The *u* probability depends on the distribution of values for a given linkage variable in the dataset. For example, for the variable sex, one could specify a probability u=0.5 or use the observed distribution of sex from study data. In Lighthouse Trust, the u probability for men is 0.422 and 0.578 for women. Similarly, for the linkage variable last name, each unique name has a u probability which is specific to the study and local context, such as geography. In the Lighthouse Trust, located in central Malawi, the u probability for last name 'Banda' is 0.0419 and 0.038 for 'Phiri'. At Queen Elizabeth Central Hospital, located in the southern region, the u probability for 'Banda' is 0.0218 and 0.018 for 'Phiri'.

More intuitively for epidemiologists, *m* and *u* may be conceptualized through a 2-by-2 table of record match status from linkage algorithms compared to the true match status measured from a hypothetical gold standard dataset (Table 3-2). Here, the *m* probability is analogous to sensitivity, while the *u* probability is analogous to 1-specificity:

m = p (match on link variable | True Record Pair)

m or *sensitivity* = True Match / (True Match+ False non-Match)

u = p (match on link variable | True non-Pair*)*

u or *1-specificity* = False Match / (False Match+ True non-Match)

Table 3-2. Concepts of *m* and *u* probabilities illustrated with an epidemiology 2-by-2 table

		Truth (unobserved)	
		Records belong to the same person (True Record Pair)	**Records do <u>not</u> belong to the same person (True Non-Pair)**
Probabilistic linkage outcomes	**HIV clinic record matches to cancer registry ('Match')**	True Match	False Match
	HIV clinic record does <u>not</u> match to cancer registry ('non-Match')	False non-Match	True non-Match

<u>Likelihood ratios.</u> The overarching goal of constructing linkage weights is to assess the performance of the match algorithms in identifying true cancer cases. The goal may be viewed as analogous to the process of measuring validity of clinical diagnostic testing. In diagnostic testing, the likelihood ratio for a positive test is the ratio of the probability a correct result to the probability of an incorrect result [137] :

$$\frac{\text{sensitivity}}{1-\text{specificity}}$$

The likelihood ratio used to assess a positive result from matching algorithms on a given variable is therefore: m / u. In our study, the likelihood ratios were logarithmically transformed to handle skewness.

<u>Constructing the weights using Bayes' Theorem.</u> For epidemiologists, the process of constructing linkage weights may also be understood by working through Bayes' theorem. Using a simple notation for Bayes' theorem, event B is the event that two records truly belong to the same person and p(B) is the probability of the event. The complement event $\overline{B}$ is the event that two records truly do not belong to the same person, and p($\overline{B}$) is the probability of the event.

A1 through A5 are the events of matching values on each of the 5 linkage variables first name (A1), last name (A2), year of birth (A3), sex (A4), residence (A5), and p(A1) through p(A5) are the probabilities of the corresponding events A1 through A5. Applying this notation to the *m* and *u* probabilistic described in the section above:

m = p (match on link variable | True Record Pair) = p (A1| B)

u = p (match on link variable | True non-Pair)= p (A1| $\overline{B}$)

Recalling our goal to obtain a likelihood ratio for a positive match on a set of identifiers, the conditional probabilities of a true record pair given agreement on variable A1 and the conditional probabilities of a true non-pair given agreement on variable A1 are:

$$p(B|A1) = \frac{p(A|B)\cdot p(B)}{p(A)} \quad \text{and} \quad p(\overline{B}|A1) = \frac{p(A|\overline{B})\cdot p(\overline{B})}{p(A)}$$

Manipulating the above probabilities to obtain the odds of getting a true record pair to a true non-record pair when we match on A1 provides the epidemiological terms posterior odds, likelihood ratio, and prior odds[138] :

$$\frac{p(B|A1)}{p(\overline{B}|A1)} = \frac{p(A1|B)}{p(A1|\overline{B})} \cdot \frac{p(B)}{p(\overline{B})}$$

posterior odds of a true pair given agreement on variable A1	likelihood ratio *m* / *u*	prior odds of a true record pair

A major assumption used a probabilistic linkage approach is conditional independence of events across the set of 5 linkage variables. For example, we may assume that two records matching on first name is independent of the same two records matching on year of birth given the records truly belong to the same person. We assume that events A1 and A2, A1 and A3, A1 and A4 and so forth, are independent and expand the calculation of likelihood ratios for the full set of 5 variables:

$$\frac{p(B|A1,A2,A3,A4,A5)}{p(\overline{B}|A1,A2,A3,A4,A5)} = \frac{p(A1|B)}{p(A1|\overline{B})} \cdot \frac{p(A2|B)}{p(A2|\overline{B})} \cdot \frac{p(A3|B)}{p(A3|\overline{B})} \cdot \frac{p(A4|B)}{p(A4|\overline{B})} \cdot \frac{p(A5|B)}{p(A5|\overline{B})} \cdot \frac{p(B)}{p(\overline{B})}$$

The final step is logarithmically transforming all terms so that posterior log odds of a true record pair, given agreement across A1 through A5 variables, is the sum of the log likelihood ratios for each variable A1 through A5 plus the prior log odds of a true record pair:

$$\log \frac{p(B|A1,A2,A3,A4,A5)}{p(\overline{B}|A1,A2,A3,A4,A5)} =$$

$$\log \frac{p(A1|B)}{p(A1|\overline{B})} + \log \frac{p(A2|B)}{p(A2|\overline{B})} + \log \frac{p(A3|B)}{p(A3|\overline{B})} + \log \frac{p(A4|B)}{p(A4|\overline{B})} + \log \frac{p(A5|B)}{p(A5|\overline{B})} + \log \frac{p(B)}{p(\overline{B})}$$

In our study, we assume that the prior probability of a match is small, and consequently the prior odds will also be small. We expected that a small proportion of all HIV patients in the cohort will develop cancer; we did not expect a 1:1 match of HIV records to the cancer registry.

Record classification

Finally, highest-weight classification is used to classify record pairs into matches, non-matches, and potential matches requiring further adjudication. The weight thresholds used during the classification process are specific both to the data at hand and the purpose of the research. Record linkages are typically run iteratively to refine weight thresholds.[136] Our study team decided on the weight thresholds after manually reviewing data quality of the record matches and the range of weights observed in our study. We choose weights equal to or greater than 23 as definite matches, and weights between 12 and 23 as potential matches. A weight less than 23 in our study is the point at which data quality issues or missing data introduce uncertainty, and extra clinical information from the HIV medical record or the cancer registry is required to make a final determination.

Outcome misclassification in probabilistic record linkage

Inherent in any probabilistic method is a degree of uncertainty in the match versus non-match outcome [139, 140], a problem we will refer to as measurement error. Excessive missing data and lack of discriminatory power among patient identifiers may cause linkage errors that result in false positive matches (records that link erroneously) and false negative matches (records that fail to link).[18, 19]

Subsequent analysis should take these errors into account, but this seldom reported.[140] Even small errors can introduce substantial bias[141], as seen in a Swiss study where missed linkages caused an underestimation of mortality rates.[142]

There are three challenges driving outcome measurement error in probabilistic record linkages. The first is measurement error on match variables, which can be viewed analogously as “exposure variables”. These errors are language and context specific. For example in Malawi, certain letters like ‘r’ and ‘l’ are commonly interchanged in Chichewa, the local Bantu-based language. Additional discrepancies may arise from typographical errors occurring during data entry, and variations in patient demographics, and the use of nicknames. Probabilistic methods are well-suited to handle minor discrepancies such as these across match variables of interest.[136, 143] Second, common surnames and first names present a major challenge because they limit the discriminatory power of linkage on name. For example, the two surnames “Banda” and “Phiri” account for nearly 7% of names in our study population therefore matching on a common surname for these records contributes little to the total summary weight for those records. In contrast, rare names contribute a high weight towards the overall summary weight of the record pair. Common first names and use of aliases during hospital registration also lower the discriminatory power of name as a matching variable. Changes in civil status and surname, and changes in place of residence over time are also problematic for the linkage process since the analyst cannot know *a priori* which record is ‘correct’. Legitimate variations in these demographics are marked as discordances during the match process which in turn lowers the summary match weights for the affected record pairs. Assuming that these record pairs make the cutoff as potential matches, they are clerically reviewed and resolved using additional information from the electronic medical records.

Missing data on match variables is third challenge in using real-world data from a resource-limited environment: fully-complete identifiers including name, date of birth and residence, are available for only a subset of all records. Country-specific circumstances such as lack of vital statistics cause substantial missing data on date of birth. Compounding the problem are healthcare interviewing techniques which introduce further measurement error in date of birth: older patients who do not know their date of birth may be interviewed by healthcare workers using well-known historical events to determine the approximate age if the patient, or worse yet, healthcare workers may estimate age based on physical

appearance. Taken together, measurement error and missing data in match variables may substantially lower summary linkage weights, creating a situation of high uncertainty in the 'true' match status of records.

Several approaches can be used to quantify uncertainty and reduce bias in record linkage studies [140]: i) conduct clerical review of all potential matches, the number of which may be prohibitive in terms of time and cost; ii) conduct validation and evaluation of linkage performance using a gold-standard dataset [144] , if such a dataset exists, and adjust weight thresholds accordingly; iii) employ analytic approaches to incorporate uncertainty into post-linkage analyses, such as Bayesian prior-informed multiple imputation of match weights.[18, 145, 146]

D. Specific Aim 1

Specific Aim 1. Adapt, implement, and evaluate a probabilistic linkage methodology suited to health systems in a resource-limited setting.

Since shared unique identifiers are not available across datasets, a probabilistic approach was used as a first pass approach to identify sets of definite and potential matches. Though ART cohorts and the cancer registry have unique patient ID numbers, a cross-walk between the two does not exist, nor does Malawi use a national ID analogous to a Social Security Number. Given the large number of potential matches that contained missing data, a second pass, labor-intensive approach of manual linkage, followed by clinical review was used to further assign potential matches into definite matches or non-matches. LT and QECH were linked to the cancer registry separately. The study workflow describing each step is summarized in Table 3-3.

Data preprocessing and probabilistic record linkage were performed in KNIME Analytics Platform Version 2.12.1 (Konstanz, Germany)[147], an open-platform data miner, utilizing the K-Link probabilistic linkage plug-in. Data pre- and post-processing were conducted in STATA 14 (Stata Corporation, College Station, Texas) and SAS 9.4 (Cary, North Carolina). Personal identifiers were made available to the linkage analytic team (Horner, Spoërri, Chasimpha). All data were encrypted using TrueCrypt 7.2 on portable drives and BitLocker on Windows 8.1 laptops.

Table 3-3. Summary of workflow in probabilistic record linkage

Step	Description
Evaluate and choose the set of linkage variables	Evaluate the proportion of missing data across potential linkage variables, choose a set of variables that are consistently recorded across datasets to serve as the basis for matching
Pre-Processing	Data harmonization, formatting, correcting errors, de-duplication of datasets, removing non-essential variables from datasets
Blocking	Define selection criteria, blocking, reduce the number of record pairs for comparison
Define rules for each linkage variable	Define weights for each variable, rules for margin of error (e.g. date of cancer diagnosis +/- 5 years of ART initiation)
Define *u* and *m* probabilities	Conditional probabilities of a record pair being a match or non-match are derived using the study data
Run probabilistic linkage	Pre-processed datasets are run through KLINK
Refine weight thresholds	Go through the iterative process of re-running the linkage algorithms to evaluate the performance of linkage weight thresholds; categorize matches and non-matches, assign potential matches for deterministic linkage and clinical review
Construct manual deterministic linkage	Logic rules parse potential matches based on dates, residence, and facility where cancer diagnosis occurred (e.g., date of cancer diagnosis< date of death; cancer diagnosis at district hospital used as a proxy for district of residence when it is missing); deterministic linkage is used to review nicknames and full date of birth when available
Conduct clinical review of potential match pairs	Review of potential cases by 3 Malawian senior clinical investigators (Drs. Dzamalala, Malisita, Masomba) using additional clinical information (e.g. cancer stage and subsequent treatment, age at cancer diagnosis compatible with HIV clinical history)
Post-processing	Relink clinical variables of interest for analysis, remove personal identifiers, create a master linkage key

*All data were encrypted during the record linkage process. Final datasets contain only anonymized data.

The set of linkage variables used for matching were last name, first name, year of birth, sex, and district of residence as these were common identifiers across datasets. Place of residence was recorded as free text in the ART datasets but was determined to be unusable as a linkage variable given the lack of physical street addresses; further, the cancer dataset only documented residence at the district level. Therefore, district was used as the geographic unit of interest for record linkage. Consistent with other African research settings, date of birth was estimated for a substantial proportion of records. Therefore, only the year of birth was used as a linkage variable, allowing for 5-year discrepancies between record pairs during subsequent iterations of the linkage.

Data pre-processing is the most time-consuming and labor intensive step of a data linkage. Sophisticated software packages and algorithms cannot correct for poor quality baseline data during the linkage. Therefore the success of record linkage is highly dependent on quality data. Pre-processing

involves de-duplicating records, standardizing formats for variables such as dates, and correcting miscoded or missing data. Patient records were consolidated when more than one observation was present per individual. Variables that were not directly used for record linkage were temporarily removed from the datasets.

Data was pre-processed to harmonize variables across datasets. Greater than 4000 distinct, free-text patient residences were recoded to the district level in each ART cohort master file. In the cancer registry, hospital facility was used as a proxy for the district of residence when these were missing, except if the hospital was one of three major referral centers: Kamuzu Central Hospital, Zomba Central Hospital, or Queen Elizabeth Central Hospital. Cancer patients are often diagnosed at the district hospital closest to their residence, prior to receiving referral for oncology services in Lilongwe or Blantyre. After assigning the proxy residence, x% of cancer cases still had missing district of residence.

Regular expressions were used for name parsing and removing prefixes, titles, salutations, special characters and handling initials. In computer science, regular expressions form the basis of search patterns for text strings; these were specifically configured to our study using Espresso 2.1 regular expression editor, then written into the KNIME Analytics data miner[147].

Pre-processing is context and language specific. Beyond issues surrounding name changes due to marriage, maiden names, and nicknames, a particular consideration in the Malawian context is the variation in the spelling of names. Certain letters are used interchangeably such a ‘r’ and ‘l’, ‘y’ and ‘ee’, among many others. A consensus on spelling variations and nicknames for >8000 Malawian common first and last was established by local data staff from Tidziwe Center, Lilongwe (Salima, Chilima, Mukatipa) for use in subsequent steps of review and deterministic linkage. Also, context specific is the use of salutations that are recorded as part of the patient name in Malawi. Salutations need to be removed from name fields during pre-processing. The challenge is balancing the true variation in name spellings that allows algorithms to discriminate between records and excessive data cleaning.

Blocking was used to define potential pairs. Blocking is also referred to as selection criteria ‘pockets’ or stratification of record comparisons according to a given value (see example). Comparisons are restricted to records in the same block, for example records with the same birth year. Blocking is a means to managing the unwieldy number of record comparisons that would otherwise occur during linkage. The

number of record comparisons is the product of the number of records in each file: $N_1 * N_2$. Given that the LT dataset has N_1 >23,000 records and the cancer registry dataset has N_2 > 60,000 records, the product easily exceeds 1,380,000,000 comparisons of potential pairs, which is computationally unfeasible in the absence of a blocking step. The goal therefore is to quickly reduce the number of potential pairs by throwing out pairs with no match in a given block, while allowing for some flexibility of error in the variables. In other words, the goal is to find the potential pairs that are worth looking at. Records pairs with low weights are unlikely to be matches, while those with high weights are more likely to be matches.

Example of Pockets

Pocket 1 Match on year of birth

OR

Pocket 2 Match on sex AND residence

OR

Packet 3 Match on last name AND year of birth

OR

Pocket 4 Match on first name AND year of birth

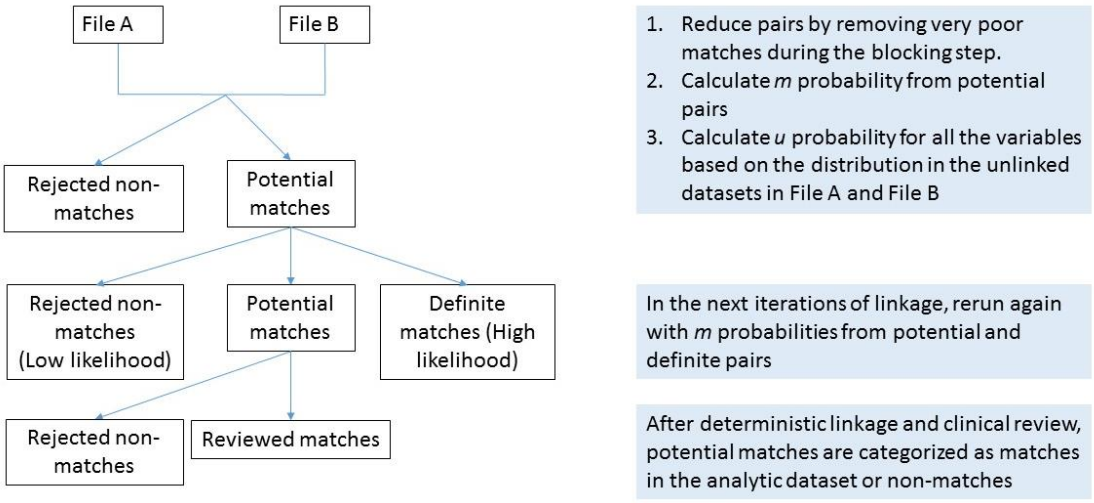
Linkage rules are then defined for each potential pair across each variable, depending on whether the variable is a date, string or numeric: agreement, disagreement, or missing if one or both records are missing information on a given variable. Partial agreements are possible depending on the *a priori* set of matching rules defined by the analytic team. Again, rules for partial agreements are context specific. In Malawi, where dates of birth are often not exact or are missing one or more fields of day, month, and year, we created a rule allowing for a 5- year margin of error in the year of birth, an approach that has been applied in other healthcare linkage studies in Uganda and South Africa. Partial agreements are allowed to account for minor variations within names (see example).

Example of rules for names in Malawi.
* indicates the approach used in the HIV-Cancer Match Study

Rule 1 *Total agreement
Rule 2 Typographical errors: transposition (mary to mayr), character mismatch (mary to mari), extra characters (mary to marry)
Rule 3 *Aliases: nicknames (beth, liz, lizzy for elizabeth), used during the second phase deterministic linkage
Rule 4 *Similarity comparison of n-grams. Other strategies include comparing text strings using similarity filtering with Multi-bit Trees (e.g. used by Google) or Bloom Filter Bit Arrays.

Linkage -*m* and *u*-probabilities are iteratively derived using the study data (Figure 3-2). The *u*-probability is the probability of getting a match on linkage variable given that two records truly do not belong to the same person, or matching by chance. The *u*-probability depends on the distribution of a given linkage variable in the dataset. For example, for the variable sex, one could specify probability *u*=0.5 or adjust it to the observed underlying frequency distribution of sex in the study population; our study did the latter.

Figure 3-2. Iterative linkage process for generating m and u probabilities



After running the first round of linkage, the next step is to refine weights, set the classification thresholds, and categorize the records pairs in one of the three groups: match, non-match, and potential match. In the initial iterations of running the linkage, record pairs are quantitatively reviewed across each variable's weight and total weight using histograms and descriptive statistics. Qualitative assessment

considers overall characteristics of the data such as the frequency of common last names, the proportion of missing data for district and misclassification of year of birth driving the distribution of weights for those respective linkage variables. Linkage rules are then fine-tuned to reflect the importance of each variable in predicting a match within our specific data. The linkage is re-run with the new specifications. Total weight thresholds are defined after an iteration of manual review of a sample of linked record pairs.

Deterministic review of records falling in the category of potential matches was the second stage of record linkage in our study. Often the potential matches had missing data, which drives down the overall total weight for the record pair. Names, year of birth and district were manually reviewed for record pairs that had the lowest weights. The directory of common name variations and nicknames was consulted during this review. Logic rules were constructed to further assign potential matches into one of two categories: those requiring clinical review and non-matches. The following logic rules were applied:

Rule during manual review	Record-pair outcome
Date of death recorded in ART cohort < date of cancer diagnosis	non-match
Date of cancer diagnosis< date of birth	non-match
Not enough information to assign outcome due to missing date of cancer diagnosis	non-match
Not enough information to assign outcome due to missing year of birth and low linkage weight and year of cancer diagnosis occurs >5 years before ART cohort enrollment	non-match
District of residence is missing, but the location of cancer diagnosis is within proximity to the ART clinic, therefore the potential match meets the “physical presence test” of being present in Lilongwe or Blantyre to receive healthcare services	requires further clinical review

Clinical review was conducted by 3 senior Malawian investigators using additional oncology treatment data collected by the cancer registry. Patient identifiers were made available to the reviewers in order to also consult the register of the External Referral Committee, which documents patients sponsored to receive oncology care abroad. Potential matches that linked with a cancer diagnosis occurring more than 5 years prior to ART cohort enrollment (LT n=105; QECH n=163) were reviewed for biologic and clinical plausibility according to the criteria below (see table). Reviewers were provided with 1) information pertaining to the cancer diagnosis: date of diagnosis, age at diagnosis, cancer type, cancer histology, tumor behavior, basis of diagnosis, facility where data was abstracted, treatment, sex; and 2)

information pertaining to the ART cohort: age at ART enrollment, ART facility, cancer diagnosis recorded by ART facility (if any: KS or cervical cancer), time between registry cancer diagnosis and ART enrollment, last date of patient contact, patient outcome as of last date of contact. The final match outcome after clinical review were categorized as match, non-match, and equivocal. Equivocal conclusions were recoded as non-matches in the final analytic dataset.

Criteria during clinical review
Review Criteria #1: Is age at cancer diagnosis compatible with the data presented? (yes, no, equivocal)
Review Criteria #2: Is the treatment information and diagnosis date compatible with survival time between the cancer diagnosis and ART enrollment? (yes, no, equivocal)
Review Criteria #3: Consider if the case was referred out of country by reviewing the register from the External Referral Committee. Is the clinical data compatible with criteria 1 and 2? (yes/no/not applicable)
Final conclusion: enter the final review conclusion. (1=match; 0=clinical data not compatible with a match; equivocal).

During post-processing, clinical information from ART was merged with the final linked analytic file. All personal identifiers were removed to protect patient confidentiality. An encrypted linkage master key containing a matrix of linked IDs and patient names is maintained by the in-country analysis team (Horner, Chasimpha) to allow for record linkage updates in the future.

Statistical analysis specific aim 1.1

Since LT and QECH were independently linked to the cancer registry, the *u*- and *m*-probabilities and subsequent linkage weights are specific to those cohorts. Therefore, all analyses in Specific Aim 1 were conducted separately for each cohort.

Analyses used the final linkage weight cut-point as the basis for classifying potential matches (12 <weight score <23) and definite matches (weight score ≥23). Potential matches were adjudicated in subsequent steps of manual review.

First, the overall proportion of matched and potentially matched records at each stage of the linkage process was graphically presented as a flowchart for each cohort separately. To illustrate potential bias in the classification of match outcomes due to missing data, the proportion of missing values for each linkage variable were presented for definite and potential matches. For each variable, we presented two-

by-two tables showing a) the number and proportion of discrepancies and perfect matches among definite and potential matches, b) the number and proportion of missing data among definite and potential matches. Among 2x2 tables for a), we described the proportion of definite matches that had partial agreement t on first and last name as compared to an exact match on name. Chi-square test were calculated to assess differences in the proportions of a) imperfect matches and b) missingness on high weight and medium weight linkage outcomes.

Lastly, to illustrate how the local distribution of names impacts our linkage study, we compared scatter plots of frequency distributions for first and last names by the linkage weights for these respective variables, stratified by imperfect and perfect matches on name. The goal was to illustrate that in a population with very common last names, even perfect record matching will have only modest linkage weights that require manual review.

E. Specific Aim 2

Specific Aim 2. Characterize cancer incidence rates and clinical timing of cancer diagnosis among ART initiators in Malawi, while accounting for outcome measurement error in a methodologically rigorous way.

Inclusion and exclusion criteria for analytic dataset

Patients newly enrolled on ART in the QECH cohort with a first clinic date occurring between January 1, 2000 and August 31, 2010 were included in the analytic dataset, regardless of prior or current cancer diagnoses. Similarly, patients newly enrolled on ART in the LT cohort with a first clinic date occurring between January 1, 2007 and August 31, 2010 were included for analysis.

Patients enrolled on ART for a single day were excluded as these represent patients who received emergency pharmacy refills. The temporary transfer of these patients into care at LT or QECH was verified against the mastercard for transfer-ins, which is electronically entered into the ART facility database and was made available to the analytic team. Patients enrolled in care at LT or QECH only for receipt of chemotherapy for KS were excluded; these patients receive ART at a separate facility, such as Partners in Hope in Lilongwe.

Outcome definition

Incidence will be restricted to the first cancer primary recorded by the cancer registry. Multiple cancer diagnoses occurring within 90 days of the first diagnosis were considered to be part of the clinical work-up of a single event. Subsequent multiple primaries occurring more than 90 days after the first primary were excluded to avoid misclassification with cancer metastases. Cancer diagnoses that were linked beyond the last date of contact or 180 days past the default date were excluded from primary analyses.

Prevalent cancers are defined as those that were clinically ascertained within the first 90 days of ART enrollment or that occurred prior to enrollment. Early incident cancers are defined as those occurring between 4-24 months days after ART enrollment; late incident occurring greater than 24 months after ART enrollment.

Person-time definition

Person-time on ART was calculated from 90 days after ART enrollment into the cohort until a cancer event, ART cessation, clinic transfer, default or death. Among defaulters, also known as lost to follow-up, the ART outcome date recorded in the electronic medical record (EMR) is the default date. Person-time among patients who defaulted, or lost to follow-up, was defined as the date of the date of the last scheduled appoint or prescription refill that was missed. Calculations of person-time at risk among defaulters included a 180- day window of active tracing past the missed appointment date (Figure 3-3).

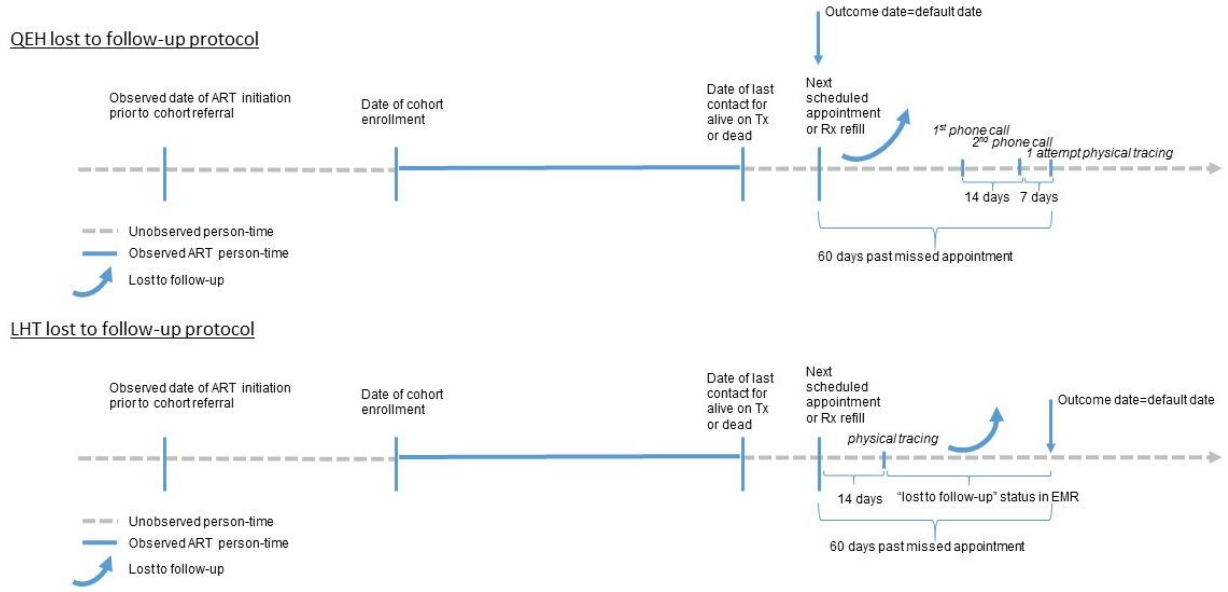
Covariates

Clinical and demographic characteristics were considered as covariates for Poisson regression rates (Table 3-4).

Table 3-4. Covariates

Variable	Description	Values
WHO stage	Categorical	1 or 2, 3 , 4
Reason for ART initiation	Categorical	CD4 level, Pregnant, Breastfeeding, Presumed severe HIV infection in infant, WHO stages 1 or 2, WHO stage 3, WHO stage 4
Age group	Categorical	0-14, 15-29, 30-39, 40-49, 50-59, 60-69, 70+ years
Sex	Binary	Female, Male
Calendar year	Categorical	2000-2003, 2004-2007, 2008-2010

Figure 3-3. Definitions of lost to follow-up dates in ART study cohorts



Descriptive statistics

Descriptive analysis of LT and QECH study populations compared patient characteristics across the periods 2000-2003 (pre-ART), 2004-2007(early ART scale-up), 2008-2010 (late ART scale-up). The distribution of cancer diagnoses according to clinic cohort, 5- year age group, sex, method of cancer diagnosis, site-specific cancers, category of cancer diagnosis (prevalent, early incident, late incident); person-time between ART enrollment and cancer diagnosis, and ART person-years were described.

Statistical analysis specific aim 2

In primary analysis, incident and prevalent cancer diagnoses will be defined relative to the date of ART enrollment.

Cancer incidence rates were calculated per 100,000 person-years and corresponding 95% confidence intervals (95% CI) among persons receiving ART for all sites combined and site-specific cancers. Incidence rates were presented by category of early (4-24 months post enrollment) and late cancer incidence (>24 months post enrollment).

Poisson regression was used for estimating cancer counts and rates by clinical and demographic covariates. People with prevalent cancers were excluded from the population at risk. We applied direct standardization for age (0-15, 16-25, 25-35, 36-45, 46-55, 56+ years) and sex (male, female) using

population weights derived from LT and QECH cohorts combined; otherwise we used sex-specific age-adjustment for male and female populations, respectively.

F. Sensitivity analysis

To quantify the impact of linkage misclassification on cancer incidence rates, we will conduct a sensitivity analysis using probabilistic methods alone compared to the second-pass linkage that uses extensive adjudication of records.

Further sensitivity analyses explored incidence rates among persons with prior ART exposure, who transferred into care at LT or QECH from another clinic. We consider a sensitivity analysis using the interval of unobserved person-time among this sub-cohort of patients with a documented date of prior ART exposure (Figure 3-4).

Figure 3-4. Hypothetical patient timeline for sensitivity analysis of prior ART exposure

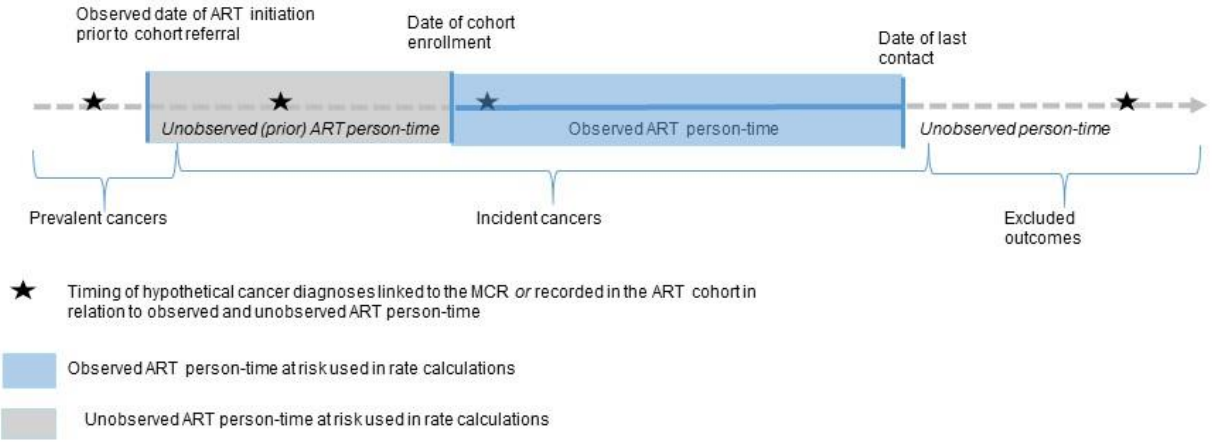


Table 3-5. Definitions of ART person-time, cancer outcomes, and limitations of each analytic approach in Specific Aim 2

Person-time calculation	Outcome classification	Limitations
Primary analysis		
▪Person-time using only observed ART person-time occurring among new ART users ▪Person-time start date=date of ART cohort enrollment ▪Person-time end date=date of last contact if dead or alive, or 180-days past date of last contact if defaulted or lost to follow-up	▪Prevalent cancers=diagnosed prior to cohort enrollment or within first 3 months of enrollment ▪Early incident cancers=diagnosed 4-24 months post-enrollment ▪Late incident cancers=diagnoses >24months post-enrollment	Restriction to subset of cohort deduces sample size
Sensitivity analysis: unobserved prior ART exposure		
▪Person-time using unobserved ART person-years + observed ART person-years among the sub-cohort of *prior* ART users who were referred to LH or QECH ▪Person-time start date=date of ART ▪Person-time end date=last date of contact if dead or alive; 180 days past date of last contact if defaulted or lost to follow-up	▪Prevalent cancers=diagnosed prior to recorded ART *initiation* or within first 3 months of initiation ▪Early incident cancers=diagnosed 4-24 months after ART initiation ▪Late incident cancers=diagnoses >24months after ART initiation	Limitations of unobserved person-time: 1.ART adherence is unobserved. It is unknown if the patient defaulted during the time they initiated ART and subsequently enrolled in the cohort 2.WHO stage at ART initiation is unknown, but may be assumed based on guidelines at the time 3.Date of initial ADC diagnosis may be unknown

CHAPTER 4. OPPORTUNITIES AND BARRIERS TO BIG DATA APPROACHES IN GLOBAL HEALTH: A CASE-STUDY OF HIV-CANCER RECORD LINKAGE IN MALAWI

A. Introduction

Big data approaches provide exciting opportunities to improve population health in low- and middle-income countries.[148] Countries in Africa are no exception and are currently experiencing a data revolution.[149, 150] In this context, novel analytical approaches to integrate data from diverse sources can help fill critical knowledge gaps with respect to clinical care and population health. For example, data linkages of electronic medical records from healthcare systems and traditional sources of disease surveillance data, such as registries, have potential to bridge infectious disease and non-communicable disease research silos in Africa.[151]

Big data broadly refers to four dimensions of data. Volume reflects the size of the data; velocity, the rate at which is it produced; variety, the complexity of sources and formats; and veracity, the certainty of the data.[152] New analytical methods, such as machine learning, increased computing power, and cloud technologies designed to handle the flow of complex data further characterize the big data revolution.[152] Medical record linkages that capitalize on existing data are increasingly deployed in resource-limited settings to strengthen health systems. Data linkages are a versatile and robust tool to efficiently merge information across distinct data sources, providing unique insights into clinical and public health issues,[123-125] but sophisticated technology solutions may not work if they are not tailored to context-specific challenges of the local environment.

Implementing data linkage of health records and registries often faces unique challenges in low-and middle-income countries. Ubiquitous and consistent forms of identification are a challenge for countries in Africa.[153] Furthermore, local technology infrastructure and statistical capacity are usually limited,[154] and systems to collect, disseminate, and use data are typically weak.[155] Patient identification numbers, such as barcoded IDs on the Malawi health passport, a portable abridged medical record belonging to the

patient, may not be consistently noted in local hospital files. Additional statistical challenges include missing data, data accuracy, and absence of gold-standard datasets for validation.

Therefore, new solutions are needed to maximize implementation of big data approaches in resource-limited settings. Case studies with careful assessment of data quality are needed to describe solutions to specific analytic and contextual obstacles encountered during data linkage studies in low-and middle-income countries.[156] In this paper, we present a novel case study from Malawi. In this case study, we specifically address the dimension of data veracity as it applies to public health research. The Malawi HIV-Cancer Match Study leverages data sources across vertical health programs: a population-based cancer registry and clinical systems that support antiretroviral therapy (ART) delivery for HIV. Since HIV status is seldom captured by the cancer registry, and cancer outcomes are not typically collected by HIV clinics in Malawi, we used a data linkage approach to merge these distinct data sources. The study design is innovative in that it employs an approach of probabilistic algorithms for record matching and an iterative process of medical chart review. The design is locally tailored to address real-world challenges of missing data, outcome misclassification, and case validation. We describe our strategy for assessing linkage performance and quality. We share the challenges and lessons learned through this process to further the dialogue of field implementation strategies for data linkage in resource-constrained settings.

B. Methods

Populations

The Malawi HIV-Cancer Match Study is comprised of two observational cohorts of HIV patients receiving ART. In the central region, Lighthouse Trust clinic, located adjacent to at Kamuzu Central Hospital, Lilongwe, is the largest public ART provider. Queen Elizabeth Central Hospital is a tertiary hospital in Blantyre and the largest referral center in the southern Region. Both HIV clinics use electronic medical records to routinely collect demographics, baseline and follow-up clinical data, laboratory tests, and ART information.[126, 127] Active tracing is used for patient follow-up and ascertainment of vital status. Confirmatory HIV diagnosis and WHO clinical staging are provided at time of entry into care.

The Malawi cancer registry uses active case-finding during cross-sectional surveys of secondary (district-level), tertiary (central-level), and major departments of other public and private hospitals. Sources of information include outpatient departments, inpatient wards, clinics, laboratory, pharmacy,

surgery, and mortuaries.[17, 157] Data are coded using the International Classification of Disease for Oncology (ICD-O).[133] Information on diagnosis and treatment are processed and stored in World Health Organization cancer registry software CanReg4.[134]

Record linkage study design

We conducted a data linkage between records from the cancer registry and EMRs from HIV clinics. All patients enrolled at Queen Elizabeth Central Hospital from January 1, 2000 – October, 1 2015 were eligible regardless of prior cancer diagnosis or length of patient follow-up (N=23,743). Similarly, all patients enrolled at Lighthouse Trust from January 1, 2007 – October 1, 2010 were eligible (N=26,977). The entire dataset for the cancer registry from 1985-2010 was used for the record linkage (N=62,944).

Match variables

A combination of non-unique identifiers (names, birth dates) were used to link the HIV cohorts with the cancer registry. In our study, unique identifiers such as a national identification number or cell phone numbers were not available, which is a common challenge in many resource limited contexts. Therefore, we used a set of demographic variables shared across the data sources to link data files: first name, last name, year of birth, sex, and district of residence. Consistent with other African research settings, date of birth was estimated for a substantial proportion of records. Therefore, we incorporated a rule allowing for a 5-year discrepancy in year of birth between pairs of records.[12]

Probabilistic data linkage

We used probabilistic data linkage to match pairs of records across datasets. We considered two main approaches: deterministic and probabilistic methods for linking the datasets together. Deterministic linkage groups records using unique or non-unique identifiers, such as names and birth dates, and relies on the *exact match* of one or more identifiers. Probabilistic linkage instead uses the *probability* of matching across a set of variables for a given record pair.

A match probability and weight is calculated for each potential record pair across each linkage variable. The match weight is the likelihood that two records truly match given agreement on a set of patient identifiers. A high weight indicates a high degree of similarity between a record pair, and a low weight indicates dissimilarity. Record pairs are then classified by their weights into one of 3 categories using a process known as highest-weight classification, resulting in three outcomes: definite match,

possible match requiring review, and non-match. Records with the highest weights above a predetermined weight threshold are categorized as definite matches, while potential matches require review. The weight thresholds used during the classification process are specific both to the data at hand and the purpose of the research. Record linkages are typically run iteratively to refine weight thresholds (Figure 4-1).[136]

Data harmonization and pre-processing

Data pre-processing is the most labor intensive step of a data linkage. Sophisticated software packages and algorithms cannot correct for poor quality baseline data during the linkage. Therefore, the success of record linkage is highly dependent on data quality. Pre-processing involves de-duplicating records, standardizing formats for variables such as dates, and correcting miscoded or missing data. Patient records were consolidated when more than one observation was present per individual.

Data was pre-processed to harmonize variables across datasets. Greater than 4000 distinct, free-text patient residences were recoded to the district level in each HIV cohort master file. In the cancer registry, hospital facility was used as a proxy for the district of residence when these were missing, except if the hospital was one of three major referral centers: Kamuzu Central Hospital, Zomba Central Hospital, or Queen Elizabeth Central Hospital. Cancer patients are often diagnosed at the district hospital closest to their residence, prior to receiving referral for oncology services in Lilongwe or Blantyre.

Regular expressions were used for name parsing and removing prefixes, titles, salutations, special characters and handling initials. Pre-processing is context and language specific. Beyond issues surrounding name changes due to marriage, maiden names, and nicknames, a particular consideration in the Malawian context is the variation in the spelling of names. Certain letters are used interchangeably such a ‘r’ and ‘l’, ‘y’ and ‘ee’. For example, ‘Graham’ is interchanged with ‘Glaham’, ‘Eviness’ with ‘Eveness’, ‘Bitya’ with ‘Bitia’, and ‘Jacqueline’ with ‘Jacqureen’. A consensus on spelling variations and nicknames for >8000 Malawian common first and last was established by local data staff for use in subsequent steps of review. Also, context specific is the use of salutations that are recorded as part of the patient name in Malawi. Preprocessing was conducted in KNIME Analytics data miner.[147]

Highest weight classification of records

Several iterations of the linkage were run to refine the matching algorithms and weight thresholds (Figure 4-2). During the initial iterations, the study team refined match criteria based on a preliminary review of randomly sampled record pairs from a range of high to low weights. The sampling was done as a qualitative appraisal of the quality of record matches. Descriptive statistics and boxplots were generated to visualize the range of weights for each match variable and the overall total weight score for a record pair. Linkage rules were then fine-tuned by the study team (MJH, SC, AS) to reflect the importance of each variable in predicting a match within our specific database. Linkage algorithms were re-run with the new specifications. For example, many people have an estimated birthdate, we reset the criteria for the birth year variable to allow a 5-year window for matching and to contribute a lower weight to the total weight score.

The total weight score for each record pair represents the log odds of obtaining a true record pair to a true non-record pair given the agreement patterns across the set of 5 match variables. A high weight indicates a high degree of similarity between a record pair, and a low weight indicates dissimilarity. Negative values represent non-matches. In our study, the total weight scores range from -28 to 39. The study team set the weight thresholds after qualitatively reviewing a random sample of matched record pairs for accuracy. Since *a priori* information of the expected number of cancer matches was unknown, we set the weight threshold to maximize sensitivity by including the maximum number of potential matches. As in other linkage studies, highest-weight classification (Figure 4-3) was used to classify record pairs as definite matches ($23 \leq weight \leq 39$), potential matches ($12 \leq weight < 23$) or non-matches ($-28 \leq weight < 12$), followed by manual adjudication of potential matches. A weight less than 23 in our study is the point at which data quality issues or missing data introduce uncertainty, and extra clinical information from the HIV medical record or cancer registry is required to make a final determination. In primary analysis, definite matches ($23 \leq weights \leq 39$) and the clinically adjudicated matches ($12 \leq weights < 23$) were used for cancer incidence rate calculations. In subsequent sensitivity analyses, we calculated rates only using definite matches ($23 \leq weights \leq 39$) to exclude the possibility of false matches; this approach is common to studies of similar design.[158]

Outcome validation

Since a gold standard dataset was not available for external validation of matches, we conducted a two-pass process of clerical and clinical review. Logic rules were constructed to validate highest weight definite matches and to further parse potential matches into one of two categories: those requiring further clinical review and non-matches. Potential matches were manually inspected to assess whether records belonged to the same person. Reviewers (MJH,SC) were blinded to the cancer type during clerical review. The study team used *a priori* criteria consisting of date of birth, when available, date of death, date of cancer diagnosis, and hospital location of the cancer diagnosis when residence information was missing (Table 4-1). For example, 29% of records in the Lighthouse Trust cohort has missing district of residence. Since 15% of people attend clinic solely for an emergency prescription refill, once could not assume that those people permanently reside in the metropolitan area of Lilongwe. Therefore, if the hospital location of the cancer diagnosis was within proximity of the HIV clinic, we took that as meeting the physical presence test of residing in the district. Name variations were reviewed against a directory of regionally specific common names and nicknames complied specifically for this purpose by the study team.

Clinical review of potential matches was conducted by three senior Malawian investigators using additional oncology treatment data collected by the cancer registry and archival patient records from National Oncology Review Board's External Referral Committee, which documents patients sponsored to receive oncology care abroad. Potential matches were reviewed for biological and clinical plausibility according to *a priori* criteria. Reviewers were provided with 1) information pertaining to the cancer diagnosis: date of diagnosis, age at diagnosis, cancer type, cancer histology, tumor behavior, basis of diagnosis, facility where data was abstracted, treatment, sex; and 2) information pertaining to the ART cohort: age at ART enrollment, ART facility, AIDS-defining cancer diagnosis recorded by HIV clinic, if any, time between registry cancer diagnosis and start of ART, last date of patient contact, patient outcome as of last date of contact. The final match outcome after clinical review were categorized as match, non-match, and equivocal. Equivocal conclusions were recoded as non-matches in the final analytical datasets.

Post-processing

During the final step of linkage post-processing, clinical information from ART clinics was merged in the analytical file. All personal identifiers were removed to protect patient confidentiality. An encrypted linkage master key containing a matrix of linked IDs and patient names is maintained by the in-country analysis team to allow for record linkage updates in the future.

Statistical analysis

First, we assessed match performance in each cohort by calculating the match rate, defined as the total number of matches divided by the total number of people in each cohort. The quality of the probabilistic matching is defined as the proportion of cases that are categorized as high weight "definite matches" (weight ≥23) versus low weight "potential matches" (12≤weight <23) requiring further review, and the proportion of cases that are discarded due to insufficient information. The study team (MJH,SC) used a random sample review of record pairs to set the weight threshold used in highest weight classification of "definite" and "potential" matches.

Second, we examined the impact of missing data on outcome classification during the first step of probabilistic matching. To assess the impact of missing data on the highest weight classification of records into high weight (weight ≥23) and low weight matches (12≤weight<23), we used Pearson's chi-square test of proportions for year of birth (missing versus not missing) and district (missing versus not missing).

Third, we examined the impact of partial agreement of identifiers on outcome classification during the first step of probabilistic matching. To evaluate the impact of data quality of the classification of records into high weight and low weight matches from the first step of probabilistic matching, we used we used Pearson's chi-square for exact and partial agreement on name and year of birth.

Fourth, we examined whether individual-level characteristics introduced a systematic bias in the linkage of each cohort. We hypothesized that people who were lost to follow-up, transferred out of care, or ceased therapy at HIV clinic would be less likely to have a match with the cancer registry.

Fifth, we qualitatively assessed the specific contextual challenge of how common first names and surnames in Malawi reduce the discriminatory power of probabilistic match algorithms. We present the

classification of records into definite and potential matches across frequencies of surname in our study populations.

Analyses were conducted in SAS 9.4 (Cary, North Carolina). All significance tests were conducted at the α=0.05 level.

C. Results

Linkage performance

Probabilistic matching identified a total of 1,269 definite and potential matches across 28,576 new ART users across both cohorts (Figure 4-2). Probabilistic linkage performance varied across cohorts, with higher performance yield at QECH compared to LH (QECH: n=731 matches; match rate=6%; LT: n=538 matches, match rate=3%). By including the additional steps of clerical and clinical review, 36% to 75% of potential matches at QECH and LT were discarded due to missing identifiers, inconsistency between date of diagnosis and vital status, or insufficient information from the clinical record to make a final determination. The post-review match rate remained higher at QECH compared to LT (4% versus 1%).

Data completeness and quality

The level of completeness of identifiers used in the matching algorithms was variable across data sources (Table 4-1). First and surname had negligible missing data in both HIV cohorts and in the cancer registry. Consistent with other clinical settings in Africa, exact birth date was not routinely recorded in health clinics. The majority of records had unknown day and month of birth. Exact month of birth was unknown for 27% of records in LT, 40% in QECH, and 85% in the cancer registry. Exact year of birth was estimated for 52% of records in LT, and missing for 36% in QECH and 8% in the cancer registry. Similarly, district of residence was missing for 9% to 88% of records across data sources.

Missing data was an important determinant of the large number of potential matches requiring labor intensive manual review and adjudication. In Tables 4-2 and 4-3, medium weight matches were associated with a greater proportion of missing birth year compared to high weight matches at LT (28% versus 2%, $p<0.001$) and QECH (19% versus 3%, $p<0.001$). To a lesser extent, medium weight matches were associated with a greater proportion of missing district compared to high weight matches at LT (32% versus 22%, $p=0.05$), but not at QECH (56% versus 64%, $p=0.06$).

Imperfect matches on name and birth year also resulted in time and labor intensive manual review efforts of medium weight matches. In Tables 4-4 and 4-5, imperfect agreement across all link fields was more common for medium weight matches compared to high weight matches for LT (98% versus 61%; $p<0.001$) and QECH (99% versus 89%; $p<0.001$). Partial agreement across each link field separately, except for sex, was consistently more common among low weight matches compared to high weight matches.

Bias analysis

Vital status and retention in care in clinic were associated with a cancer match outcome at LT Table 4-6). Among deceased persons the odds of a matched outcome were nearly twice that of people who were alive and in care (OR=1.9; 95%CI 1.3, 2.8), after adjusting for age and sex. A greater odds of a match outcome relative to those who remained in care at LT was observed for people who transferred to another clinic (OR= 1.5, 95%CI 0.9, 2.2), were lost to follow-up (OR=2.0, 95%CI 1.4, 3.0), or ceased antiretroviral therapy (OR=2.6, 95%CI 1.5, 4.3), after adjusting for age and sex.

People <30 years at the time of enrollment has a reduced odds of matching to the cancer registry (OR=0.6, 95%CI 0.4, 0.9) compared to people ages 30-44 years, after adjusting for vital status and sex. There was no association between ages older than 45 years (OR=1.0, 95%CI 0.7, 1.5) and female sex (OR=1.0, 95%CI 0.8, 1.4) and matching in the cancer registry.

Onomastics

Distribution of surnames varied by geographical location of the HIV cohorts in our study. Two surnames comprised 7.6% of 9,359 unique names in Lilongwe and 4.1% of 8,517 unique names in Blantyre. Because probabilistic weights are a function of 1) the degree of agreement between identifiers and 2) the frequency at which identifiers occur in the data, records with common last names tend to cluster in the range of medium weight scores (Figure 4-4). At Lighthouse, the low overall match rate was in part due to the high proportion of discarded matches (40%) with common surnames occurring at frequencies >0.5% (Figure 4-3).

D. Discussion

Our case study describes common analytic barriers that may be encountered during complex healthcare data linkage studies in LMIC. We evaluated analytic challenges encountered during probabilistic linkage of healthcare records from two observational HIV cohorts to a population-based cancer registry in Malawi. We used a hybrid approach for the record linkage: probabilistic matching on demographic identifiers, followed by labor- and time-intensive manual adjudication of records and clinical review. Our hybrid approach was designed to address three main analytic challenges encountered in our setting: missing data, reduced resolution of link fields in Malawi, and absence of a gold standard dataset for validation of outcomes.

Data quality and completeness are real-world, practical concerns for the design of linkage studies, particularly in resource-limited environments. Missing data and misreporting of patient identifiers significantly impacted the overall performance of probabilistic matching in our study. Country-specific circumstances such as lack of birth certificates cause substantial missing data on date of birth in clinical records in Malawi. In probabilistic matching, the degree of similarity between values of a given variable and the weighted contribution of that variable together drive the total weight score for a given record pair.[159] When either of two values is missing, it is not possible to calculate the similarity distance, and that variable contributes nothing to the total weight score. Detailed physical location of patient residence is recorded by HIV cohorts for the purpose of active tracing, and is therefore a quasi-unique identifier. However, detailed residence information is not available to the cancer registry. The linkage design therefore relies on the most common denominator information for patient residence, which is a loss of resolution for this highly informative patient identifier. Lastly, time-varying patient demographics such as maiden name and residence are generally complex to handle with algorithms alone; manual adjudication with extra longitudinal data is usually required. Striving towards consistent and high resolution patient identifiers would improve linkage outcomes in the long-term.

Data errors in match variables may be language and context specific. For example, in Malawi, certain letters like ‘r’ and ‘l’ are commonly interchanged in Chichewa, the local Bantu-based language. Additional discrepancies may arise from typographical errors occurring during data entry, variations in patient demographics, and the use of nicknames. Probabilistic methods are well-suited to handle minor

discrepancies such as these[136, 143]. However, the algorithms also weight the frequency at which identifiers occur in the study sample. Common surnames and first names present a major challenge in Malawi because they reduce the discriminatory power of linkage on patient name. For example, the two surnames “Banda” and “Phiri” account for 7% of names in our study population. Even exact agreement on a common name will contribute little to the total summary weight for that record pair. In contrast, rare names contribute a high weight towards the overall summary weight of the record pair. Anecdotally, clinic staff also reported the use of nicknames and aliases among people enrolling in care due to social stigma associated with HIV, creating cultural and disease-specific challenge for our study. This issue may also apply to similar efforts focused on stigmatized disease states in low- and middle-income countries.

A critical step in the linkage process is to assess the performance and quality of linkage, especially when the number of expected cases is unknown.[160] We evaluated the match rate at each step of the linkage process as an indicator of performance. Variations in data quality and completeness of linkage fields across study sites likely affected the linkage performance. The greater match rate in Queen Elizabeth Central Hospital is likely due to higher completeness of cancer registration activities in that region of the country, and richer sources of secondary clinical information with which to triangulate matched cases. Insufficient data to make a final decision resulted in a large proportion of cases being discarded. The reduced match rate, and large number of “missed links” likely reduced the sensitivity of the linkage at Lighthouse, but it was not feasible to directly measure this.

Taken together, missing data, data errors, legitimate variations in demographic identifiers, and country-specific name frequencies jointly lower linkage weights, creating a scenario of uncertainty in the ‘true’ match status of records. The use of probabilistic algorithms alone in low-and middle-income countries may be overly conservative in assigning HIV-cancer matches due to the aforementioned underlying analytical challenges in the local healthcare system. A validation step is important to gauge the extent of possible measurement error in linkage outcomes, and calibrate analyses accordingly. Since a gold standard dataset was not available, our approach therefore used a time and labor intensive review process to adjudicate potential matches using additional information abstracted from the medical record. In low-and middle-income countries, high quality review may be difficult when secondary information on a case is simply not available. Lack of triangulating data to verify links may be a major setback[156], and the

availability of additional sources of data with which to validate cases should be considered carefully during the study planning phase of a healthcare linkage.

Adding a supplemental review step with external clinical data is a field-tested approach to address real-world statistical limitations in a resource-limited setting. When available, biometric-based identifiers may also improve the quality of data linkage products. In South Africa, fingerprint scans used as IDs in large demographic surveys [161] and in select healthcare centers were used to validate the identify of cases in a population-based data linkage.[162] Rural Ghana used fingerprint scans as a primary linkage identifier between a demographic survey and health centers.[163] As fingerprint scans are launched as a primary means of national identification in select African countries[153], with biometric-based forms of identification leapfrogging paper-based methods, it remains to be seen how to best incorporate these new technologies into big data approaches in public health.

E. Conclusions

We demonstrated that a big data approach to public health research is feasible in a low-income country. Record linkage is a powerful tool, but presents limitations [160] that warrant transparent discussion within the research community. Reporting linkage results and limitations in a transparent manner[164] may assist with field implementation of similar studies in other countries in Africa. Despite statistical and other challenges present in resource-constrained environments, record linkages are an opportunity to invest in strengthening health systems. Continued and sustained investments in data quality and completeness should be long-term priorities for public health research in the era of big data medical informatics.

Panel. Criteria used during clerical review of definite and potential matches.

Applies to	Logic rules	Outcome
All matches (12<weight<40)	Date of death recorded in ART cohort prior to date of cancer diagnosis	non-match
All matches (12<weight<40)	Date of cancer diagnosis prior to date of birth	non-match
All matches (12<weight<40)	Not enough information to assign outcome due to missing date of cancer diagnosis	non-match
Potential matches (12< weight< 23)	Not enough information to assign outcome due to missing year of birth	non-match
Potential matches (12< weight< 23)	District of residence is missing, but the location of cancer diagnosis is within proximity to the ART clinic, therefore the potential match meets the “physical presence test” of being present in Lilongwe or Blantyre to receive healthcare services	requires further clinical review

Lilongwe hospitals: ABC Clinic, Kamuzu Central Hospital, Mtengowanthenga, Likuni, Kasungu, City Centre Clinic

Blantyre hospitals: Mwaiwathu Private Hospital, Queen Elizabeth Central Hospital, Mlambe, Chitawira Private, Mtengoumodzi, Malamulo (Makwasa And Amina), St.Joseph's (Nguludi), Blantyre Adventist Hospital

Table 4-1. Completeness of identifiers used in probabilistic matching

	Lighthouse Trust (2007-2010)		Queen Elizabeth Central Hospital (2000-2010)		Malawi Cancer Registry (1985-2010)	
	N	%	N	%	N	%
Total records	26,961		23,510		62,944	
Complete surname	26,958	100.0%	23,509	100.0%	62,943	100.0%
Surname initial	2	0.0%	0	0.0%	0	0.0%
Missing	1	0.0%	1	0.0%	1	0.0%
Complete First name	26,933	99.9%	23,505	100.0%	62,912	99.9%
First name initial	27	0.1%	4	0.0%	32	0.1%
Missing	1	0.0%	1	0.0%	220	0.3%
Estimated birthdate *	14,106	52.3%	n/a		n/a	
Birth year, recorded	26,961	100.0%	15,149	64.4%	57,850	91.9%
Birth Month, estimated July	7,519	27.9%	9,501	40.4%	53,391	84.8%
Birth Day, estimated 1st	6,957	25.8%	18,037	76.7%	56,749	90.2%
Birth Day, estimated 15th	1,364	5.1%	n/a			n/a
District of residence						
Blantyre	11	0.0%	749	3.2%	18,478	29.4%
Lilongwe	18,517	68.7%	38	0.2%	8,137	12.9%
Other districts	770	2.9%	2,108	9.0%	30,758	48.9%
Missing	7,663	28.4%	20,615	87.7%	5,571	8.9%
Sex						
male	6,626	24.6%	10,278	43.7%	29,630	47.1%
female	9,819	36.4%	13,232	56.3%	33,312	52.9%
missing	10,516	39.0%	n/a		2	0.0%

*one or more elements of date of birth are estimated

Table 4-2. Missing characteristics across link fields during each step of the record linkage process, Lighthouse Trust

	Probabilistic linkage only							Probabilistic linkage and review						
	Total matches		High weight ≥23		Medium weight >12 and <23			Total		High weight ≥23		Medium weight >12 and <23		
	N	%	N	%	N	%	*p*	N		N	%	N	%	*p*
Year of Birth														
Partial or exact agreement	414	77%	98	98%	316	72%		200	99%	93	98%	107	100%	
Missing	124	23%	2	2%	122	28%	<0.001	2	1%	2	2%	0	0%	0.13
District														
Partial or exact agreement	375	70%	78	78%	297	68%		170	84%	75	79%	95	89%	
Missing	163	30%	22	22%	141	32%	0.05	32	16%	20	21%	12	11%	0.85

Table 4-3. Missing characteristics across link fields during each step of the record linkage process, Queen Elizabeth Central Hospital

	Probabilistic linkage only							Probabilistic linkage and review						
	Total matches		High weight ≥23		Medium weight >12 and <23			Total		High weight ≥23		Medium weight >12 and <23		
	N	%	N	%	N	%	*p*	N	%	N	%	N	%	*p*
Year of Birth														
Partial or exact agreement	626	86%	206	97%	420	81%		449	97%	205	97%	244	96%	
Missing	105	14%	6	3%	99	19%	<0.001	15	3%	6	3%	9	4%	0.67
District														
Partial or exact agreement	305	42%	77	36%	228	44%		213	46%	77	36%	136	54%	
Missing	426	58%	135	64%	291	56%	0.06	251	54%	134	64%	117	46%	<0.001

Table 4-4. Degree of similarity across link fields during each step of the record linkage process, Lighthouse Trust

	Probabilistic linkage only							Probabilistic linkage and review						
	Total matches		High weight ≥23		Medium weight >12 and <23			Total matches		High weight ≥23		Medium weight >12 and <23		
	N	%	N	%	N	%	*p*	N	%	N	%	N	%	*p*
All identifiers (name, year of birth, district, sex)														
Full agreement	46	9%	39	39%	7	2%		44	22%	39	41%	5	5%	
Partial agreement	492	91%	61	61%	431	98%	<0.001	158	78%	56	59%	102	95%	<0.001
Full name and year of birth														
Full agreement	83	20%	55	56%	28	9%		58	29%	51	55%	7	7%	
Partial agreement	331	80%	43	44%	288	91%	<0.001	142	71%	42	45%	100	93%	<0.001
Year of Birth														
Full agreement	198	48%	71	72%	127	40%		102	51%	69	74%	33	31%	
Partial agreement	216	52%	27	28%	189	60%	<0.001	98	49%	24	26%	74	69%	<0.001
Full name														
Full agreement	313	58%	82	82%	231	53%		139	69%	80	84%	59	55%	
Partial agreement	225	42%	18	18%	207	47%	<0.001	63	31%	15	16%	48	45%	<0.001
District														
Full agreement	368	98%	71	91%	297	100%		160	94%	67	89%	93	98%	
Partial agreement	7	2%	7	9%	0	0%	<0.001	10	6%	8	11%	2	2%	0.02
Sex														
Full agreement	536	100%	100	100%	436	81%		200	99%	95	100%	105	98%	
Disagreement	2	0%	0	0%	2	0%	0.49	2	1%	0	0%	2	2%	0.18

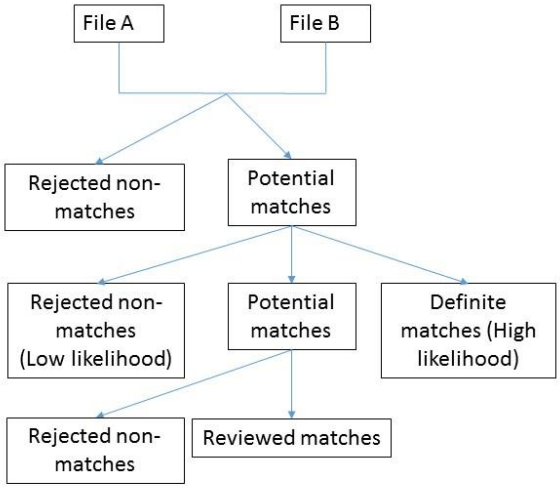
Table 4-5. Degree of similarity across link fields during each step of the record linkage process, Queen Elizabeth Central Hospital

	Probabilistic linkage only							Probabilistic linkage and review						
	Total matches		High weight ≥23		Medium weight >12 and <23			Total matches		High weight ≥23		Medium weight >12 and <23		
	N	%	N	%	N	%	*p*	N	%	N	%	N	%	*p*
All identifiers (name, year of birth, district, sex)														
Full agreement	31	4%	24	11%	7	1%	<0.001	31	7%	24	11%	7	3%	
Partial agreement	700	96%	188	89%	512	99%		433	93%	187	89%	246	97%	<0.001
Full name and year of birth														
Full agreement	124	20%	99	48%	25	6%	<0.001	116	26%	98	48%	18	7%	
Partial agreement	502	80%	107	52%	395	94%		333	74%	107	52%	226	93%	<0.001
Year of Birth														
Full agreement	184	29%	114	55%	70	17%		162	36%	113	55%	49	20%	
Partial agreement	442	71%	92	45%	350	83%	<0.001	287	64%	92	45%	195	80%	<0.001
Full name														
Full agreement	548	75%	190	90%	358	69%		359	77%	189	90%	170	67%	
Partial agreement	183	25%	22	10%	161	31%	<0.001	105	23%	22	10%	83	33%	<0.001
District														
Full agreement	305	100%	77	100%	228	100%		213	100%	77	100%	136	100%	
Partial agreement	0	0%	0	0%	0	0%		0	0%	0	0%	0	0%	
Sex														
Full agreement	726	99%	212	100%	514	99%		462	100%	211	100%	251	99%	
Disagreement	5	1%	0	0%	5	1%	0.15	2	0%	0	0%	2	1%	0.20

Table 4-6. Association of individual-level characteristics among matched and non-matched records, Lighthouse Trust

	Match	Non-match	
	N	N	Odds Ratio (95%CI)
Patient-level follow-up status			
Alive, in care	74	8066	1.
Lost to follow-up	40	2299	2.0 (1.4, 3.0)
Stopped therapy	18	783	2.6 (1.5, 4.3)
Transfer to another clinic	31	2342	1.5 (0.9, 2.2)
Deceased	39	2224	1.9 (1.3, 2.8)
missing	0	3	
Age group (years)			
<30	53	5522	0.6 (0.4, 0.9)
30-44	117	8009	1.
45+	32	2187	1.0 (0.7, 1.5)
Sex			
Female	91	6622	1. 03 (0.8, 1.4)
Male	111	9096	1.

Figure 4-1. Record linkage workflow



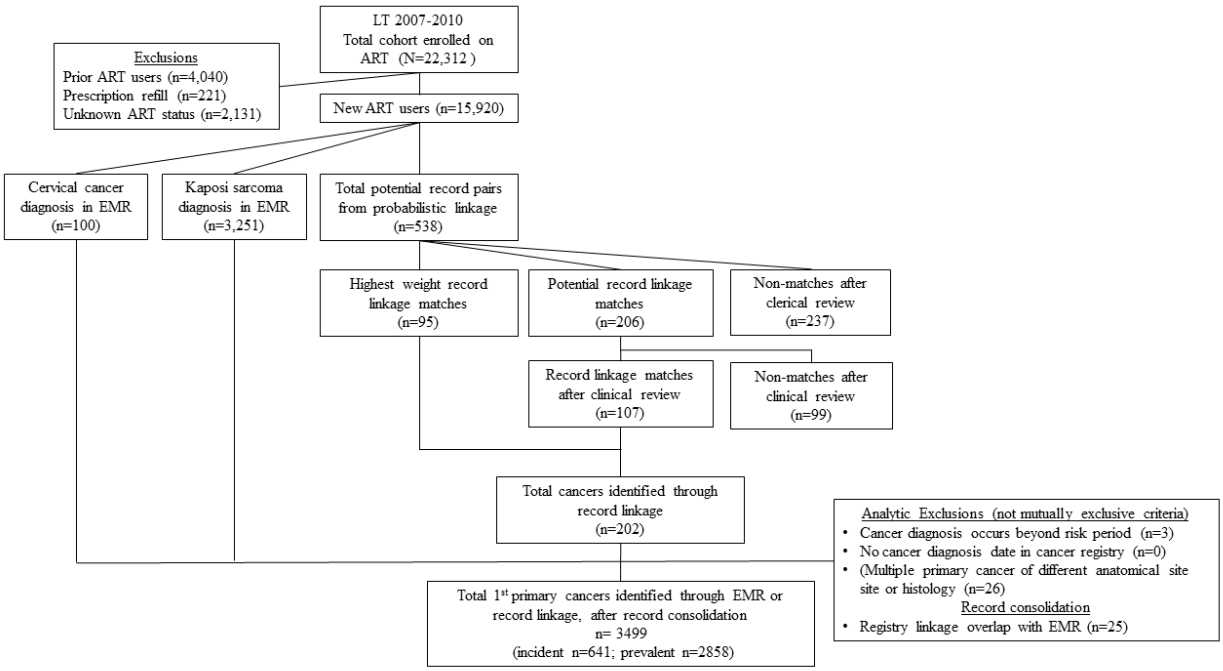
1. Reduce pairs by removing very poor matches during the blocking step.
2. Calculate *m* probability from potential pairs
3. Calculate *u* probability for all the variables based on the distribution in the unlinked datasets in File A and File B

In the next iterations of linkage, rerun again with *m* probabilities from potential and definite pairs

After deterministic linkage and clinical review, potential matches are categorized as matches in the analytic dataset or non-matches

Figure 4-2. Flowchart of cancer registry linkage to HIV patient records (A: Lighthouse Trust; B: Queen Elizabeth Central Hospital)

A.



B.

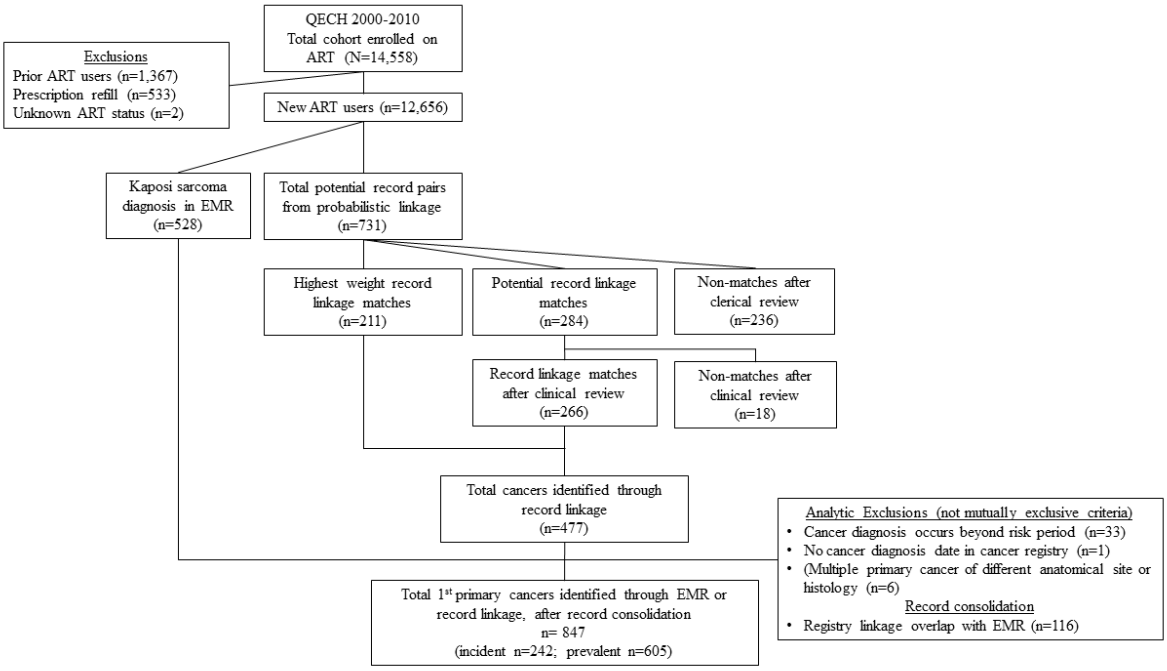


Figure 4-3. Highest weight classification of linkage outcomes as a function of perfect and partial agreement (A: Lighthouse Trust; B: Queen Elizabeth Central Hospital)

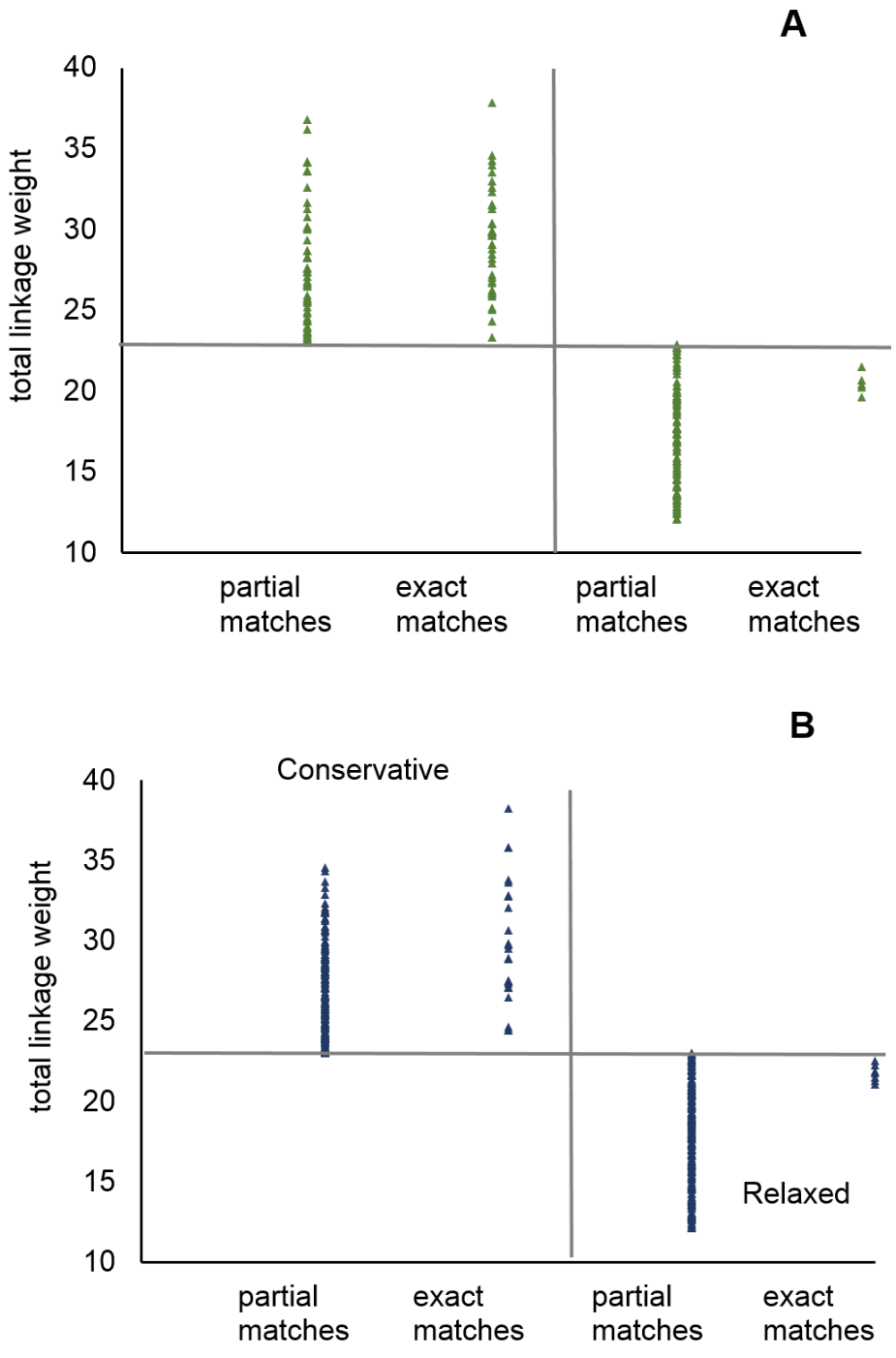
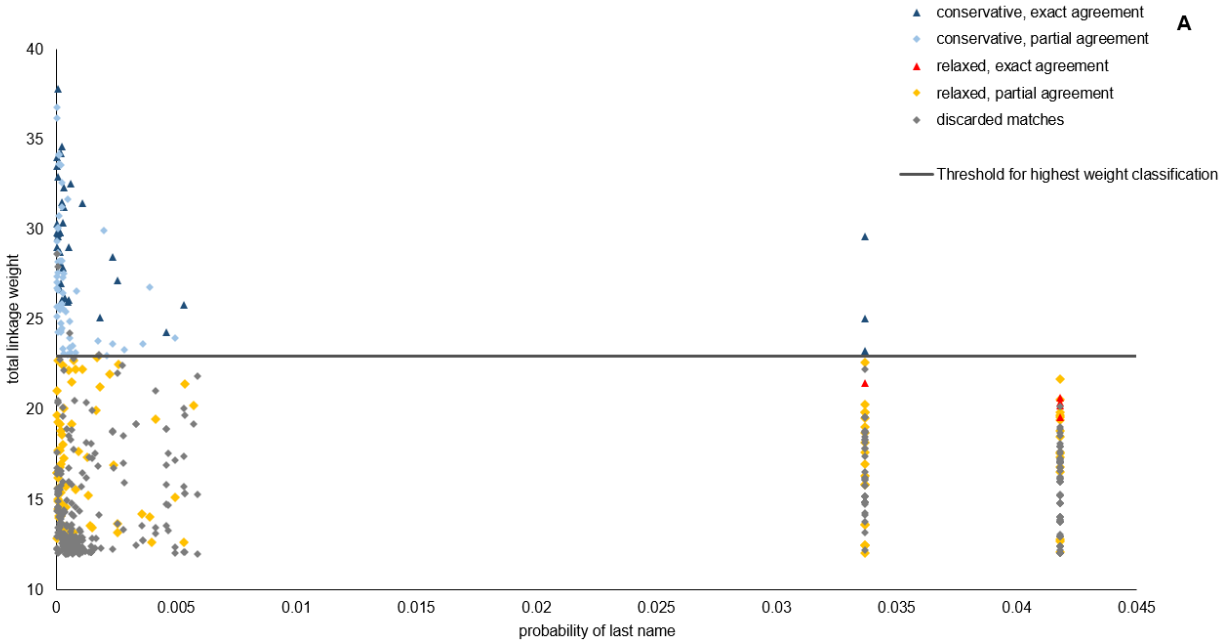
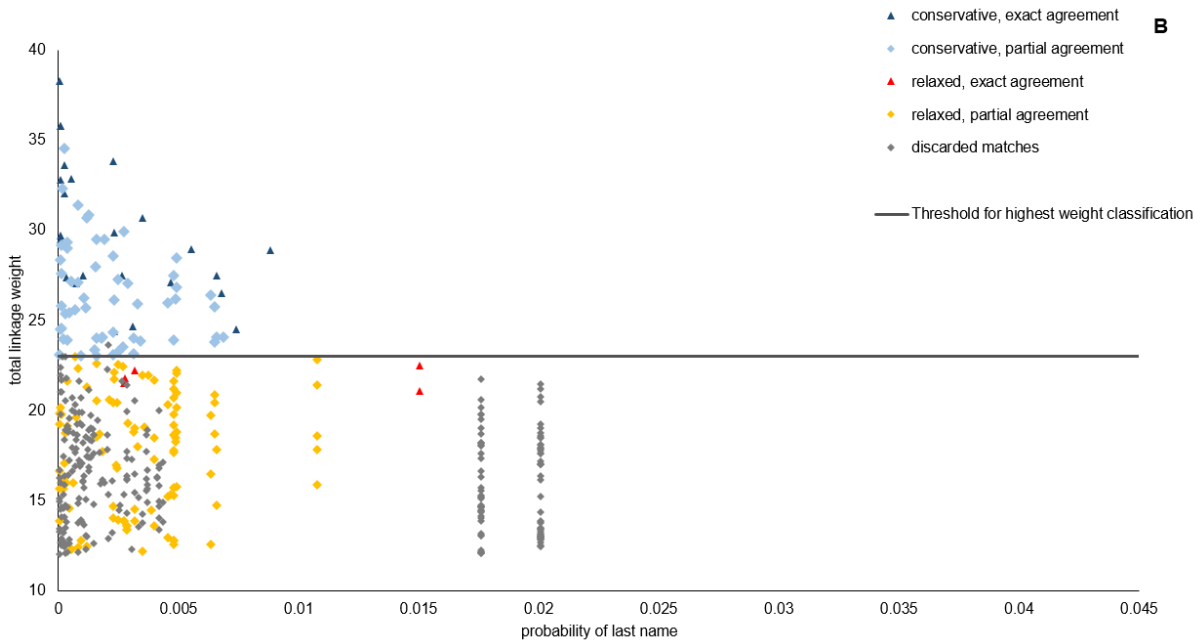


Figure 4-4. Highest weight classification of linkage outcomes as a function of perfect and partial agreement and distribution of surnames in HIV cohorts (A: Lighthouse Trust; B: Queen Elizabeth Central Hospital)





CHAPTER 5. HIGH CANCER BURDEN AMONG ANTIRETROVIRAL THERAPY USERS IN MALAWI: A RECORD LINKAGE STUDY USING OBSERVATIONAL HIV COHORTS AND CANCER REGISTRY DATA

A. Introduction

Three AIDS-defining malignancies, Kaposi sarcoma (KS), cervical cancer, and non-Hodgkin lymphoma (NHL) are among the top 10 cancers in sub-Saharan Africa (SSA), where 70% of global HIV burden is concentrated.[89,20] Rapid scale-up of ART availability over the past decade[89] is likely to affect regional cancer burden. In high-income countries, cancer risk among HIV populations is evolving, with notable declines in risk of KS, NHL, and some non-AIDS defining cancers (NADC) over the course of ART expansion since 1996.[165] The burden of certain NADC is now projected to increase and surpass that of AIDS-defining cancers (ADC) due to growth and aging of the population living with HIV.[166] But extrapolations from high-income countries may not apply to SSA, where delays in accessing HIV care are substantial, advanced immunosuppression at ART initiation is common[167], and prevalence of oncogenic viral infections is high.[168] Epidemiological evidence is therefore needed to understand current cancer trends specifically in SSA.

In the Malawi HIV-Cancer Match Study, we aim to characterize cancer incidence among ART initiators. In Malawi, HIV prevalence is 9% and ART coverage has reached 67%, using a threshold for ART eligibility of 500 CD4 cells/µL or WHO clinical stages 3 and 4.[14] Our work differs from previous studies in the region [12, 116, 117], in that we used two of the largest, actively traced cohorts of ART users in the country. We also conducted cross-sectional linkage of cancer outcomes using the population-based national cancer registry. Finally, we examined a study period spanning the initiation of national ART scale-up from 2000 through 2010.

B. Methods

Populations

Study subjects were HIV-infected people receiving ART at Lighthouse Trust (LT) and the Queen Elizabeth Central Hospital (QECH) HIV clinics. In the central region, LT in the capital, Lilongwe, is Malawi's largest public ART provider. In the south, the QECH HIV clinic is situated in Blantyre, Malawi's second largest city. LT and QECH use electronic monitoring systems to routinely collect demographic information, WHO stage at clinic enrollment, drug regimens, and patient outcomes.[126, 127] Active tracing is used for patient follow-up and ascertainment of vital status. In Malawi, CD4 count measurement was historically restricted to stage 1 and 2 patients who were not clinically eligible for ART (e.g. stages 3 and 4). For QECH prior to 2011, CD4 counts were not captured in the electronic monitoring system and therefore were not available for analysis. Routine HIV RNA monitoring in Malawi did not begin until 2011. Therefore, limited CD4 count data and no HIV RNA data were available during the time period of our study reflecting practice within the Malawi national HIV program.

The population-based Malawi Cancer Registry (henceforth, the registry) is a founding member of the African Cancer Registry Network and one of only five cancer registries from SSA included in the WHO *Cancer Incidence on Five Continents* monograph.[20] Active case-finding is conducted through cross-sectional surveys of secondary (district-level) and tertiary (central-level) hospitals, and major departments of other public and private hospitals. Population-based catchment areas and data collection procedures have been described previously.[17] During the study period, QECH in Blantyre housed the sole pathology laboratory for the entire country; a second laboratory open at Kamuzu Central Hospital, Lilongwe in 2011.[169] Thus, in the 2010 national cancer survey, only 18% of cases were pathologically confirmed, with most cancer diagnoses supported by clinical, radiological, and/or laboratory data.[17] The registry pathology confirmation rate is comparable to that of other population-based cancer registries from SSA.[170, 171]

Electronic medical record linkage

In the absence of unique personal identifiers, we used probabilistic algorithms to link electronic medical records from ART cohorts with cancer records over periods of geographical overlap between HIV clinics and cancer registration. All HIV-infected people initiating ART at QECH from 2000 to 2010 or LT

from 2007 to 2010 were eligible based on years of registry coverage in Blantyre and Lilongwe, respectively. Electronic medical records (EMR) were matched on first and last name, year of birth, sex, and district of residence as these identifiers were shared across datasets. Consistent with other SSA settings, date of birth was estimated for a substantial proportion of records. Therefore, only year of birth was used as a linkage variable, allowing for 5-year discrepancies between record pairs during subsequent iterations of the linkage.[12] As in other linkage studies, highest-weight classification was used to classify record pairs as matches, potential matches or non-matches, followed by manual adjudication of potential matches.[12] Potential matches were manually reviewed and validated according to *a priori* criteria. Potential matches were further validated through clinical review by three senior Malawian investigators using additional oncology treatment data collected by the registry and National Oncology Review Board, when available. Data preprocessing and probabilistic record linkage were performed in KNIME Analytics Platform Version 2.12.1 (Konstanz, Germany)[147], an open-platform data miner, utilizing the K-Link probabilistic linkage plug-in. Data pre- and post-processing were conducted in Stata 14 (Stata Corporation, College Station, Texas). Analyses were conducted in SAS software 9.4 for Windows (Cary, NC, USA). The study was approved by the University of North Carolina Institutional Review Board and University of Malawi College of Medicine Research Ethics Committee.

Study design and statistical analysis

We used an observational multi-cohort design. ART-naïve persons were eligible for analysis if they had a first clinic date occurring between January 1, 2000 and August 30, 2010 at QECH and January 1, 2007 and August 30, 2010 at LT,and were followed for at least 90 days. The event of interest was a first primary cancer occurring greater than 90 days after ART start. Cancers were identified either through registry linkage or from the cohort EMRs as part of the medical workup for WHO clinical staging and patient monitoring. We further classified new cancer as early incident (4 to 24 months after ART start) and late incident (>24 months). Person-years (py) at risk for incident cancer was calculated from 90 days after ART enrollment to the earliest of cancer diagnosis or censor due to ART cessation, clinic transfer, last contact, death, or October 1, 2015 administrative censor. For those lost to follow-up, person-time at risk included a 180-day window past the missed appointment date. Cancer diagnoses that were linked beyond the last date of contact or 180-day window were excluded from primary analyses (n=36).

Subsequent multiple primaries of different anatomical site or histology (n=32) and prevalent malignancies defined as a diagnosis prior to enrollment or within 90 days of ART start (n= 3,463) were excluded. Cancer data were coded using the International Classification of Disease for Oncology (Table 5-4).[133]

Incidence rates (IR) and 95% confidence intervals (CI) were estimated with Poisson regression, separately for each cohort, sex, individual cancer sites, early versus late incidence periods, and WHO stage at clinic enrollment. We applied direct standardization for age (0-15, 16-25, 25-35, 36-45, 46-55, 56+ years) and sex (male, female) using population weights from the combined cohorts. For sex-specific cancers, we used age-standardization for male and female populations, respectively. We further conducted a sensitivity analysis using only cancer matches with the highest linkage weights to estimate a conservative lower bound on IR (Table 5-6).

C. Results

Our study included 28,576 new ART users who initiated care at QECH (n=12,656; Figure 5-2) and LT (n=15,920; Figure 5-3). Median age at enrollment was 33 years (Table 5-1). New patients tended to initiate ART at an advanced WHO stage (LT stage 3: 41%, stage 4: 14%; QECH stage 3: 50%, stage 4: 16%). WHO stage distribution differed significantly between clinic sites (p<0·0001), where LT had greater proportion of persons with missing WHO stage information (16%) in the EMRs used for our study.

Overall, 4,346 cancers were identified: 16% were identified through record linkage (LT n=202; QECH n=477), 84% through the EMR (LT n=3351; QECH n=528). Pathological confirmation of cancer diagnosis was low and varied by clinic site: 3% at LT and 19% at QECH, reflecting diagnostic pathology availability in Lilongwe and Blantyre, respectively, during the study period. Prevalence of cancer at time of enrollment differed substantially by cohort: 18% of the LT cohort and 5% of QECH presented with malignancies, with prevalent cancers being predominantly ADC (98% and 87%, respectively).

A total of 23,655 cancer-free persons were followed for 100,815 py at risk (LT: n=12,464 individuals; n=49,981 py; QECH n=11,191; n=50,834 py; Table 5-1). Most incident cancers occurred within 4 to 24 months after starting ART (early incident: n=618; late incident n=265). Observed cancer incidence rates in our study varied in magnitude across cohorts, and therefore are presented separately rather than as a combined point estimate. Overall cancer incidence rate for all sites combined ranged from 488/100 000

py (CI: 431, 553) at QECH to 1,257/100,000 py (CI: 1,162, 1,359) at LT (Figure 5-1). Incidence was greatest during the early period of 4-24 months following ART initiation).

AIDS-defining cancers

KS, cervical cancer and NHL accounted for 98% of new malignancies among patients at LT and 85% at QECH (Table 5-4). The majority of KS and cervical cancer cases were diagnosed clinically; 10% and 27% received a pathological confirmation of diagnosis, respectively. Squamous cell carcinoma was the most common type of cervical cancer (42%), followed by non-specified histological types based on clinical diagnosis. All NHL cases were pathologically confirmed.

KS incidence rates ranged from 347 (CI: 288, 402) to 1,204 (CI: 1,111, 1,304) at QECH and LT, respectively (Figure 5-1; Table 5-5). KS occurred most frequently in the early incidence period of 4-24 months after starting ART for both men and women (Table 5-2). Early incident KS ranged from 413 to 964/100,000 py among men and 267 to 840/100,000 py among women (Table 5-2). Men and women at QECH experienced 2 to 9-fold increased incidence of KS at advanced WHO stage relative to early stage (Table 5-3); however, this association was not observed among LT patients.

Cervical cancer was the second most commonly occurring cancer, with an incidence rate ranging from 39 (CI: 22, 69) to 108 (CI: 77,153) at LT and QECH (Figure 5-1). No discernable pattern was observed in early versus late incidence of cervical cancer (Table 5-2). Women with advanced WHO stage had 30-80% lower rates of cervical cancer than women with early stage HIV (Table 5-3). NHL was detected at a low rate in our study (IR: 1.8 to 1.9).

Non-AIDS-defining cancers

NADC accounted for 15% of the total cancer burden at QECH and 2% at LT (Table 5-4). At QECH, the highest IR were for cancers of the esophagus (13.7), breast (12.9), female reproductive cancers (9.9), eye/conjunctiva (8.5), non-melanoma skin (6.3), penis (5.6), colorectum (4.1), unspecified lymphoma (3.9), uterus (3.5) and anus (3.0; Figure 5-1; Table 5-5. At LT, the highest IR were for bladder cancer (7.9), esophagus (2.2), eye/conjunctiva (1.9), unspecified lymphoma (1.9), larynx (1.9), and lip/oral cavity (1.9).

Pathological confirmation varied across NADC sites: esophagus (19%), cervix (27%), breast (53%), anus (75%), lymphomas (100%), oral cavity/pharynx (100%), eye/conjunctiva (100%). Squamous cell

carcinoma was the most common type of lip/oral cavity (40%), anus (75%), and eye/conjunctiva (85%) malignancy. Squamous cell carcinoma of the esophagus was also predominant (89%). Among breast cancers, 71% were infiltrating ductal carcinoma. Infection-associated cancers linked to *H. pylori* (stomach), hepatitis B and C virus (liver), schistosomiasis (bladder), and Epstein-Barr virus (Hodgkin lymphoma subtypes) were rarely detected in our study (Figure 5-1).

D. Discussion

Our goal in the Malawi HIV-Cancer Match Study was to characterize the burden and spectrum of cancer among ART users in Malawi, where HIV prevalence is 9% and one in twenty Malawian adults is now on ART.[14] In our study of nearly 29,000 ART users, the overall cancer burden was high and predominantly driven by ADC, even during the era of improving access to ART. KS was the most common cancer, and ADC were a common reason for presenting to care: 4% to 17% of patients presented to care with prevalent malignancies. The incidence of new KS was most pronounced during the first two years of ART, but remained high over long-term follow-up. Our KS incidence estimates in Malawi are among the highest for ART users in SSA, with other reported rates including 77 per 100,000 py in Zimbabwe, 169 in Zambia[172], 270 in Kenya, 340 in Uganda[173], and 432 in South Africa.[116] High incidence of KS at LT in 2007-2010 and QECH in 2000-2010 are similar in magnitude to ART users with CD4 count <50 (IR: 1,523 per 100,000py) and CD4 count 51-100 (IR: 716) in East African ART populations.[173] High KS burden in Malawi is likely attributable to the 35% to 88% prevalence of the causative virus human herpesvirus-8 (HHV-8) in Southern Africa[174], and typically advanced HIV stage with late presentation to care among ART initiators. This is especially true during the relatively early period of ART scale-up analyzed, which began in Malawi in 2004. KS burden may therefore still decline as the national ART program matures with earlier and more widespread application of ART.

Reflecting late presentation to care, advanced WHO stage was associated with increased KS incidence, although this association was not observed as strongly across both cohorts perhaps owing to differences in competing risk of death and prevalence of persons who already had KS at ART initiation. At QECH, advanced HIV stage was associated with a 2- to 10-fold increase in KS incidence. Cervical cancer was the second commonest cancer, and more advanced WHO stage was not associated with increased incidence of cervical cancer in our study. These findings might suggest KS burden will be more

immediately impacted by earlier ART application in SSA than cervical cancer, as also suggested by early epidemiological data from Uganda and Botswana showing modest incidence declines for KS but not cervical cancer.[11, 13]

We also observed a range of NADC even among Malawians with relatively advanced HIV prior to ART initiation. While NADC incidence rates were low overall, our findings highlight the heterogeneous cancer burden among Malawian ART initiators beyond KS and cervical cancer, as also suggested by other regional studies.[11, 13, 175] The highest incidence rates observed were for breast, esophageal, other female reproductive, eye/conjunctiva, and bladder cancers. Of these, only other female reproductive cancers and bladder cancer have confirmed etiologic associations with infectious pathogens (human papilloma virus and schistosomiasis, respectively), but associations with HIV in SSA are uncertain. For esophageal cancer, there is no known infectious etiology[176], although a Zambia case-control study suggested possible association with HIV[177]. For breast cancer, large studies from high-income countries have repeatedly shown reduced risk among HIV-infected persons.[165] However, SSA studies from Botswana and South Africa have reported HIV prevalence among breast cancer patients that is substantially higher than the general population. [25,32] Our work and that of others thus highlight the need for larger and more definitive epidemiological studies to define relationships between HIV and cancer which may be unique to SSA, to inform comprehensive, holistic cancer screening and prevention programs in regional ART clinics.

Differences in screening, diagnostic confirmation and referral patterns likely contributed to geographical variation in observable KS and cervical cancer incidence in our cohorts, and overall completeness of cancer ascertainment by the registry. This observation is similar to the overall context of cancer ascertainment in SSA, for which regional limitations have been extensively described.[178] The lower than expected incidence of lymphomas and NADC in our study is likely due to underdiagnosis. During the study period, a single pathology center in Blantyre provided services to the entire country, and only approximately one-fifth of all cancer cases in the registry were thus pathologically confirmed. Lymphomas may be particularly susceptible to misdiagnosis, and studies from Uganda and Malawi have shown that lymphomas are commonly clinically misdiagnosed as tuberculosis.[179,180] Our study may also be subject to underreporting of cancer incidence during late patient follow-up. The last registry survey

occurred in 2010; therefore, diagnoses occurring during follow-up beyond 2010 are not captured by the registry. Furthermore, study cohorts are young, and longer follow-up and additional linkages may be needed to monitor these populations as they age into a demographic group where NADC are most common. In Malawi, a demographic shift among people living with HIV is underway, where patients older than 50 years represent a growing proportion of ART users.[167]

The Malawi HIV-Cancer Match Study used probabilistic record linkage algorithms to ascertain cancer outcomes at centers of excellence for HIV care. This is one of the largest epidemiological studies of its kind in SSA to provide a comprehensive overview of the cancer burden among persons receiving ART. We used information on active patient follow-up to construct a retrospective cohort of approximately 29,000 new ART initiators. Given typically low-quality data sources for robust epidemiological studies in SSA, innovative approaches are needed to overcome cancer surveillance obstacles. Our study leveraged probabilistic methods and extensive clinical review to link data from Malawi's national cancer registry with existing electronic medical systems supporting ART delivery within large HIV clinics. This strategy is highly efficient for Malawi in terms of time and cost and imposed little additional burden on the health care system. We performed validation of cancer outcomes through extensive clerical and clinical review of matched cases. Our results on the occurrence of a broad spectrum of malignancies are an important baseline against which to monitor potential future shifts in evolving cancer burden among persons living with HIV.

Our study has downstream implications for strengthening health systems in Malawi, and improving data quality and completeness is a long-term priority. Our study used high quality data from HIV clinics participating in the International epidemiologic Databases to Evaluate AIDS (IeDEA) consortium. However, a substantial proportion of KS, cervical cancer and lymphoma may not be systematically recorded in electronic monitoring outside of HIV centers of excellence. Querying new data sources and validation studies using other clinics for HIV, palliative care, and women's health may improve cancer surveillance among HIV-infected populations in SSA. Together, these lessons underscore the importance of interdisciplinary collaboration[181] between HIV and cancer systems to build efficient and complete public health data resources in low-income settings.

In conclusion, we provide the first comprehensive baseline description of cancer burden against which to monitor cancer control efforts for HIV-infected populations in Malawi. Our findings demonstrate an ongoing high burden of KS and cervical cancer in a young, urban patient population, and the importance of integrated screening and management of KS and cervical cancer in ART programs. Validation of our findings through companion studies in other parts of SSA is needed, as well as longer-term studies to monitor potential shifts in cancer distribution with ART scale-up. Continued investment in high-quality cancer surveillance will be essential to inform national cancer control efforts in SSA.

Table 5-1. Characteristics of naïve ART initiators enrolled at Lighthouse Trust HIV Clinic (2007-2010) and Queen Elizabeth Hospital HIV clinic (2000-2010).

	LT (2007-2010)				QECH (2000-2010)			
	Total cohort		Person-years at risk		Total cohort		Person-years at risk	
	N	%	N	%	N	%	N	%
Total	15920		49980		12656		50833	
Sex								
Male	6713	42%	19557	39.1%	5529	43.7%	20590	40.5%
Female	9207	58%	30422	60.9%	7127	56.3%	30243	59.5%
Age category (years)								
<16	706	5%	2181	5%	1730	14%	6182	12%
16-25	1641	11%	4881	10%	1203	910%	4048	8%
26-35	6273	41%	19980	42%	4628	37%	18641	37%
36-45	4527	30%	14421	30%	3283	26%	14217	28%
46-55	1624	11%	5012	10%	1321	10%	5799	11%
56+	595	4%	1652	3%	491	‘%	1944	4%
missing	554	-	1851	-		-		-
Age at enrollment, years, median (IQR)	33.3 (27.9, 39.8)				33.5 (16.7, 40.9)			
WHO stage								
1 or 2	4852	31%	17693	35%	4181	33%	18792	37%
3	6499	41%	19362	39%	6261	50%	26122	51%
4	2207	14%	4855	10%	2010	16%	5476	11%
Not applicable/ unknown	2362	15%	8070	16%	204	2%	441	0.4%

Figure 5-1. Site-specific cancer incidence rates, by HIV clinical cohort

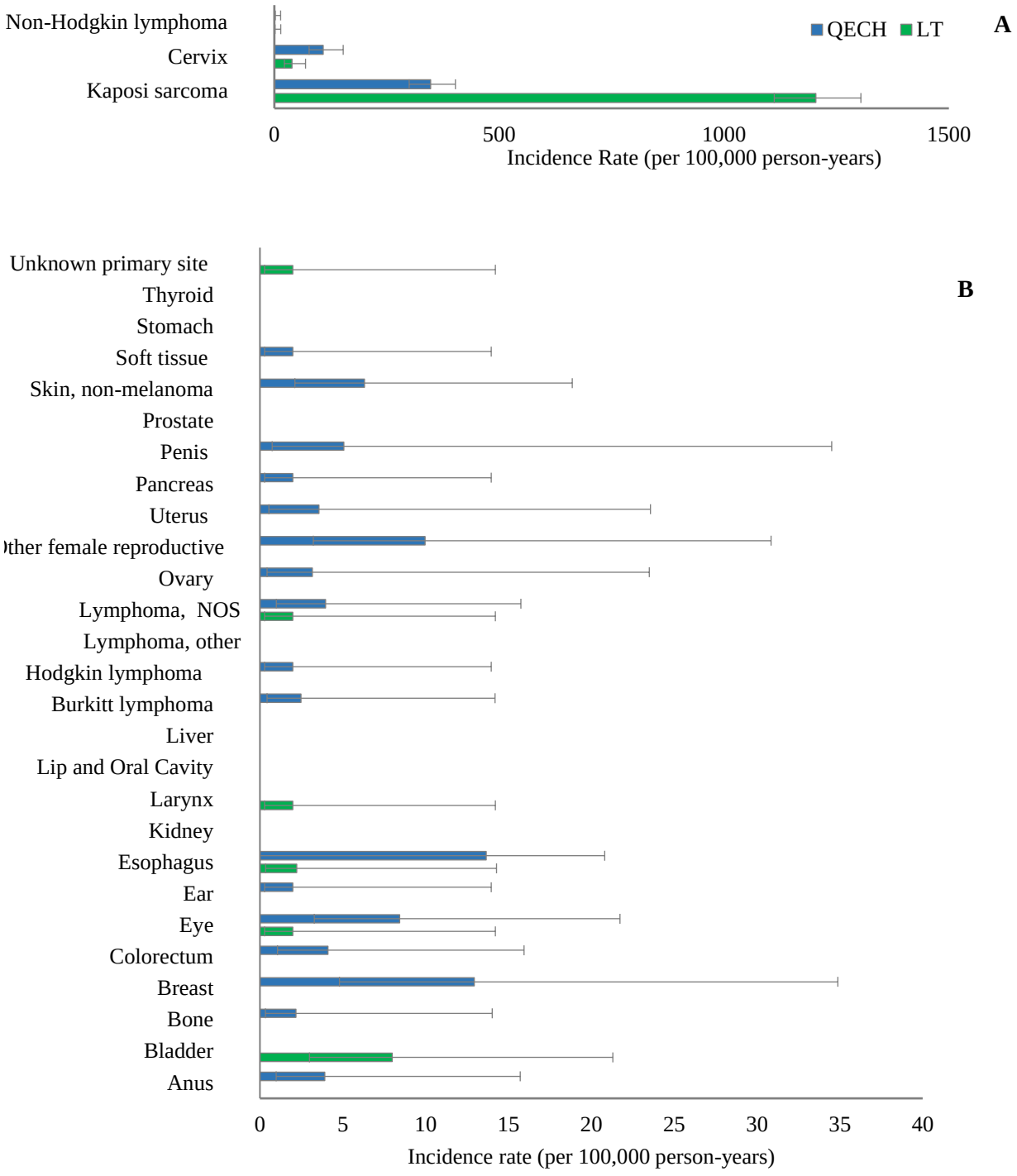


Table 5-2. Cancer incidence rates by timing of diagnosis after ART initiation

	Men				Women			
	LH		QECH		LH		QECH	
	IR	(95% CI)	IR	(95% CI)	IR	(95% CI)	IR	(95% CI)
All sites, total	1255.0	(1107.7, 1421.8)	585.0	(489.3, 699.3)	1257.3	(1137.3, 1390.1)	423.0	(355.7, 503.1)
Early incidence	964.6	(836.6, 1112.1)	413.3	(334.2, 511.1)	840.2	(743.1, 949.9)	267.4	(215.0, 332.5)
Late incidence	290.4	(224.0, 376.4)	171.7	(123.5, 238.7)	417.2	(350.5, 496.6)	155.7	(117.0, 207.2)
Kaposi sarcoma								
Early incidence	926.1	(800.8, 1070.9)	388.5	(312.0, 483.7)	797.8	(703.3, 904.9)	184.7	(142.0, 240.3)
Late incidence	280.7	(215.6, 365.4)	140.5	(97.6, 202.3)	404.7	(339.0, 483.0)	44.3	(25.9, 75.8)
Cervical cancer								
Early incidence	-		-		30.1	(15.7, 57.9)	43.1	(25.0, 74.3)
Late incidence	-		-		9.8	(3.1, 30.8)	66.2	(42.7, 102.8)

Early incidence: 4 - 24 months after beginning ART; late incidence: >24 months after beginning ART
IR: incidence rates per 100,000 person-years on ART, direct standardized for sex-age categories using combined cohorts as standard population
95% CI : 95% Confidence Interval
Female cancers were direct standardized for age, separately by sex

Table 5-3. Cancer incidence rates and incidence ratios by WHO clinical stage

	Men						Women					
	LH			QECH			LH			QECH		
Kaposi sarcoma	IR	(95% CI)	IRR	IR	(95% CI)	IRR	IR	(95% CI)	IRR	IR	(95% CI)	IRR
stage 1 or 2	393.4	(315.3, 490.9)	1.	43.7	(22.6, 84.6)	1.	299.5	(243.8, 367.8)	1.	40.4	(23.1, 70.8)	1.
stage 3	236.2	(177.5, 314.3)	0.6	86.1	(53.8, 137.8)	2.0	138.9	(102.5, 188.1)	0.5	38.6	(21.8, 68.1)	1.0
stage 4	568.5	(472.9, 683.4)	1.4	399.8	(321.4, 497.4)	9.2	25.4	(12.5, 51.7)	0.1	149.0	(111.2, 240.3)	3.7
Cervical cancer												
stage 1 or 2	-	-		-	-		19.0	(8.4, 43.0)	1.	44.3	(25.9, 75.7)	1.
stage 3	-	-		-	-		13.1	(4.9, 35.0)	0.7	57.3	(35.8, 91.8)	1.3
stage 4	-	-		-	-		-	-	-	7.1	(1.9, 27.1)	0.2

IR: incidence rates per 100,000 person-years on ART, direct standardized for sex-age categories using combined cohorts as standard population
95% CI : 95% Confidence Interval
IRR: Incidence Rate Ratio
Cervical cancer IR were direct standardized for age in the female population

E. Supplementary materials

Figure 5-2. Flowchart of cancer registry linkage to Queen Elizabeth Central Hospital

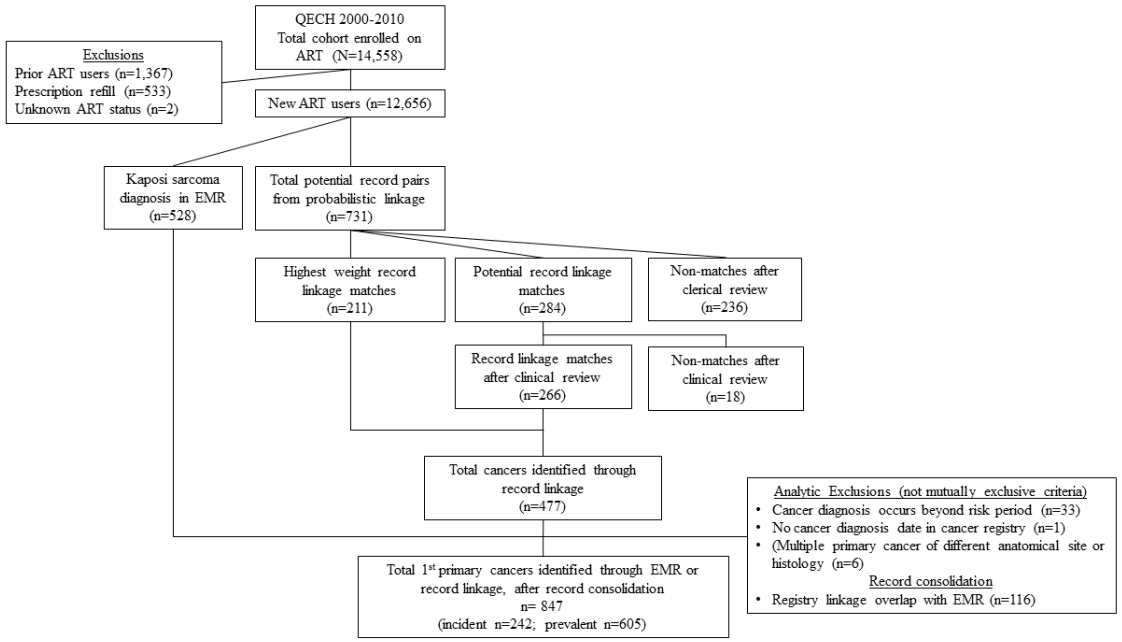
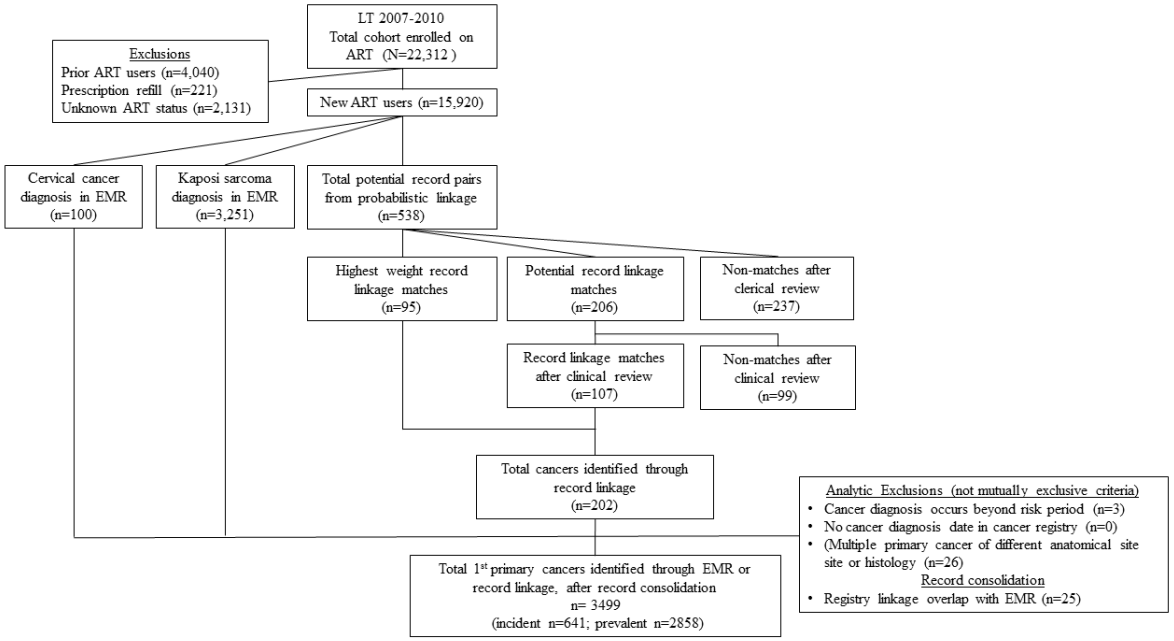


Figure 5-3. Flowchart of cancer registry linkage to Lighthouse Trust



Highest weight classification and validation

The analytic team set the final weight thresholds using iterative review of a random sample of record pairs and manually reviewed potential matches while blinded to the cancer site. Highest weight classification was used to group record pairs into “definite matches” (weight ≥23) and “potential matches” requiring further review (weights 12 to 23), and “definite non-matches” (weights <12). In a first round of clerical review, date of HIV clinic enrollment, date of cancer diagnosis, date of death, location of cancer diagnosis, and a Malawi-specific names thesaurus were used to adjudicate potential matches. Potential matches with missing birth year were excluded due to insufficient information. A second round of clinical review was conducted by three senior Malawian clinicians using additional oncology treatment data collected by the cancer registry and National Oncology Review Board. Potential matches were reviewed for biologic and clinical plausibility using additional information 1) pertaining to the cancer diagnosis: date of diagnosis, age, cancer type, cancer histology, tumor behavior, basis of diagnosis, hospital facility, treatment, sex; and 2) information pertaining to the HIV cohort: age at start of ART, AIDS-defining cancer diagnosis recorded by the facility, if any, time between MCR cancer diagnosis and ART start, last date of patient contact and outcome. The final match outcome after clinical review were categorized as match, non-match, and equivocal. Equivocal conclusions were recoded as non-matches in the final analytic dataset.

Table 5-4. Frequency of observed cancers by HIV clinical cohort

			LT					
			Incident		Prevalent		Total	
Primary site	ICD-O site	ICD-O morphology	N	%	N	%	N	%
Kaposi sarcoma	-	9140	614	(95.8%)	2709	(94.8%)	3323	(95.0%)
Cervix uteri	C53.0-C53.9	8000, 8010, 8070, 8071, 8140, 8384	14	(2.2%)	115	(4.0%)	129	(3.7%)
Non-Hodgkin lymphoma	-	9591	1	(0.2%)	0	(0.0%)	1	(0.0%)
Anus	C21.0-C21.8	8070, 8140	0	(0.0%)	2	(0.1%)	2	(0.1%)
Bladder	C67.0-C67.9	8070	4	(0.6%)	0	(0.0%)	4	(0.1%)
Bone	C40.0-C40.9	8810, 9731	0	(0.0%)	0	(0.0%)	0	(0.0%)
Breast	C50.0-C50.9	8010, 8140, 8500	0	(0.0%)	4	(0.1%)	4	(0.1%)
Colorectum	C18.2-C20.9	8140	0	(0.0%)	1	(0.0%)	1	(0.0%)
Conjunctiva, Retina, Eye	C69.0-C69.9	8010, 8070, 9510	1	(0.2%)	3	(0.1%)	4	(0.1%)
Ear	C30.1	8001	0	(0.0%)	0	(0.0%)	0	(0.0%)
Esophagus	C15.0-C15.0	8000, 8010, 8070	1	(0.2%)	6	(0.2%)	7	(0.2%)
Kidney	64.9	8321, 8960	0	(0.0%)	0	(0.0%)	0	(0.0%)
Larynx	C32.0-C32.9	8070	1	(0.2%)	0	(0.0%)	1	(0.0%)
Lip, Oral Cavity	C00.0-C10.1	8010, 8070, 8071, 8200, 8941	0	(0.0%)	3	(0.1%)	3	(0.1%)
Liver	C22.0-C22.1	8170	0	(0.0%)	1	(0.0%)	1	(0.0%)
Lymphoma Burkitt lymphoma	-	9687	0	(0.0%)	1	(0.0%)	1	(0.0%)
Hodgkin lymphoma	-	9650	0	(0.0%)	0	(0.0%)	0	(0.0%)
Lymphoma, other specified	-	9714, 9727	0	(0.0%)	0	(0.0%)	0	(0.0%)
Lymphoma, unspecified	-	9590	1	(0.2%)	2	(0.1%)	3	(0.1%)
Ovary	C56.9	8070, 8140, 9100	0	(0.0%)	1	(0.0%)	1	(0.0%)
Other female reproductive, Vulva, Vagina	C57.0-C57.0, C51.0-C52.9	8010, 8070	3	(0.5%)	1	(0.0%)	4	(0.1%)
Uterus	C54.0-C54.9		0	(0.0%)	2	(0.1%)	2	(0.1%)
Pancreas	C25.0-C25.0	8140	0	(0.0%)	0	(0.0%)	0	(0.0%)

			LT					
			Incident		Prevalent		Total	
Primary site	ICD-O site	ICD-O morphology	N	%	N	%	N	%
Penis	C60.0-C60.9	8070	0	(0.0%)	0	(0.0%)	0	(0.0%)
Prostate	C61.9	8140	0	(0.0%)	0	(0.0%)	0	(0.0%)
Skin malignancies, non-melanoma	C44.0-C44.9	8070, 8409,8832,8833, 9590	0	(0.0%)	5	(0.2%)	5	(0.1%)
Soft tissue	C49.0-C49.9	8002, 8901	0	(0.0%)	0	(0.0%)	0	(0.0%)
Stomach	C16.0-C16.9	8240	0	(0.0%)	1	(0.0%)	1	(0.0%)
Thyroid	C73.9	8331	0	(0.0%)	1	(0.0%)	1	(0.0%)
Unknown primary site	C80.9	8000, 8070	1	(0.2%)	0	(0.0%)	1	(0.0%)
Total			641		2858		3499	(100.0%)

			QECH					
			Incident		Prevalent		Total	
Primary site	ICD-O site	ICD-O morphology	N	%	N	%	N	%
Kaposi sarcoma	-	9140	172	(71.1%)	493	(81.5%)	665	(78.5%)
Cervix uteri	C53.0-C53.9	8000, 8010, 8070, 8071, 8140, 8384	32	(13.2%)	32	(5.3%)	64	(7.6%)
Non-Hodgkin lymphoma	-	9591	1	(0.4%)	8	(1.3%)	9	(1.1%)
Anus	C21.0-C21.8	8070, 8140	2	(0.8%)	0	(0.0%)	2	(0.2%)
Bladder	C67.0-C67.9	8070	0	(0.0%)	0	(0.0%)	0	(0.0%)
Bone	C40.0-C40.9	8810, 9731	1	(0.4%)	1	(0.2%)	2	(0.2%)
Breast	C50.0-C50.9	8010, 8140, 8500	4	(1.7%)	9	(1.5%)	13	(1.5%)
Colorectum	C18.2-C20.9	8140	2	(0.8%)	1	(0.2%)	3	(0.4%)
Conjunctiva, Retina, Eye	C69.0-C69.9	8010, 8070, 9510	4	(1.7%)	18	(3.0%)	22	(2.6%)
Ear	C30.1	8001	1	(0.4%)	0	(0.0%)	1	(0.1%)
Esophagus	C15.0-C15.0	8000, 8010, 8070	7	(2.9%)	4	(0.7%)	11	(1.3%)
Kidney	64.9	8321, 8960	0	(0.0%)	3	(0.5%)	3	(0.4%)
Larynx	C32.0-C32.9	8070	0	(0.0%)	0	(0.0%)	0	(0.0%)
Lip, Oral Cavity	C00.0-C10.1	8010, 8070, 8071, 8200, 8941	0	(0.0%)	2	(0.3%)	2	(0.2%)
Liver	C22.0-C22.1	8170	0	(0.0%)	2	(0.3%)	2	(0.2%)
Lymphoma Burkitt lymphoma	-	9687	2	(0.8%)	0	(0.0%)	2	(0.2%)
Hodgkin lymphoma	-	9650	1	(0.4%)	3	(0.5%)	4	(0.5%)
Lymphoma, other specified	-	9714, 9727	0	(0.0%)	2	(0.3%)	2	(0.2%)
Lymphoma, unspecified	-	9590	2	(0.8%)	8	(1.3%)	10	(1.2%)

			QECH					
			Incident		Prevalent		Total	
Primary site	ICD-O site	ICD-O morphology	N	%	N	%	N	%
Ovary	C56.9	8070, 8140, 9100	1	(0.4%)	6	(1.0%)	7	(0.8%)
Other female reproductive, Vulva, Vagina	C57.0-C57.0, C51.0-C52.9	8010, 8070	3	(1.2%)	1	(0.2%)	4	(0.5%)
Uterus	C54.0-C54.9		1	(0.4%)	1	(0.2%)	2	(0.2%)
Pancreas	C25.0-C25.0	8140	1	(0.4%)	0	(0.0%)	1	(0.1%)
Penis	C60.0-C60.9	8070	1	(0.4%)	1	(0.2%)	2	(0.2%)
Prostate	C61.9	8140	0	(0.0%)	2	(0.3%)	2	(0.2%)
Skin malignancies, non-melanoma	C44.0-C44.9	8070, 8409,8832,8833, 9590	3	(1.2%)	4	(0.7%)	7	(0.8%)
Soft tissue	C49.0-C49.9	8002, 8901	1	(0.4%)	1	(0.2%)	2	(0.2%)
Stomach	C16.0-C16.9	8240	0	(0.0%)	0	(0.0%)	0	(0.0%)
Thyroid	C73.9	8331	0	(0.0%)	0	(0.0%)	0	(0.0%)
Unknown primary site	C80.9	8000, 8070	0	(0.0%)	3	(0.5%)	3	(0.4%)
Total			242		605		847	(100.0%)

Table 5-5. Cancer incidence rates by HIV clinical cohort

	LH (2007-2010)			QECH (2000-2010)		
	IR	**(95% CI)**	**Rank**	**IR**	**(95% CI)**	**Rank**
All sites	1257.1	(1162.5, 1359.4)		488.3	(431.2, 553.0)	
Kaposi Sarcoma	1204.2	(1111.7, 1304.3)	1	347.5	(299.8, 402.7)	1
Cervix	39.3	(22.3, 69.4)	2	108.8	(77.3, 153.1)	2
Non-Hodgkin lymphoma	1.8	(0.2, 14.2)	9	1.9	(0.3, 14.0)	20
Anus	-			3.9	(1.0, 15.7)	11
Bladder	8.0	(3.0, 21.3)	3	-		
Bone	-			2.2		15
Breast	-			12.9	(4.8, 34.9)	4
Colorectum	-			4.1	(1.1, 15.9)	9
Conjunctiva, Retina, Eye	2.0	(0.3, 14.2)	5	8.4	(3.3, 21.7)	6
Ear	-			2.0	(0.3, 14.0)	16
Esophagus	2.2	(0.3, 14.3)	4	13.7	(6.5, 28.7)	3
Kidney	-			-		
Larynx	2.0	(0.3, 14.2)	5	-		
Lip, Oral Cavity	-			-		
Liver	-			-		
Lymphoma						
Burkitt lymphoma	-			2.5	(0.4, 14.2)	14
Hodgkin lymphoma	-			2.0	(0.3, 14.0)	16
Lymphoma, other specified	-			-		
Lymphoma, not otherwise specified	2.0	(0.3, 14.2)	5	4.0	(1.0, 15.8)	10
Ovary	-			3.2	(0.4, 23.5)	13
Other female reproductive, Vulva, Vagina	-			10.0	(3.2, 30.9)	5
Uterus	-			3.6	(0.5, 23.6)	12
Pancreas	-			2.0	(0.3, 14.0)	16
Penis	-			5.1	(0.7, 34.5)	8
Prostate	-			-		
Skin malignancies, non-melanoma	-			6.3	(2.1, 18.9)	7
Soft tissue	-			2.0	(0.3, 14.0)	16
Stomach	-			-		
Thyroid	-			-		
Unknown primary site	2.0	(0.3, 14.2)	5	-		

IR: Incidence rates per 100,000 person-years on ART, direct standardized for sex-age categories using combined cohorts as standard population; 95% CI : 95% Confidence Interval
Female and male cancers were direct standardized for age, separately by sex

Sensitivity analysis

We estimated a conservative lower bound of cancer incidence rates using only “definite matches” identified through highest weight classification or the HIV clinic electronic medical record (LT n=3420; QECH n= 657).

Table 5-6. Sensitivity analysis: site-specific cancer incidence rates, by HIV clinical cohort

	LH (2007-2010)			QECH (2000-2010)		
	IR	**(95% CI)**	**Rank**	**IR**	**(95% CI)**	**Rank**
All sites	1224.5	(1131.2, 1325.5)		359.9	(311.4, 416.0)	
Kaposi Sarcoma	1194.9	(1102.8, 1294.7)	1	295.5	(251.8, 346.7)	1
Cervix	20.0	(9.0, 44.3)	2	45.9	(27.1, 77.7)	2
Non-Hodgkin lymphoma	1.8	(0.2, 14.2)	6	-	-	
Anus	-			3.9	(1.0, 15.7)	7
Bladder	4.0	(1.0, 16.0)	3	-		
Bone	-			-	-	
Breast	-			9.9	(3.2, 30.7)	3
Colorectum	-			2.2	(0.3, 14.0)	
Conjunctiva, Retina, Eye	2.0	(0.3, 14.2)	5	4.3	(1.2, 16.2)	6
Ear	-			-		
Esophagus	2.2	(0.3, 14.3)	4	5.7	(1.8, 18.0)	5
Kidney	-			-		
Larynx	-			-		
Lip, Oral Cavity	-			-		
Liver	-			-		
Lymphoma						
Burkitt lymphoma	-			-		
Hodgkin lymphoma	-			2.0	(0.3, 14.0)	10
Lymphoma, other specified	-			-		
Lymphoma, not otherwise specified	-			2.0	(0.3, 14.0)	10
Ovary	-			-		
Other female reproductive, Vulva, Vagina	-			7.1	(1.9, 27.1)	4
Uterus	-			3.6	(0.5, 23.6)	8
Pancreas	-			2.0	(0.3, 14.0)	10
Penis	-			-		
Prostate	-			-		
Skin malignancies, non-melanoma	-			2.2	(0.3, 14.0)	9
Soft tissue	-			-		
Stomach	-			-		
Thyroid	-			-		
Unknown primary site	-			-		

IR: Incidence rates per 100,000 person-years on ART, direct standardized for sex-age categories using combined cohorts as standard population; 95% CI : 95% Confidence Interval
Female and male cancers were direct standardized for age, separately by sex

CHAPTER 6. CONCLUSIONS

A. Summary of findings

We conducted one of the largest epidemiological studies of its kind in SSA to provide a comprehensive overview of the spectrum of cancers among people receiving ART. Our multicohort study is comprised of nearly 29,000 actively traced patients receiving ART and is a unique resource for SSA, where high-quality cancer registries cover only 1% of the population with HIV status rarely captured. To overcome local constraints in healthcare infrastructure, we applied innovative and efficient data solutions to link cases from the Malawi Cancer Registry with centers of excellence for HIV care in Malawi. Our statistical design meant to address the primary analytic challenges encountered in our setting: missing data, uncertain linkage because of loss of resolution for patient identifiers in Malawi and other factors, and absence of a gold standard dataset for validation of outcomes. Such issues related to data quality and completeness are practical concerns for the design of linkage studies, particularly in resource-limited environments. Missing data and misreporting of patient identifiers likely resulted in a high number of records that failed to link, and perhaps to a lesser degree, false positive links. Consequently, missing data on potential matches may have diminished sensitivity of the linkage algorithms, although quantifying the impact of missing data on sensitivity and specificity was not feasible due to lack of a gold-standard dataset with which to cross-validate our results. We therefore employed an additional review step using supplemental information from the medical record as a field-tested approach to address linkage measurement error in a rigorous way when sufficient secondary clinical information was available. Our study is a proof of concept of a linkage study design tailored to observational, routinely collected data in SSA.

We identified 883 incident malignancies among 23,655 persons who were cancer-free at time of enrollment and followed for 100,815 person-years. KS and cervical cancer were the most common cancers in this young population of ART users who tend to present to care with severe immunosuppression. Rates of KS are among the highest reported to date in the region and are of similar

magnitude to what is observed among profoundly immunocompromised people in East Africa. Regarding the timing of disease, most incident KS occurred within the first two years of starting ART, and elevated incidence rates persisted over the course of patient follow-up in spite of therapy. Men and women presenting with advanced stage HIV at the start of ART experienced elevated incidence of KS relative to those with early stage of disease. However, more advanced WHO clinical stage was not associated with increased incidence of cervical cancer in our study. Our findings suggest that the burden of KS burden may be more immediately impacted by earlier ART application in Malawi than cervical cancer, which is consistent with data from Uganda and Botswana showing modest incidence declines for KS but not cervical cancer.[11, 13] Our results not withstanding, earlier application of ART, and therefore the timing of earlier immune reconstitution relative to HPV infections and cervical lesions, is critical to control the progression of neoplastic cervical lesions to invasive cancer. [182]

AIDS-associated non-Hodgkin lymphomas and a heterogeneous spectrum of NADC were also observed, but at low incidence rates. The overall young age distribution of the study population and underdiagnosis resulting from limited cancer diagnostic capacity in Malawi are likely factors that lead to low rates of NADC and lymphomas observed in our study.

B. Public health implications

Recent attention has focused on evolving cancer risk in aging HIV populations in the United States.[165] An emerging public health question is whether these cancer trends will be replicated in SSA now that access to ART has become more widespread, and demographic shifts among African ART populations are underway. Epidemiologic studies addressing this question have been few within the region, and historically, have been hampered by a paucity of data. Public health data linkages are a cost-efficient strategy for bridging the knowledge gaps between clinical and population sciences and infectious disease and noncommunicable disease research silos in SSA.

Our study shows that cancer burden among Malawian ART users does not yet mirror high-income countries. Malawi bears a high burden of AIDS-defining malignancies among young patient populations of ART users. It is worth emphasizing that Malawi's HIV treatment program is still maturing relative to that of high income countries, where HAART became widely implemented into clinical practice in 1996. In contrast, Malawi's national ART program began nearly a decade later in 2004, with ART coverage

reaching half of all eligible patients by 2011. In our study, median age at the start of therapy was 33 years and mean time on ART was 4.5 years, which together may not be a sufficient window of time to observe reductions in KS incidence and the development of NADC in our cohorts. Our results reflect clinical practice through 2010, when ART was reserved for those with severe immunosuppression. Therapy guidelines have since been revised to include patients with higher CD4 counts.

Our work has implications for monitoring the impact of earlier application of ART on KS incidence in Malawi, where expanded eligibility criteria among adults now include WHO stage 1 or 2 and CD4 counts ≤500 cells/mm^3 , WHO stage 3 or 4 regardless of CD4 count, and universal ART for HIV-infected pregnant and breastfeeding women (Option B+).[16] ART plays a vital role in the treatment and primary prevention of AIDS-related KS, and improves survival among KS patients.[183] Early application of ART has been shown to dramatically reduce KS incidence at population level in ecologic studies[11] and at the individual level in IeDEA and other prospective cohorts in Africa[172, 173, 184], but additional studies are needed to confirm and validate the impact in other SSA populations. East African ART users in Uganda and Kenya experienced 80% and 50% reductions in KS incidence rates compared to HIV-infected ART-naïve persons, though absolute rates of KS remained very high at CD4 counts <350 cells/mm^3 even among ART users, and significantly elevated relative to ART users in high-income settings.[173]

Our work also has implications for monitoring the impact of cervical cancer screening and management programs among ART users in Malawi. Effective screening and early treatment of precancerous cervical lesions are critical to reduce the burden of cervical cancer. In low- and middle-income countries, WHO recommends a “screen and treat” approach of visual inspection with acetic acid (VIA) and crytotherapy or LEEP for premalignant lesions during a single-day visit, and screening for all women and girls with HIV upon a positive HIV result.[185] Malawi’s cervical cancer screening program began in the 1980s, and since 2004, Malawi’s Ministry of Health introduced a program to scale-up VIA and cryotherapy treatment of precancerous lesions to all 29 district hospitals and 3 central hospitals.[186] The program targets women ages 30-49 and those receiving ART, with recommended screening once every five years for HIV-uninfected women and once every two to three years for HIV-infected women. There are now more than 100 facilities currently delivering VIA, 32 offering cryotherapy, and 3 offering

LEEP in Malawi.[187, 188] However, uptake of services remains low with approximately one in four eligible women accessing screening.[189]

Recently, some have called for a standalone policy for cervical cancer prevention and treatment in Malawi. [189,190, 191] Many policy and health systems challenges persist.[192] Current policy does not specifically address poor screening uptake among vulnerable populations such as women living with HIV nor screening of women younger than 30 who live with HIV. Another major concern is failure to treat a large proportion of VIA-positive women with cryotherapy/LEEP. [2,189] Lastly, there is no quality assurance or mandate currently to monitor the effectiveness of the screening program.[193]

Further strengthening and integrating cervical cancer screening into HIV ART programs is a critical direction for cancer control and prevention in the era of ART.[194] Women on ART are under regular clinical observation and therefore ART facilities may help address the challenge of high lost-to-follow rates of people who do not return for a one year follow-up visit after cryotherapy.[189] Operationalizing further data linkages is also an opportunity to monitor and evaluate the effectiveness of screening and continued ART scale-up on invasive cervical cancer incidence in Malawi. Updating the design of the cancer registry to include VIA data from sentinel HIV clinics would contribute towards the goal of more complete cancer surveillance and evidence-based evaluation of screening programs for women living with HIV in Malawi.

C. Strengths

Our study has several design strengths which improve upon the limitations of previous work describing HIV-associated cancer trends in SSA. Linkage of existing data from EMRs is efficient in terms of time and cost, and imposes no additional burden on routine delivery of care by local clinicians. We used data from the Malawi cancer registry, which is one of only four population-based cancer registries in SSA that contributes to global estimates of disease in IARC's *Cancer In V Continents.*[20]

Our longitudinal cohort design uses a well-enumerated denominator of HIV patients. The selected ART cohorts are participating members in the IeDEA worldwide consortium. Since the registry does not routinely collect HIV status, we linked patient-level data from the largest tertiary ART clinics within Malawi to the cancer registry. This new dataset incorporates HIV clinical assessments, drug regimens, active follow-up, and patient vital status. The study's large sample size was advantageous for exploring the distribution of site-specific non-AIDS defining cancers. Similarly, active patient follow-up in the cohorts

provided a unique opportunity to study the timing of cancer outcomes in a population of ART users in SSA.

Our study has several analytic strengths that contribute new insights into the implementation of healthcare record linkages for resource-limited settings. Our approach to handling record linkage measurement error is transparent, and tailored to directly address data concerns and ethnographic considerations specific to the SSA research setting, and may serve as a practical guide to other groups seeking to implement and scale healthcare linkages in low- and middle-income countries.

D. Limitations

Our study does not have a gold-standard dataset against which to validate the record linkage performance. The majority of cancer diagnoses recorded in the cancer registry and in the ART cohorts do not overlap due to local patterns of care in Malawi. Therefore, the construction of an *ad hoc* validation dataset was not possible.

Possible cancer underascertainment at both ends of the study period is a limitation to our estimations of cancer prevalence and incidence rates. The 2000-2003 period may underestimate the true burden of *prevalent* NADC among patients receiving ART in Blantyre. Prior to 2000, coverage by the registry was restricted to the immediate vicinity of Blantyre city and was mostly pathology-based. Therefore, HIV patients who received a clinical cancer diagnosis only or those who were diagnosed with a NADC outside of QECH prior to starting ART were not captured by the registry. Clinically diagnosed ADC, which are used for WHO staging of HIV/AIDS, in principle would still be recorded in the ART cohorts, therefore prevalent ADC are less likely to be underreported. The 2007-2010 period is not able to capture late incident NADC developing after diagnosis year 2010, even though prospective follow-up in the ART cohorts extends beyond 2010. Therefore, the later period of the study captures only prevalent and early incident NADC diagnoses. Consequently, incidence rates of late incident cancer may be underreported for persons entering the cohorts during the later years of study.

In terms of generalizability, our study includes ART cohorts from centers of excellence for HIV care in the two largest cities in Malawi. The clinics are also adjacent to tertiary centers where the majority of oncology care is delivered within the country. Our findings may not be representative of ART delivery and cancer care in rural settings. However, public health implications are expected to remain generalizable to

patients receiving ART at large clinics in Malawi. Findings in our descriptive epidemiological work did not address confounding by cancer treatment nor by behavioral risk factors such as smoking or alcohol abuse, since this information was not available. In Malawi, surveys report 14.6% prevalence of heavy drinking among men and 1.4% among women, and 18.0 % prevalence of tobacco smoking among men and 1.2% among women.[195, 196] However, the prevalence of these risk factors specifically among people living with HIV in Malawi has yet to be characterized. The study did not have data with which to examine cohort-level cancer screening effects for cervical cancer and KS.

There were contextual limitations to conducting this study in a low-income country. We were limited in our ability to correlate CD4 count and HIV viral load with cancer incidence, with consequently reduced ability to study important clinical predictors of cancer incidence. Limited CD4 count data and no HIV RNA data were available during the period of our study reflecting practice within the Malawi national HIV program. CD4 count was historically restricted to stage 1 and 2 patients who were not clinically eligible for ART (e.g. stages 3 and 4). For Queen Elizabeth Central Hospital prior to 2011, CD4 counts were not captured in the electronic monitoring system and therefore were not available for analysis. Routine HIV RNA monitoring in Malawi did not begin until 2011. Viral load is available for a small number of patients, thereby precluding its evaluation in our analyses. Due to the high cost of HIV RNA assessment ($20-60 per test), WHO guidelines historically have not recommended routine viral load testing among persons already on ART in low- and middle-income countries. Current WHO guidelines recommend viral load testing 6 months after starting ART, and then at 12 months and yearly if people have achieved viral suppression [197]; Malawi guidelines recommend viral load testing every other year for people receiving ART.[16] Continued efforts to improve capacity of viral load testing are underway in Malawi, where one in five ART patients received at least one viral load test in 2015-2016 and 82%-89% of those tested achieved viral suppression.[198] Viremia may contribute directly to cancer pathogenesis of KS and NHL, and possibly indirectly for virally associated NADCs including HL, anal cancer, and liver cancer .[40, 199-202] Moving forward, understanding the roles of long duration of viremia and viral suppression in cancer development may have practical implications for risk stratification among people receiving ART.

Our ability to detect lymphomas and certain solid tumors such as liver, lung, prostate, and colorectal cancers may be limited. These cancer sites may be underdiagnosed even in the general population

because the availability of pathology and medical imaging were limited in Malawi during the study period. Therefore, these types of cancers in the registry are likely an underestimate of the true burden of disease both in the overall population and in our study population. The cancer incidence rates we observed in our study may represent the “tip of the iceberg” of the disease burden which is truly occurring in Malawi. For similar reasons, detailed histology and cancer subtyping were not available for analysis. Despite these limitations, it is important to emphasize that the Malawi cancer registry remains one of the best regional resources for population-based data contributing to IARC global burden of disease estimates, in a part of the world where deficits in high-quality cancer registration have been abundantly described. [106, 203]

E. Conclusions and future directions

We implemented a healthcare data linkage in a low-resource setting to construct a *de novo* resource for public health research. We built a process for integrating health information from two independent sources, EMRs and a disease registry, and studied the validity and performance of this process in the field. Our approach uses evidence from observational HIV cohorts at high-volume clinics and locally collected cancer outcomes. In alignment with this approach, our findings reflect actual clinical practice and the realities of health systems in Malawi.

Results from the linkage have several implications for strengthening health systems in Malawi. First, the data linkage identified little overlap of KS and cervical cancer diagnoses between HIV EMRs and the Malawi cancer registry. Updating Malawi’s cancer surveillance design to include sentinel HIV clinics within a small, yet well-enumerated geographic region such as Blantyre or Lilongwe, would more thoroughly capture the full burden of ADC that are being diagnosed primarily in HIV point-of-care clinics, not district hospitals. Ideally, such a design would involve more direct data sharing through data transfers directly from EMR databases at large clinics and paper-based, cross-sectional collection at smaller clinics. Additionally, future cancer surveillance could incorporate KS diagnoses captured by the national ART registry, which is still being refined. Further harmonizing EMRs and paper-based records that feed into the contemporary ART national registry is important if the ART registry is to be used for public health research. At the time of our study, the Malawi ART registry was fraught with non-unique national IDs and duplication of patient records, among other issues. Our findings also emphasize improving the collection

of accurate and complete personal identifiers and demographics in EMRs, as well as the cancer registry, so long as this is not done at the expense of maintaining high quality clinical information.

Missing data in the EMR and cancer registry reduced performance of linkage algorithms. There are novel analytic approaches that could potentially strengthen the performance of algorithms in our subsequent work. An algorithm that uses distance imputation between two linkage fields, rather than imputation of missing values, estimates the extent of agreement or disagreement of two linkage fields when one or both are missing.[204] This type of imputation uses the records with full data to calculate a conditional probability of agreement of the fields with missing data in a given record pair, given similarity across the other linkage fields. A different approach may consider an algorithm that uses weight redistribution, which removes the weight assigned to the variable with missingness and redistributes it across the other linkage fields that are complete.[204]

We used the linked HIV-cancer dataset to describe the cancer burden among people receiving ART during the period of ART scale-up in Malawi. Further studies should be conducted to both validate and expand upon the descriptive epidemiology of cancer occurrence among contemporary HIV populations in Malawi. Malawi's national ART program is still young and our work focused on the period of early scale-up when eligibility for therapy was restricted to persons with very advanced HIV. Future work may update the data linkage so that contemporary HIV populations who are initiating therapy under expanded eligibility guidelines may be studied. Since 2010, revised guidelines now recommend starting therapy at higher CD4 counts and lifelong therapy for HIV-infected pregnant women regardless of HIV stage or CD4 count.

Furthermore, contemporary data from QECH and LT include information on CD4 levels and HIV viral load which may be used to explore associations with cancer development and replicate findings from Western settings. Associations between immunodeficiency and viremia have been described in European and American populations, where low CD4 counts are associated with virally- related NADCs [38-41], and long duration of immune suppression increases risk of HPV-related cancers.[43, 4] HIV viremia is associated with increased risk of NHL[205, 206], anal cancer[200-202], Hodgkin lymphoma[200, 201] . Describing associations between clinically relevant biomarkers and cancer development, specifically in Malawian ART patient populations, may be useful for cancer risk stratification during a time frame of current WHO guidelines.[197]

Our study is a baseline against which to monitor temporal trends and to compare with the contemporary burden of cancer among people who are starting therapy at earlier stages of HIV, when ART is more likely to have a substantial impact on cancer incidence. Descriptive epidemiological data, specifically focusing among people living with HIV, is critical for decision makers at the Malawi Ministry of Health to further develop evidence-based cancer control plans targeting high-risk HIV populations.

APPENDIX A. STANDARDIZED INCIDENCE RATIOS

Appendix A Table 1. Standardized incidence ratios for AIDS-defining cancers (2007-2010 only)

	LT					QECH				
	Crude rate	Reference rate (Center)	SIR (95%CI)		Standardized rate (95%CI)	Crude rate	Reference rate (South)	SIR (95%CI)		Standardized rate (95%CI)
Kaposi sarcoma	1224.4	36.2	14.5	(13.3, 15.6)	524.0 (482.4, 565.5)	148.2	66.8	1.0	(0.7, 1.3)	67.2 (47.1, 87.3)
Cervical cancer	39.4	54.6	0.4	(0.2, 0.6)	20.3 (8.8, 31.8)	106.6	73.1	0.7	(0.4, 1.02)	51.21 (27.6, 74.9)
NHL	2.0007	9.0	0.4	(0, 1.2)	3.76 (0, 11.2)	0	22.004	0		0 (0,0)

SIR: Standardized Incidence Ratio, sex-age indirect standardization with population weights from the 2008 census, using central and southern regions of Malawi as referent populations for LT and QECH, respectively
LT: Lighthouse Trust
QECH: Queen Elizabeth Central Hospital
NHL: Non-Hodgkin Lymphoma

Ancillary analyses will evaluate whether risk of NADC among ART users is elevated relative to the general population of Malawi. To date, we conducted preliminary analysis of standardized incidence ratios (SIRs) for KS and cervical cancer. We used sex-age indirect standardization with population weights from the 2008 census, with central and southern regions of Malawi as referent populations for LT and QECH, respectively. Compared to the general population, we observed elevated risk of KS among ART users at LT (SIR=14.5, 95%CI 13.3, 15.6) but paradoxically, not at QECH (SIR=1.0, 95%CI 0.7, 1.3). We observed a reduced risk for cervical cancer among ART users in both cohorts (LT SIR=0.4, 95%CI 0.2, 0.6; QECH SIR=0.7, 95%CI 0.4, 1.02).

APPENDIX B. CUMULATIVE RISK OF CANCER AMONG NEW ART USERS

We examined the cumulative risk of cancer in the presence of 1) competing risks and, 2) the large proportion of individuals who already have the disease at the start of follow-up. Ignoring high baseline prevalence of AIDS-defining cancers KS and cervical cancer, which are clinical indications to initiate ART, may lead to underestimation of the true risk of developing cancer among ART users. The overall age-sex-adjusted cancer incidence was calculated using direct standardization. We used a modified Kaplan-Meier approach to estimate adjusted cumulative incidence (ACI) without conditioning on event-free status at baseline enrollment and accounting for competing risk of death. The ACI was estimated at index ages 20-29, 30-39, and 40-49 years assuming a scenario of the youngest age of cancer diagnosis at 20 years and age-sex-specific cancer prevalence derived from the cohort. Age in discrete years was used as the time scale, and number of years at risk were a function of the width of the age group. Survival time was censored at the last date of contact or administratively on 31/12/2010.

Appendix B Table 1. Cancer prevalence at ART start (Lighthouse Trust)

	Ages 20-29 years		Ages 30-39 years		Ages 40-49 years	
	Naïve ART users (N)	Prevalence any cancer (%)	Naïve ART users (N)	Prevalence any cancer (%)	Naïve ART users (N	Prevalence any cancer (%)
Prevalence						
Men	1028	19.2	2969	21.6	1592	22.7
Women	3160	14.1	3642	16.2	1279	18.8
	At risk (N)	Person-years at risk‡	At risk (N)	Person-years at risk‡	At risk (N)	Person-years at risk‡
Incidence						
Men	831	1218	2328	3524	1231	1939
Women	2714	4203	3051	4887	1279	1637

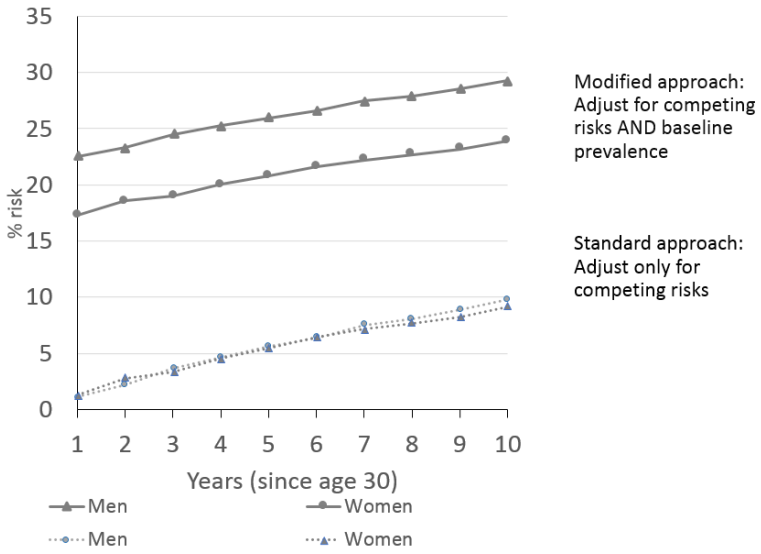
† prevalence of any cancer at cohort enrollment. Kaposi sarcoma and cervical cancer account for >98% of prevalent malignancies.
‡ person-years at risk among cancer-free subjects are administratively censored at Oct 30, 2010.

Appendix B Table 2. Cumulative incidence of ever cancer among ART users, Lighthouse Trust

	+10 years		+10 years	
	Estimated risk (%) Adjusting for competing risk and baseline prevalence of cancer	(95%CI)	Estimated risk (%) Adjusting for competing risk only	(95%CI)
Men, Current Age				
20	25.3	(21.4, 29.1)	8.8	(4.6, 13.0)
30	28.2	(26.4, 30.1)	9.2	(7.5, 10.9)
40	28.8	(25.8, 30.4)	7.7	(5.9, 9.4)
Women, Current Age				
20	21.4	(19.4, 23.5)	9.2	(7.5, 10.9)
30	23.1	(21.6, 24.6)	7.7	(5.9, 9.4)
40	23.6	(21.3, 6.0)	9.2	(7.5, 10.9)

Results: In Appendix B Appendix B Table 2, estimated risk is the age-conditional probability of having *or* developing cancer (i.e. ever). In the left column, for a 20-year old man who started ART at Lighthouse Trust during 2007-2010, the risk of already having *or* developing cancer by age 29 was 25.3%. In the right column, without considering baseline prevalence or setting the baseline hazard to 0, our estimate reduces to the cumulative incidence adjusting only for competing risk.

Appendix B Figure 1. Incidence of cancer among 30-year old ART users: modified Kaplan Meier versus standard approach



We plan to extend this work to the QECH cohort, and also parse cumulative incidence for ADC and NADC groups, while simultaneously accounting for competing risks and prevalence of cancer at time of clinic enrollment. Our preliminary results estimated the probability of having or developing cancer over a 10-year time horizon, given a specific age at ART start. For example, we show that for a 20-year old man, the risk of already having *or* developing cancer by age 29 was 25% and 21% for a 20-year old woman. By age 30, the ACI increases to 28% for men and 23% for women. Our estimates of ACI use the most current age-specific cancer incidence rates in a population of young ART users, are less susceptible to historic calendar trends in cancer incidence, and may be more reflective of risk experienced by those alive today. On a policy level, cumulative risk estimates are important to plan for the future burden of cancer while also accounting for the current burden of disease.

APPENDIX C. TEMPORAL TRENDS IN INCIDENCE OF KAPOSI SARCOMA

For KS and cervical cancer, incidence rate ratios (IRR) were calculated for calendar period of clinic enrollment categorized as 2000-2003 (limited ART), 2004-2006 (early national ART scale-up), 2007-2010 (later national ART scale-up) to describe secular trends in cancer incidence at QECH during ART scale-up. To examine temporal trends in cancer incidence, an update to the linkage with new cancer registry data collected in Blantyre from 2011 through 2015 will be conducted. This will enable a trend analysis of incidence rates comparing the 2007-2010 period to the 2011-2015 period at QECH in Blantyre. A preliminary trend analysis was conducted for the 2000-2010 period but results are subject to calendar artifacts in the early cancer registry design and therefore should be interpreted with reservation.

AIDS-defining cancers across calendar period

The trend of KS incidence among ART users at QECH was not linear across calendar period of enrollment: IR peaked in 2004-2006 at 661·1 (95%CI: 554·5, 788·0), with most cases diagnosed among stage 4 patients, followed by a 76% decline in incidence by 2007-2010 (IRR=0·24; 95%CI: 0·11, 0·51; Appendix C Table 1). At the same time, a shift in WHO stage was noted across calendar periods: in 2000-2003, two-thirds of patients initiated therapy at clinical stage 1 or 2; by 2004-2006, this shifted to 79% at clinical stages 3 or 4. By 2007-2010, approximately half of patients were stage 3 or 4 and this was distribution was similar to LH during the same period (Appendix C Table 2; Appendix C Figure 1).

At QECH, KS incidence among men during 2004-2006 (IR: 1083; 96%CI: 863, 1360) was comparable to that observed at LH during 2007-2010 (IRR: 1.11; Appendix C Table 1). Among women, the IR of KS during 2004-2006 was less than half that of men (IR: 418·4; 96%CI: 317·1, 552·0).

The calendar trend in KS incidence is likely an artifact of both evolving KS screening at QECH and a calendar effect of underascertainment by the registry. Historically, a large proportion of KS cases from QECH were diagnosed in the out-patient department (OPD) and referred to an external palliative care center (Tiyanjane), both of which were not primary sources of information during the 2000-2003 expansion of population-based cancer registration. Accordingly, the cancer registry recorded a surge of KS cases once the OPD was added as a datasource in 2003 (Dzamalala personal communication). Furthermore, in Blantyre, suspected KS patients are often referred to dermatology clinic (LEPRA) for diagnostic confirmation rather than pathology. Once confirmed, cases are in turn referred directly to

palliative care or again to oncology. Given the fragmented nature of medical record keeping in low-resource settings and under-resourced healthcare system, a substantial proportion of KS cases may be lost along this chain of referrals, and subsequently invisible to public health surveillance activities by the registry. Calendar trends in KS incidence recorded during the early years of the registry should be interpreted with reservation, as these are likely to be artifactual due to the evolving design of the registry.

Appendix C Table 1. Kaposi sarcoma and cervical cancer incidence rates by WHO clinical stage, early versus late timing of cancer diagnosis, and calendar period of ART initiation

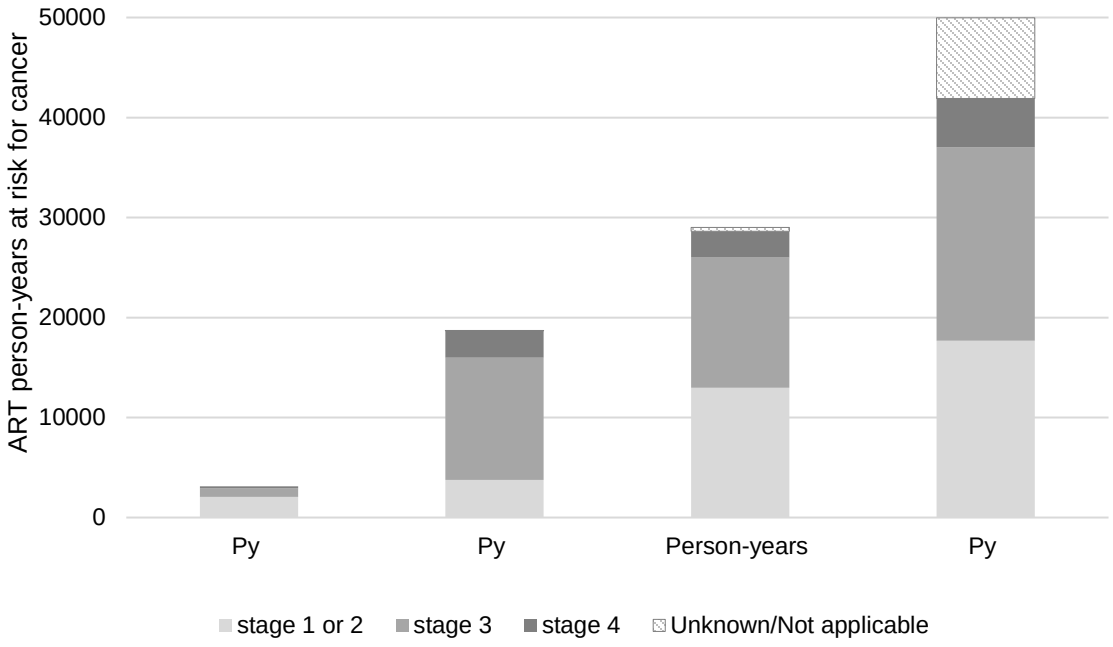
	QECH 2000-2003			QECH 2004-2006			QECH 2007-2010			LH 2007-2010		
	IR	(95% CI)	IRR (95%CI)	IR	(95% CI)	IRR (95%CI)	IR	(95% CI)	IRR (95%CI)	IR	(95% CI)	IRR (95%CI)
Both sexes												
All stages	**224.1**	**(106.8, 470.0)**	**0.34 (0.16, 0.72)**	**661.1**	**(554.5, 788.0)**	**1.**	**156.9**	**(117.1, 209.6)**	**0.24 (0.11, 0.51)**	**1204.2**	**(1111.7, 1304.3)**	**1.82**
stage 1 or 2	**-**	**-**		**16.4**	**(5.4, 50.1)**	**1.**	**62.8**	**(39.6, 99.5)**	**3.82**	**336.7**	**(289.5, 391.6)**	**20.48**
early incident	-	-		-	-		55.0	(33.6, 89.9)	7.05	233.6	(194.9, 280.1)	
late incident	-	-		16.5	(5.4, 50.4)		7.8	(2.1, 28.8)		103.2	(78.6, 135.6)	
stage 3	**20.5**	**(1.7, 243.8)**	**0.35**	**59.1**	**(32.7, 106.6)**	**1.**	**60.7**	**(38.0, 96.9)**	**1.03**	**177.7**	**(144.3, 218.8)**	**3.01**
early incident	20.4	(1.7, 242.7)		18.0	(6.2, 52.4)		46.3	(27.1, 79.2)	3.23	155.2	(124.2, 193.9)	
late incident	-	-		41.1	(20.3, 83.4)		14.4	(5.5, 37.6)		22.6	(12.6, 40.6)	
stage 4	**207.0**	**(95.2, 449.9)**	**0.35**	**589.8**	**(489.4, 710.8)**	**1.**	**33.5**	**(17.8, 62.9)**	**0.16**	**242.1**	**(202.6, 289.4)**	**0.41**
early incident	-	-		481.2	(391.4, 591.6)		29.5	(15.1, 57.8)	7.57	185.2	(151.1, 227.1)	
late incident	207.6	(95.5, 451.2)		107.8	(69.7, 166.8)		3.9	(0.6, 24.7)		57.0	(39.5, 82.4)	
Men												
All stages	**332.7**	**(135.6, 816.2)**	**0.31**	**1083.5**	**(863.3, 1360.0)**	**1.**	**232.1**	**(159.7, 337.4)**	**0.21**	**1207.6**	**(1064.3, 1370.2)**	**1.11**
stage 1 or 2	**-**	**-**		**14.7**	**(2.1, 103.4)**	**1.**	**65.8**	**(32.6, 132.8)**	**4.49**	**393.4**	**(315.3, 490.9)**	
early incident	-	-		14.7	-		46.8			281.2	(216.4, 365.4)	
late incident	-	-		-	-		19.0			112.4	(74.3, 170.1)	
stage 3	**43.7**	**(3.7, 519.8)**	**0.67**	**65.1**	**(25.7, 164.4)**	**1.**	**103.4**	**(59.1, 181.1)**	**1.59**	**236.2**	**(177.5, 314.3)**	
early incident		(0.0, 0.0)		14.7			-	-		212.1	(156.9, 286.8)	
late incident	-	-		50.5	(17.6, 144.7)		26.7	(8.9, 80.4)		24.2	(9.9, 59.0)	
stage 4	**289.0**	**(110.3, 756.9)**	**0.29**	**1003.8**	**(792.7, 1271.1)**	**1.**	**62.9**	**(30.7, 129.0)**	**0.06**	**568.5**	**(472.9, 683.4)**	**0.57**
early incident	-	-		-	-		53.4	(24.5, 116.5)		425.8	(344.2, 526.8)	
late incident	291.2	(111.2, 762.8)		-	-		9.5	(1.5, 60.3)		142.9	(99.0, 206.3)	
Women						0.3						
All stages	**132.0**	**(35.5, 490.1)**	**0.32**	**418.4**	**(317.1, 552.0)**	**1.**	**104.3**	**(65.6, 166.0)**	**0.25**	**1201.9**	**(1084.0, 1332.5)**	**2.87**
stage 1 or 2	**-**	**-**		**17.5**	**(4.5, 67.8)**	**1.**	**60.5**	**(32.9, 111.4)**	**3.47**	**299.0**	**(243.1, 367.7)**	**17.13**
early incident	-	-		17.6	-		60.6	(33.0, 111.6)		202.1	(157.1, 259.9)	
late incident	-	-		**-**	**-**		**-**	**-**		97.1	(67.6, 139.7)	
stage 3	**-**	**-**		**54.9**	**(25.6, 118.0)**	**1.**	**30.9**	**(13.2, 72.6)**	**0.56**	**138.9**	**(102.5, 188.1)**	**2.53**
early incident	-	-		19.8	(5.5, 71.1)		19.8	(5.5, 71.1)		117.3	(84.3, 163.3)	
late incident	-	-		35.6	(13.7, 92.3)		5.8	(0.8, 41.7)		21.6	(10.0, 46.7)	
stage 4	**132.0**	**(35.5, 502.9)**	**0.38**	**346.0**	**(255.1, 469.3)**	**1.**	**12.9**	**(3.5, 48.3)**	**0.04**	**25.4**	**(12.5, 51.7)**	**0.07**
early incident	-	-		297.3	(213.7, 413.5)		12.9	(3.5, 48.4)		25.4	(12.5, 51.7)	
late incident	132.0	(35.5, 502.9)		51.3	(23.2, 113.5)		-	-		-	-	

IR: incidence rates per 100,000 person-years on ART, direct standardized for sex-age categories using combined cohorts as standard population
95% CI : 95% Confidence Interval
IRR: incidence rate ratio
Calendar periods include 2000-2003 (limited availability fee-for-service ART, Blantyre only), 2004-2006 (early period national scale-up of ART access), 2007-2010 (late period national scale-up of ART access). IRR for calendar period (all stages): Incidence Rate Ratio, referent group is calendar period 2004-2006

Appendix C Table 2. Individuals at risk and ART person-years at risk for incident cancer, by WHO stage and calendar period

	QECH									LH		
	2000-2003			2004-2006			2007-2010			2007-2010		
WHO stage	**N**	**Py**	**% Py**	**N**	**Py**	**% Py**	**N**	**Py**	**% Py**	**N**	**Py**	**% Py**
			66%									
stage 1 or 2	267	2059		708	3768	20%	2933	12965	45%	4197	17693	35%
stage 3	128	878	28%	2424	12217	65%	3150	13026	45%	4919	19362	39%
stage 4	27	183	6%	654	2674	14%	759	2618	9%	1367	4854	10%
Unknown/Not applicable	0	0.0	0%	3	28	0%	138	412	1%	1981	8070	16%
Total	422	3121		3789	18689		6980	29022		12464	49980	

Appendix C Figure 1. Person-years at risk for cancer, by WHO clinical stage, calendar period, and cohort



REFERENCES

1. GLOBOCAN 2012 v1.0, Cancer Incidence and Mortality Worldwide: IARC CancerBase No. 11 [Internet]. Lyon, France: International Agency for Research on Cancer; 2013. [Accessed at http://globocan.iarc.fr on September 21, 2017].

2. American Cancer Society. Global Cancer Facts & Figures 2nd Edition. Cancer in Africa. Atlanta: American Cancer Society; 2011.

3. UNAIDS. Access to antiretroviral therapy in Africa. Status report on progress towards the 2015 targets. [Accessed at http://www.unaids.org/sites/default/files/media_asset/20131219_AccessARTAfricaStatusReportProgresstowards2015Targets_en_0.pdf on January 28, 2018].

4. Hammer SM, Squires KE, Hughes MD, Grimes JM, Demeter LM, Currier JS, et al. A controlled trial of two nucleoside analogues plus indinavir in persons with human immunodeficiency virus infection and CD4 cell counts of 200 per cubic millimeter or less. AIDS Clinical Trials Group 320 Study Team. The New England journal of medicine. 1997;337(11):725-33.

5. Patel P, Hanson DL, Sullivan PS, Novak RM, Moorman AC, Tong TC, et al. Incidence of types of cancer among HIV-infected persons compared with the general population in the United States, 1992-2003. Annals of internal medicine. 2008;148(10):728-36.

6. Shiels MS, Cole SR, Wegner S, Armenian H, Chmiel JS, Ganesan A, et al. Effect of HAART on incident cancer and noncancer AIDS events among male HIV seroconverters. Journal of acquired immune deficiency syndromes (1999). 2008;48(4):485-90.

7. Franceschi S, Lise M, Clifford GM, Rickenbach M, Levi F, Maspoli M, et al. Changing patterns of cancer incidence in the early- and late-HAART periods: the Swiss HIV Cohort Study. British journal of cancer. 2010;103(3):416-22.

8. Pipkin S, Scheer S, Okeigwe I, Schwarcz S, Harris DH, Hessol NA. The effect of HAART and calendar period on Kaposi's sarcoma and non-Hodgkin lymphoma: results of a match between an AIDS and cancer registry. Aids. 2011;25(4):463-71.

9. Shiels MS, Pfeiffer RM, Gail MH, Hall HI, Li J, Chaturvedi AK, et al. Cancer burden in the HIV-infected population in the United States. Journal of the National Cancer Institute. 2011;103(9):753-62.

10. Robbins HA, Shiels MS, Pfeiffer RM, Engels EA. Epidemiologic contributions to recent cancer trends among HIV-infected people in the United States. Aids. 2014;28(6):881-90.

11. Dryden-Peterson S, Medhin H, Kebabonye-Pusoentsi M, Seage GR, 3rd, Suneja G, Kayembe MK, et al. Cancer Incidence following Expansion of HIV Treatment in Botswana. PloS one. 2015;10(8):e0135602.

12. Mbulaiteye SM, Katabira ET, Wabinga H, Parkin DM, Virgo P, Ochai R, et al. Spectrum of cancers among HIV-infected persons in Africa: the Uganda AIDS-Cancer Registry Match Study. International journal of cancer Journal international du cancer. 2006;118(4):985-90.

13. Mutyaba I, Phipps W, Krantz EM, Goldman JD, Nambooze S, Orem J, et al. A Population-Level Evaluation of the Effect of Antiretroviral Therapy on Cancer Incidence in Kyadondo County, Uganda, 1999 - 2008. Journal of acquired immune deficiency syndromes (1999). 2015.

14. Government of Malawi. Malawi AIDS Response Progress Report. April 2015. [Accessed at http://www.unaids.org/sites/default/files/country/documents/MWI_narrative_report_2015.pdf on October 31, 2015].

15. National Statistical Office (NSO) and ICF Macro. 2011. Malawi Demographic and Health Survey 2010. Zomba, Malawi, and Calverton, Maryland, USA: NSO and ICF Macro.

16. 2014 Clinical Management of HIV in Children and Adults. Malawi Integrated Guidelines for Providing HIV Services in Antenatal Care, Maternity Care, Under 5 Clinics, Family Planning Clinics, HIV Exposed Child/pre-ART Clinics, ART Clinics. Second Edition. Ministry of Health, Malawi. Available at www.hivunitmohmw.org, Department for HIV and AIDS of the Ministry of Health.

17. Msyamboza KP, Dzamalala C, Mdokwe C, Kamiza S, Lemerani M, Dzowela T, et al. Burden of cancer in Malawi; common types, incidence and trends: national population-based cancer registry. BMC research notes. 2012;5:149.

18. Goldstein H, Harron K, Wade A. The analysis of record-linked data using multiple imputation with data value priors. Statistics in medicine. 2012;31(28):3481-93.

19. Moore CL, Amin J, Gidding HF, Law MG. A new method for assessing how sensitivity and specificity of linkage studies affects estimation. PloS one. 2014;9(7):e103690.

20. Forman D, Bray F, Brewster DH, Gombe Mbalawa C, Kohler B, Piñeros M, et al., editors. Cancer Incidence in Five Continents, Vol. X (electronic version). Lyon: International Agency for Research on Cancer. 2013 Available at: http://ci5.iarc.fr [Accessed on September 7, 2017].

21. Engels EA. Human immunodeficiency virus infection, aging, and cancer. Journal of clinical epidemiology. 2001;54 Suppl 1:S29-34.

22. Engels EA, Biggar RJ, Hall HI, Cross H, Crutchfield A, Finch JL, et al. Cancer risk in people infected with human immunodeficiency virus in the United States. International journal of cancer Journal international du cancer. 2008;123(1):187-94.

23. Engels EA, Pfeiffer RM, Goedert JJ, Virgo P, McNeel TS, Scoppa SM, et al. Trends in cancer risk among people with AIDS in the United States 1980-2002. Aids. 2006;20(12):1645-54.

24. Silverberg MJ, Abrams DI. AIDS-defining and non-AIDS-defining malignancies: cancer occurrence in the antiretroviral therapy era. Current opinion in oncology. 2007;19(5):446-51.

25. Silverberg MJ, Chao C, Abrams DI. New insights into the role of HIV infection on cancer risk. The Lancet Oncology. 2009;10(12):1133-4.

26. Centers for Disease Control (CDC). Kaposi's sarcoma and pneumocystis pneumonia among homosexual men: New York City and California. MMWR Morb Mortal Wkly Rep. 1981;30:305-8.

27. Shiels MS, Engels EA. Increased risk of histologically defined cancer subtypes in human immunodeficiency virus-infected individuals: clues for possible immunosuppression-related or infectious etiology. Cancer. 2012;118(19):4869-76.

28. Grulich AE, van Leeuwen MT, Falster MO, Vajdic CM. Incidence of cancers in people with HIV/AIDS compared with immunosuppressed transplant recipients: a meta-analysis. Lancet. 2007;370(9581):59-67.

29. Appay V, Sauce D. Immune activation and inflammation in HIV-1 infection: causes and consequences. J Pathol. 2008;214(2):231-41.

30. Dubrow R, Silverberg MJ, Park LS, Crothers K, Justice AC. HIV infection, aging, and immune function: implications for cancer risk and prevention. Current opinion in oncology. 2012;24(5):506-16.

31. Desai S, Landay A. Early immune senescence in HIV disease. Curr HIV/AIDS Rep. 2010;7(1):4-10.

32. Deeks SG. HIV infection, inflammation, immunosenescence, and aging. Annual review of medicine. 2011;62:141-55.

33. Deeks SG. Immune dysfunction, inflammation, and accelerated aging in patients on antiretroviral therapy. Top HIV Med. 2009;17(4):118-23.

34. Plaeger SF, Collins BS, Musib R, Deeks SG, Read S, Embry A. Immune activation in the pathogenesis of treated chronic HIV disease: a workshop summary. AIDS research and human retroviruses. 2012;28(5):469-77.

35. Franceschi S, Maso LD, Rickenbach M, Polesel J, Hirschel B, Cavassini M, et al. Kaposi sarcoma incidence in the Swiss HIV Cohort Study before and after highly active antiretroviral therapy. British journal of cancer. 2008;99(5):800-4.

36. Polesel J, Clifford GM, Rickenbach M, Dal Maso L, Battegay M, Bouchardy C, et al. Non-Hodgkin lymphoma incidence in the Swiss HIV Cohort Study before and after highly active antiretroviral therapy. Aids. 2008;22(2):301-6.

37. Franceschi S, Polesel J, Rickenbach M, Dal Maso L, Probst-Hensch NM, Fux C, et al. Hepatitis C virus and non-Hodgkin's lymphoma: Findings from the Swiss HIV Cohort Study. British journal of cancer. 2006;95(11):1598-602.

38. Krishnan S, Schouten JT, Jacobson DL, Benson CA, Collier AC, Koletar SL, et al. Incidence of non-AIDS-defining cancer in antiretroviral treatment-naive subjects after antiretroviral treatment initiation: an ACTG longitudinal linked randomized trials analysis. Oncology. 2011;80(1-2):42-9.

39. Prosperi MC, Cozzi-Lepri A, Castagna A, Mussini C, Murri R, Giacometti A, et al. Incidence of malignancies in HIV-infected patients and prognostic role of current CD4 cell count: evidence from a large Italian cohort study. Clinical infectious diseases : an official publication of the Infectious Diseases Society of America. 2010;50(9):1316-21.

40. Bruyand M, Thiebaut R, Lawson-Ayayi S, Joly P, Sasco AJ, Mercie P, et al. Role of uncontrolled HIV RNA level and immunodeficiency in the occurrence of malignancy in HIV-infected patients during the combination antiretroviral therapy era: Agence Nationale de Recherche sur le Sida (ANRS) CO3 Aquitaine Cohort. Clinical infectious diseases : an official publication of the Infectious Diseases Society of America. 2009;49(7):1109-16.

41. Reekie J, Kosa C, Engsig F, Monforte A, Wiercinska-Drapalo A, Domingo P, et al. Relationship between current level of immunodeficiency and non-acquired immunodeficiency syndrome-defining malignancies. Cancer. 2010;116(22):5306-15.

42. Monforte A, Abrams D, Pradier C, Weber R, Reiss P, Bonnet F, et al. HIV-induced immunodeficiency and mortality from AIDS-defining and non-AIDS-defining malignancies. Aids. 2008;22(16):2143-53.

43. Kesselring A, Gras L, Smit C, van Twillert G, Verbon A, de Wolf F, et al. Immunodeficiency as a risk factor for non-AIDS-defining malignancies in HIV-1-infected patients receiving combination antiretroviral therapy. Clinical infectious diseases : an official publication of the Infectious Diseases Society of America. 2011;52(12):1458-65.

44. De Vuyst H, Lillo F, Broutet N, Smith JS. HIV, human papillomavirus, and cervical neoplasia and cancer in the era of highly active antiretroviral therapy. Eur J Cancer Prev. 2008;17(6):545-54.

45. Silverberg MJ, Lau B, Justice AC, Engels E, Gill MJ, Goedert JJ, et al. Risk of anal cancer in HIV-infected and HIV-uninfected individuals in North America. Clinical infectious diseases : an official publication of the Infectious Diseases Society of America. 2012;54(7):1026-34.

46. Chaturvedi AK, Madeleine MM, Biggar RJ, Engels EA. Risk of human papillomavirus-associated cancers among persons with AIDS. Journal of the National Cancer Institute. 2009;101(16):1120-30.

47. Guiguet M, Boue F, Cadranel J, Lang JM, Rosenthal E, Costagliola D, et al. Effect of immunodeficiency, HIV viral load, and antiretroviral therapy on the risk of individual malignancies (FHDH-ANRS CO4): a prospective cohort study. The Lancet Oncology. 2009;10(12):1152-9.

48. Crum-Cianflone NF, Hullsiek KH, Marconi VC, Ganesan A, Weintrob A, Barthel RV, et al. Anal cancers among HIV-infected persons: HAART is not slowing rising incidence. Aids. 2010;24(4):535-43.

49. Hanahan D, Weinberg RA. The hallmarks of cancer. Cell. 2000;100:57-70.

50. Hanahan D, Weinberg RA. Hallmarks of cancer: the next generation. Cell. 2011;144(646-674).

51. de Visser KE, Eichten A, Coussens LM. Paradoxical roles of the immune system during cancer development. Nature Reviews Cancer. 2006;6:24-37.

52. Dvorak HF. Tumors: wounds that do not heal. Similarities between tumor stroma generation and wound healing. The New England journal of medicine. 1986;315(26):1650-9.

53. Burnet FM. Cancer—a biological approach. Br Med J.1:841-7.

54. Thomas L. Cellular and Humoral Aspects of the Hypersensitive States. . New York: Hoeber-Harper; 1959.

55. Ehrlich P. Ueber den jetzigen stand der Karzinomforschung. Ned Tijdschr Geneeskd. 1909;5:273–90.

56. Chow MT, Moller A, Smyth MJ. Inflammation and immune surveillance in cancer. Seminars in cancer biology. 2011.

57. Dunn GP, Old LJ, Schreiber RD. The immunobiology of cancer immunosurveillance and immunoediting. Immunity. 2004;21:137–48.

58. Schreiber RD, Old LJ, Smyth MJ. Cancer immunoediting: integrating immunity's roles in cancer suppression and promotion. Science. 2011;331:1565-70.

59. Finn OJ. Immuno-oncology: understanding the function and dysfunction of the immune system in cancer. Annals of Oncology 2012;23(Supplement 8):viii6–viii9.

60. Boon T, van der Bruggen P. Human tumor antigens recognized by T lymphocytes. J Exp Med. 1996;183(3):725-9.

61. Rosenberg SA. A new era for cancer immunotherapy based on the genes that encode cancer antigens. Immunity. 1999;10(3):281-7.

62. Frisch M, Biggar RJ, Engels EA, Goedert JJ, Group AI-CMRS. Association of cancer with AIDS-related immunosuppression in adults. Jama. 2001;285(13):1736-45.

63. Angeletti PC, Zhang L, Wood C. The viral etiology of AIDS-associated malignancies. Advances in pharmacology. 2008;56:509-57.

64. Shiels MS, Cole SR, Kirk GD, Poole C. A meta-analysis of the incidence of non-AIDS cancers in HIV-infected individuals. Journal of acquired immune deficiency syndromes (1999). 2009;52(5):611-22.

65. Silverberg MJ, Chao C, Leyden WA, Xu L, Tang B, Horberg MA, et al. HIV infection and the risk of cancers with and without a known infectious cause. Aids. 2009;23(17):2337-45.

66. Shiels MS, Goedert JJ, Moore RD, Platz EA, Engels EA. Reduced risk of prostate cancer in U.S. Men with AIDS. Cancer epidemiology, biomarkers & prevention : a publication of the American Association for Cancer Research, cosponsored by the American Society of Preventive Oncology. 2010;19(11):2910-5.

67. Tesoriero JM, Gieryic SM, Carrascal A, Lavigne HE. Smoking among HIV positive New Yorkers: prevalence, frequency, and opportunities for cessation. AIDS Behav. 2010;14(4):824-35.

68. Clifford GM, Polesel J, Rickenbach M, Dal Maso L, Keiser O, Kofler A, et al. Cancer risk in the Swiss HIV Cohort Study: associations with immunodeficiency, smoking, and highly active antiretroviral therapy. Journal of the National Cancer Institute. 2005;97(6):425-32.

69. Hessol NA, Seaberg EC, Preston-Martin S, Massad LS, Sacks HS, Silver S, et al. Cancer risk among participants in the women's interagency HIV study. Journal of acquired immune deficiency syndromes (1999). 2004;36(4):978-85.

70. Kirk GD, Merlo C, P OD, Mehta SH, Galai N, Vlahov D, et al. HIV infection is associated with an increased risk for lung cancer, independent of smoking. Clinical infectious diseases : an official publication of the Infectious Diseases Society of America. 2007;45(1):103-10.

71. Shiels MS, Cole SR, Mehta SH, Kirk GD. Lung cancer incidence and mortality among HIV-infected and HIV-uninfected injection drug users. Journal of acquired immune deficiency syndromes (1999). 2010;55(4):510-5.

72. Engels EA, Brock MV, Chen J, Hooker CM, Gillison M, Moore RD. Elevated incidence of lung cancer among HIV-infected individuals. Journal of clinical oncology : official journal of the American Society of Clinical Oncology. 2006;24(9):1383-8.

73. Chaturvedi AK, Pfeiffer RM, Chang L, Goedert JJ, Biggar RJ, Engels EA. Elevated risk of lung cancer among people with AIDS. Aids. 2007;21(2):207-13.

74. Engels EA. Inflammation in the development of lung cancer: epidemiological evidence. Expert review of anticancer therapy. 2008;8(4):605-15.

75. Silverberg MJ, Abrams DI. Do antiretrovirals reduce the risk of non-AIDS-defining malignancies? Current opinion in HIV and AIDS. 2009;4(1):42-51.

76. HIV-Causal Collaboration, Ray M, Logan R, Sterne JA, Hernandez-Diaz S, Robins JM, et al. The effect of combined antiretroviral therapy on the overall mortality of HIV-infected individuals. Aids. 2010;24(1):123-37.

77. Biggar RJ, Engels EA, Ly S, Kahn A, Schymura MJ, Sackoff J, et al. Survival after cancer diagnosis in persons with AIDS. Journal of acquired immune deficiency syndromes (1999). 2005;39(3):293-9.

78. Palella FJ, Jr., Baker RK, Moorman AC, Chmiel JS, Wood KC, Brooks JT, et al. Mortality in the highly active antiretroviral therapy era: changing causes of death and disease in the HIV outpatient study. Journal of acquired immune deficiency syndromes (1999). 2006;43(1):27-34.

79. Palella FJ, Jr., Delaney KM, Moorman AC, Loveless MO, Fuhrer J, Satten GA, et al. Declining morbidity and mortality among patients with advanced human immunodeficiency virus infection. HIV Outpatient Study Investigators. The New England journal of medicine. 1998;338(13):853-60.

80. Goedert JJ, Cote TR, Virgo P, Scoppa SM, Kingma DW, Gail MH, et al. Spectrum of AIDS-associated malignant disorders. Lancet. 1998;351(9119):1833-9.

81. Edwards BK, Ward E, Kohler BA, Eheman C, Zauber AG, Anderson RN, et al. Annual report to the nation on the status of cancer, 1975-2006, featuring colorectal cancer trends and impact of interventions (risk factors, screening, and treatment) to reduce future rates. Cancer. 2010;116(3):544-73.

82. Altekruse SF, Kosary CL, Krapcho M, Neyman N, Aminou R, Waldron W, Ruhl J, Howlader N, Tatalovich Z, Cho H, Mariotto A, Eisner MP, Lewis DR, Cronin K, Chen HS, Feuer EJ, Stinchcomb DG, Edwards BK (eds). SEER Cancer Statistics Review, 1975-2007, National Cancer Institute. Bethesda, MD, http://seer.cancer.gov/csr/1975_2007/, based on November 2009 SEER data submission, posted to the SEER web site, 2010.

83. Shiels MS, Copeland G, Goodman MT, Harrell J, Lynch CF, Pawlish K, et al. Cancer stage at diagnosis in patients infected with the human immunodeficiency virus and transplant recipients. Cancer. 2015.

84. Coghill AE, Newcomb PA, Madeleine MM, Richardson BA, Mutyaba I, Okuku F, et al. Contribution of HIV infection to mortality among cancer patients in Uganda. Aids. 2013;27(18):2933-42.

85. Coghill AE, Engels EA. Are cancer outcomes worse in the presence of HIV infection? Cancer epidemiology, biomarkers & prevention : a publication of the American Association for Cancer Research, cosponsored by the American Society of Preventive Oncology. 2015;24(8):1165-6.

86. Coghill AE, Shiels MS, Suneja G, Engels EA. Elevated Cancer-Specific Mortality Among HIV-Infected Patients in the United States. Journal of clinical oncology : official journal of the American Society of Clinical Oncology. 2015;33(21):2376-83.

87. Simard EP, Engels EA. Cancer as a cause of death among people with AIDS in the United States. Clinical infectious diseases : an official publication of the Infectious Diseases Society of America. 2010;51(8):957-62.

88. Bonnet F, Lewden C, May T, Heripret L, Jougla E, Bevilacqua S, et al. Malignancy-related causes of death in human immunodeficiency virus-infected patients in the era of highly active antiretroviral therapy. Cancer. 2004;101(2):317-24.

89. UNAIDS. Global AIDS update 2016. [Accessed at http://www.unaids.org/sites/default/files/media_asset/global-AIDS-update-2016_en.pdf on January 28, 2018].

90. United Nations, Department of Economic and Social Affairs, Population Division (2015). World Population Prospects: The 2015 Revision, Volume II: Demographic Profiles (ST/ESA/SER.A/380). .

91. World Health Organization. Guideline on when to start antiretroviral therapy and on preexposure prophylaxis for HIV. Septmember 2015 early-release guideline. ISBN 978 92 4 150956 5.

92. Joint United Nations Programme on HIV/AIDS (UNAIDS). The Gap Report. UNAIDS / JC2656 (English original, July 2014, updated September 2014). ISBN 978-92-9253-062-4.

93. Cohen MS, Chen YQ, McCauley M, Gamble T, Hosseinipour MC, Kumarasamy N, et al. Prevention of HIV-1 infection with early antiretroviral therapy. The New England journal of medicine. 2011;365(6):493-505.

94. http://www.unicef.org/esaro/5482_HIV_AIDS.html.

95. United Nations Development Program (UNDP). 2014 Human Development Report. Sustaining Human Progress: Reducing Vulnerabilities and Building Resilience. [Accessed at http://www.undp.org/content/undp/en/home/librarypage/hdr/2014-human-development-report.html on Febrary 11, 2016].

96. The World Bank. Malawi Country Profile [Accessed at http://data.worldbank.org/country/malawi. on February 11, 2016].

97. Treatment Of Aids Guidelines For The Use Of Antiretroviral Therapy In Malawi. First Edition: September 2003. National AIDS Commission of Malawi and Ministry of Health and Population, Malawi.

98. Lowrance DW, Makombe S, Harries AD, Shiraishi RW, Hochgesang M, Aberle-Grasse J, et al. A public health approach to rapid scale-up of antiretroviral treatment in Malawi during 2004-2006. Journal of acquired immune deficiency syndromes (1999). 2008;49(3):287-93.

99. World Health Organization. The global burden of disease: 2004 update.

100. World Health Organization. World Cancer Report 2008. Lyon: International Agency for Research on Cancer; 2008.

101. United Nations, Department of Economic and Social Affairs, Population Division (2011). World Population Prospects: The 2010 Revision, Volume I: Comprehensive Tables. ST/ESA/SER.A/313.

102. Aging population challenges in Africa. African Development Bank, Chief Economist Complex. Vol 1, Issue 1. November 2011.

103. World Health Organization. National Cancer Control Programmes: policies and managerial guidelines. 2nd ed. Geneva: World Health Organization; 2002.

104. Msyamboza KP, Ngwira B, Dzowela T, Mvula C, Kathyola D, Harries AD, et al. The burden of selected chronic non-communicable diseases and their risk factors in Malawi: nationwide STEPS survey. PloS one. 2011;6(5):e20316.

105. Parkin DM. The role of cancer registries in cancer control. Int J Clin Oncol. 2008;13(2):102-11.

106. Valsecchi MG, Steliarova-Foucher E. Cancer registration in developing countries: luxury or necessity? The Lancet Oncology. 2008;9(2):159-67.

107. Parkin DM. The evolution of the population-based cancer registry. Nature reviews Cancer. 2006;6(8):603-12.

108. Adesina A, Chumba D, Nelson AM, Orem J, Roberts DJ, Wabinga H, et al. Improvement of pathology in sub-Saharan Africa. The Lancet Oncology. 2013;14(4):e152-7.

109. Dryden-Peterson S, Medhin H, Kebabonye-Pusoentsi M, Seage GR, 3rd, Suneja G, Kayembe MK, et al. Correction: Cancer Incidence following Expansion of HIV Treatment in Botswana. PloS one. 2015;10(9):e0138742.

110. Lahuerta M, Ue F, Hoffman S, Elul B, Kulkarni SG, Wu Y, et al. The problem of late ART initiation in Sub-Saharan Africa: a transient aspect of scale-up or a long-term phenomenon? J Health Care Poor Underserved. 2013;24(1):359-83.

111. Wabinga HR, Nambooze S, Amulen PM, Okello C, Mbus L, Parkin DM. Trends in the incidence of cancer in Kampala, Uganda 1991-2010. International journal of cancer Journal international du cancer. 2014;135(2):432-9.

112. Meireles P, Albuquerque G, Vieira M, Foia S, Ferro J, Carrilho C, et al. Kaposi sarcoma incidence in Mozambique: national and regional estimates. Eur J Cancer Prev. 2015;24(6):529-34.

113. Cook-Mozaffari P, Newton R, Beral V, Burkitt DP. The geographical distribution of Kaposi's sarcoma and of lymphomas in Africa before the AIDS epidemic. British journal of cancer. 1998;78(11):1521-8.

114. Paramsothy P, Jamieson DJ, Heilig CM, Schuman PC, Klein RS, Shah KV, et al. The effect of highly active antiretroviral therapy on human papillomavirus clearance and cervical cytology. Obstet Gynecol. 2009;113(1):26-31.

115. Sengayi M, Spoerri A, Kielkowski, D., et al. The use of computerized record linkage for cancer ascertainment in a South African HIV cohort. (Presented at AORTIC International Cancer Conference, November 22, 2013, Duban, South Africa.).

116. Sengayi M, Spoerri A, Egger M, Kielkowski D, Crankshaw T, Cloete C, et al. Record linkage to correct under-ascertainment of cancers in HIV cohorts: The Sinikithemba HIV clinic linkage project. International journal of cancer Journal international du cancer. 2016;139(6):1209-16.

117. Akarolo-Anthony SN, Maso LD, Igbinoba F, Mbulaiteye SM, Adebamowo CA. Cancer burden among HIV-positive persons in Nigeria: preliminary findings from the Nigerian AIDS-cancer match study. Infectious agents and cancer. 2014;9(1):1.

118. Ateenyi-Agaba C. Conjunctival squamous-cell carcinoma associated with HIV infection in Kampala, Uganda. Lancet. 1995;345(8951):695-6.

119. Mbulaiteye SM, Parkin DM, Rabkin CS. Epidemiology of AIDS-related malignancies an international perspective. Hematology/oncology clinics of North America. 2003;17(3):673-96, v.

120. Guech-Ongey M, Engels EA, Goedert JJ, Biggar RJ, Mbulaiteye SM. Elevated risk for squamous cell carcinoma of the conjunctiva among adults with AIDS in the United States. International journal of cancer Journal international du cancer. 2008;122(11):2590-3.

121. Parkin DM, Nambooze S, Wabwire-Mangen F, Wabinga HR. Changing cancer incidence in Kampala, Uganda, 1991-2006. International journal of cancer Journal international du cancer. 2010;126(5):1187-95.

122. Kim HJ, Fay MP, Feuer EJ, Midthune DN. Permutation tests for joinpoint regression with applications to cancer rates. Statistics in medicine. 2000;19(3):335-51.

123. Jutte DP, Roos L, Brownell MD. Administrative record linkage as a tool for public health research. Annu Rev Public Health. 2011;31:91–108.

124. Meray N, Reitsma JB, Ravelli AC, Bonsel GJ. Probabilistic record linkage is a valid and transparent tool to combine databases without a patient identification number. Journal of clinical epidemiology. 2007;60(9):883-91.

125. Harron K, Gilbert R. Research: increasing value, reducing waste. Lancet. 2014;383(9923):1124.

126. Tweya H, Gareta D, Chagwera F, Ben-Smith A, Mwenyemasi J, Chiputula F, et al. Early active follow-up of patients on antiretroviral therapy (ART) who are lost to follow-up: the 'Back-to-Care' project in Lilongwe, Malawi. Tropical medicine & international health : TM & IH. 2010;15 Suppl 1:82-9.

127. Sloan DJ, van Oosterhout JJ, Malisita K, Phiri EM, Lalloo DG, O'Hare B, et al. Evidence of improving antiretroviral therapy treatment delays: an analysis of eight years of programmatic outcomes in Blantyre, Malawi. BMC public health. 2013;13:490.

128. Malamba SS, Morgan D, Clayton T, Mayanja B, Okongo M, Whitworth J. The prognostic value of the World Health Organisation staging system for HIV infection and disease in rural Uganda. Aids. 1999;13(18):2555-62.

129. World Health Organization. Interim WHO clinical staging of HIV/AIDS and HIV/AIDS case definitions for surveillance: African region. Switzerland: World Health Organization; 2005.

130. Kassa E, Rinke de Wit TF, Hailu E, Girma M, Messele T, Mariam HG, et al. Evaluation of the World Health Organization staging system for HIV infection and disease in Ethiopia: association between clinical stages and laboratory markers. Aids. 1999;13(3):381-9.

131. Kagaayi J, Makumbi F, Nakigozi G, Wawer MJ, Gray RH, Serwadda D, et al. WHO HIV clinical staging or CD4 cell counts for antiretroviral therapy eligibility assessment? An evaluation in rural Rakai district, Uganda. Aids. 2007;21(9):1208-10.

132. Banda LT, Parkin DM, Dzamalala CP, Liomba NG. Cancer incidence in Blantyre, Malawi 1994-1998. Tropical medicine & international health : TM & IH. 2001;6(4):296-304.

133. Percy, C.L, Van Holten, V. & Muir, C.S., eds (1990). International Classification of Diseases for Oncology, 2nd edition (ICD-O-2). Geneva, World Health Organization.

134. Cooke, A.P., Parkin, D.M., Ferlay, J. CanReg4: Computer Software for Cancer Registries. Available at http://www.iacr.com.fr.

135. Sayers A, Ben-Shlomo Y, Blom AW, Steele F. Probabilistic record linkage. International journal of epidemiology. 2015.

136. Clark D. Practical introduction to record linkage for injury research. Injury Prev. 2004;10(3):186-91.

137. Weiss NS. Chapter 32. Clinical Epidemiology. In: Rothman K, Greenland S, Lash T, editors. Modern Epidemiology. 3 ed. Philadelphia: Lipponcott Williams & Wilkans; 2008. p. 642-6.

138. Greenland S, Rothman K. Chapter 13. Fundamentals of Epidemiologic Data Analysis In: Rothman K, Greenland S, Lash T, editors. Modern Epidemiology. 3 ed. Philadelphia: Lipponcott Williams & Wilkans; 2008. p. 230-1.

139. Bohensky MA, Jolley D, Sundararajan V, Evans S, Pilcher DV. Data Linkage: A powerful research tool with potential problems. BMC Health Serv Res. 2010;10:346-52.

140. Harron K, Wade A, Muller-Pebody B, Goldstein H, Gilbert R. Opening the black box of record linkage. J Epidemiol Community Health. 2012;66(12):1198.

141. Neter J, Maynes E, Ramanathan R. The effect of mismatching on the measurement of response error. J Am Stat Assoc. 1965;60(312):1005-7.

142. Schmidlin K, Clough-Gorr KM, Spoerri A, Egger M, Zwahlen M, Swiss National C. Impact of unlinked deaths and coding changes on mortality trends in the Swiss National Cohort. BMC medical informatics and decision making. 2013;13:1.

143. Jaro M. Probabilistic linkage of large public health data files. Statistics in medicine. 1995;14(5-7):491-8.

144. Silveira DP, Artmann E. Accuracy of probabilistic record linkage applied to health databases: systematic review. Rev Saude Publica. 2009;43(5):875-82.

145. McGlincy M. A Bayesian Record Linkage Methodology for Multiple Imputation of Missing Links. ASA Section on Survey Research Methods. 2004. [Accessed at http://www.amstat.org/sections/srms/Proceedings/y2004/files/Jsm2004-000683.pdf on August 21 2015].

146. Harron K, Wade A, Gilbert R, Muller-Pebody B, Goldstein H. Evaluating bias due to data linkage error in electronic healthcare records. BMC Med Res Methodol. 2014;14:36.

147. R. M, Berthold MR, Cebron N, Dill F, Gabriel TR, Kotter T, et al. KNIME: The Konstanz Information Miner. In: Studies in Classification, Data Analysis, and Knowledge Organization (GfKL 2007): Springer; 2007.

148. Wyber R, Vaillancourt S, Perry W, Mannava P, Folaranmi T, Celi L. Big data in global health: improving health in low- and middle-income countries. Bulletin of the World Health Organization.

149. The African Data Revolution Report 2016. Highlighting development in African data ecosystems.

150. The Africa Data Consensus. Addis Ababa, 30 and 31 March 2015. [Accessed at https://wwwunecaorg/sites/default/files/PageAttachments/final_adc_-_englishpdf on January 1, 2018].

151. Marquez PV, Farrington JL. No more disease silos for sub-Saharan Africa. BMJ. 2012;345:e5812.

152. Stewart J, Barker A. Undefined By Data: A Survey of Big Data Definitions. Ithaca: Cornell University Library arXiv preprint arXiv:13095821. 2013.

153. World Bank. 2017. Identification for Development ID4D. The State of Identification Systems in Africa, Washington, DC. [Accessed at http://www.worldbank.org/en/programs/id4d#5 on 1 March, 2018].

154. Statistical Capacity Indicator's Dashboard. 2015. World Bank. [Accessed at http://datatopicsworldbankorg/statisticalcapacity/ on January 1, 2018].

155. Devarajan S. Africa's Statistical Tragedy. Review of Income and Wealth. 2013;59(S1):S9-S15

156. Public Health Research Forum. Wellcome Trust. Enabling data linkage to maximize the value of public health research data: full report. March 2015.

157. Chasimpha SJD, Parkin DM, Masamba L, Dzamalala CP. Three-year cancer incidence in Blantyre, Malawi (2008-2010). International journal of cancer Journal international du cancer. 2017;141(4):694-700.

158. Paixao ES, Harron K, Andrade K, Teixeira MG, Fiaccone RL, Costa M, et al. Evaluation of record linkage of two large administrative databases in a middle income country: stillbirths and notifications of dengue during pregnancy in Brazil. BMC Med Inform Decis Mak. 2017;17(1):108.

159. DuVall SL, Kerber RA, Thomas A. Extending the Fellegi-Sunter probabilistic record linkage method for approximate field comparators. J Biomed Inform. 2010;43(1):24-30.

160. Bohensky MA, Jolley D, Sundararajan V, Evans S, Pilcher DV, Scott I, et al. Data linkage: a powerful research tool with potential problems. BMC Health Serv Res. 2010;10:346.

161. Kahn K, Tollman SM, Collinson MA, Clark SJ, Twine R, Clark BD, et al. Research into health, population and social transitions in rural South Africa: data and methods of the Agincourt Health and Demographic Surveillance System. Scand J Public Health Suppl. 2007;69:8-20.

162. Kabudula CW, Clark BD, Gomez-Olive FX, Tollman S, Menken J, Reniers G. The promise of record linkage for assessing the uptake of health services in resource constrained settings: a pilot study from South Africa. BMC Med Res Methodol. 2014;14:71.

163. Odei-Lartey EO, Boateng D, Danso S, Kwarteng A, Abokyi L, Amenga-Etego S, et al. The application of a biometric identification technique for linking community and hospital data in rural Ghana. Glob Health Action. 2016;9:29854.

164. Bohensky MA, Jolley D, Sundararajan V, Evans S, Ibrahim J, Brand C. Development and validation of reporting guidelines for studies involving data linkage. Australian and New Zealand journal of public health. 2011;35(5):486-9.

165. Hernandez-Ramirez RU, Shiels MS, Dubrow R, Engels EA. Cancer risk in HIV-infected people in the USA from 1996 to 2012: a population-based, registry-linkage study. Lancet HIV. 2017.

166. Islam JY, Rosenberg PS, Hall HI, Jacobson EU, Engels EA, Shiels MS. Projections of cancer incidence and burden among the HIV-positive population in the United States through 2030. Cancer Res 2017;77(13 Supplement):5302.

167. Tweya H, Feldacker C, Heller T, Gugsa S, Ng'ambi W, Nthala O, et al. Characteristics and outcomes of older HIV-infected patients receiving antiretroviral therapy in Malawi: A retrospective observation cohort study. PloS one. 2017;12(7):e0180232.

168. de Martel C, Ferlay J, Franceschi S, Vignat J, Bray F, Forman D, et al. Global burden of cancers attributable to infections in 2008: a review and synthetic analysis. The Lancet Oncology. 2012;13(6):607-15.

169. Gopal S, Krysiak R, Liomba NG, Horner MJ, Shores CG, Alide N, et al. Early experience after developing a pathology laboratory in Malawi, with emphasis on cancer diagnoses. PloS one. 2013;8(8):e70361.

170. Chokunonga E, Borok MZ, Chirenje ZM, Nyakabau AM, Parkin DM. Trends in the incidence of cancer in the black population of Harare, Zimbabwe 1991-2010. International journal of cancer Journal international du cancer. 2013;133(3):721-9.

171. Jedy-Agba E, Curado MP, Ogunbiyi O, Oga E, Fabowale T, Igbinoba F, et al. Cancer incidence in Nigeria: a report from population-based cancer registries. Cancer epidemiology. 2012;36(5):e271-8.

172. Rohner E, Valeri F, Maskew M, Prozesky H, Rabie H, Garone D, et al. Incidence rate of Kaposi sarcoma in HIV-infected patients on antiretroviral therapy in Southern Africa: a prospective multicohort study. Journal of acquired immune deficiency syndromes (1999). 2014;67(5):547-54.

173. Martin J, Wenger M, Busakhala N, Buziba N, Bwana M, Muyindike W, et al. Prospective evaluation of the impact of potent antiretroviral therapy on the incidence of Kaposi's Sarcoma in East Africa: findings from the International Epidemiologic Databases to Evaluate AIDS (IeDEA) Consortium. Infectious agents and cancer. 2012;7(Suppl 1):O19.

174. Begre L, Rohner E, Mbulaiteye SM, Egger M, Bohlius J. Is human herpesvirus 8 infection more common in men than in women? Systematic review and meta-analysis. International journal of cancer Journal international du cancer. 2016;139(4):776-83.

175. Sengayi MM, Spoerri A, Egger M, Giddy J, Maskew M, Singh E, et al. Risk of Cancer in HIV-Positive Adults on ART in South Africa: A Record Linkage Study. (Presented at CROI February 22-23 2016, Boston, MA Conference abstract 613).

176. Liu W, Snell JM, Jeck WR, Hoadley KA, Wilkerson MD, Parker JS, et al. Subtyping sub-Saharan esophageal squamous cell carcinoma by comprehensive molecular analysis. JCI Insight. 2016;1(16):e88755.

177. Kayamba V, Bateman AC, Asombang AW, Shibemba A, Zyambo K, Banda T, et al. HIV infection and domestic smoke exposure, but not human papillomavirus, are risk factors for esophageal squamous cell carcinoma in Zambia: a case-control study. Cancer Med. 2015;4(4):588-95.

178. Morhason-Bello IO, Odedina F, Rebbeck TR, Harford J, Dangou JM, Denny L, et al. Challenges and opportunities in cancer control in Africa: a perspective from the African Organisation for Research and Training in Cancer. The Lancet Oncology. 2013;14(4):e142-51.

179. Masamba LPL, Jere Y, Brown ERS, Gorman DR. Tuberculosis Diagnosis Delaying Treatment of Cancer: Experience From a New Oncology Unit in Blantyre, Malawi. J Glob Oncol. 2016;2(1):26-9.

180. Buyego P, Nakiyingi L, Ddungu H, Walimbwa S, Nalwanga D, Reynolds SJ, et al. Possible misdiagnosis of HIV associated lymphoma as tuberculosis among patients attending Uganda Cancer Institute. AIDS Res Ther. 2017;14(1):13.

181. Mbulaiteye SM, Bhatia K, Adebamowo C, Sasco AJ. HIV and cancer in Africa: mutual collaboration between HIV and cancer programs may provide timely research and public health data. Infectious agents and cancer. 2011;6(1):16.

182. Kelly H, Weiss HA, Benavente Y, de Sanjose S, Mayoud P. Association of antiretroviral therapy with high-risk human papillomavirus, cervical intraepithelial neoplasia, and invasive cervical cancer in women living with HIV: a systematic review and meta-analysis. The Lancet HIV. 2018;5:e45-58.

183. Hoffmann C, Sabranski M, Esser S. HIV-Associated Kaposi's Sarcoma. Oncol Res Treat. 2017;40(3):94-8.

184. Semeere A, Wenger M, Busakhala N, Buziba N, Bwana M, Muyindike W, et al. A prospective ascertainment of cancer incidence in sub-Saharan Africa: The case of Kaposi sarcoma. Cancer Med. 2016;5(5):914-28.

185. WHO guidelines for screening and treatment of precancerous lesions for cervical cancer prevention. World Health Organization. 2013. ISBN 978 92 4 154869 [Accessed at http://www.who.int/reproductivehealth/publications/cancers/screening_and_treatment_of_precancerous_lesions/en/ on January 15, 2018].

186. Fort VK, Makin MS, Siegler AJ, Ault K, Rochat R. Barriers to cervical cancer screening in Mulanje, Malawi: a qualitative study. Patient Prefer Adherence. 2011;5:125-31.

187. Ministry of Health- Sexual and Reproductive Health Services: VIA Programme Report. Lilongwe: Ministry of Health; 2013. .

188. Government of Malawi. Reproductive Health Directorate. Malawi: Lilongwe; 2014. Cervical Cancer Statistics.

189. Msyamboza KP, Phiri T, Sichali W, Kwenda W, Kachale F. Cervical cancer screening uptake and challenges in Malawi from 2011 to 2015: retrospective cohort study. BMC public health. 2016;16(1):806.

190. Maseko FC, Chirwa ML, Muula AS. Cervical cancer control and prevention in Malawi: need for policy improvement. The Pan African medical journal. 2015;22:247.

191. Rudd P, Gorman D, Meja S, Mtonga P, Jere Y, Chidothe I, et al. Cervical cancer in southern Malawi: A prospective analysis of presentation, management, and outcomes. Malawi medical journal : the journal of Medical Association of Malawi. 2017;29(2):124-9.

192. Maseko FC, Chirwa ML, Muula AS. Health systems challenges in cervical cancer prevention program in Malawi. Glob Health Action. 2015;8:26282.

193. Bruni L, Barrionuevo-Rosas L, Albero G, Serrano B, Mena M, Gómez D, Muñoz J, Bosch FX, de Sanjosé S. ICO Information Centre on HPV and Cancer (HPV Information Centre). Human Papillomavirus and Related Diseases in the World. Summary Report 27 July 2017. [Accessed on at http://www.hpvcentre.net/statistics/reports/XWX.pdf] on January 15, 2018].

194. Franceschi S, Jaffe H. Cervical cancer screening of women living with HIV infection: a must in the era of antiretroviral therapy. Clinical infectious diseases : an official publication of the Infectious Diseases Society of America. 2007;45(4):510-3.

195. World Health Organization. Malawi Country Profile. Global Status on Alcohol and Health. 2014 ed. [Accessed at

http://apps.who.int/iris/bitstream/10665/112736/1/9789240692763_eng.pdf?ua=1 on March 7, 2018].

196. World Health Organization. Malawi Country Profile. WHO report on the global tobacco epidemic, 2017. [Accessed at http://www.who.int/tobacco/surveillance/policy/country_profile/mwi.pdf on March 7, 2018].

197. World Health Organization. Technical and operational considerations for implementing HIV viral load testing. Geneva. Switzerland: World Health Organization; 2014. [Accessed at http://www.who.int/hiv/pub/arv/viral-load-testing-technical-update/en/ on March 2, 2018].

198. Lecher S, Williams J, Fonjungo PN, Kim AA, Ellenberger D, Zhang G, et al. Progress with Scale-Up of HIV Viral Load Monitoring - Seven Sub-Saharan African Countries, January 2015-June 2016. MMWR Morb Mortal Wkly Rep. 2016;65(47):1332-5.

199. Vallet-Pichard A, Pol S. Hepatitis viruses and human immunodeficiency virus co-infection: pathogenesis and treatment. J Hepatol. 2004;41(1):156-66.

200. Kowalkowski MA, Day RS, Du XL, Chan W, Chiao EY. Cumulative HIV viremia and non-AIDS-defining malignancies among a sample of HIV-infected male veterans. Journal of acquired immune deficiency syndromes (1999). 2014;67(2):204-11.

201. Riedel DJ, Rositch AF, Redfield RR. Patterns of HIV viremia and viral suppression before diagnosis of non-AIDS-defining cancers in HIV-infected individuals. Infectious agents and cancer. 2015;10:38.

202. Chiao EY, Hartman CM, El-Serag HB, Giordano TP. The impact of HIV viral control on the incidence of HIV-associated anal cancer. Journal of acquired immune deficiency syndromes (1999). 2013;63(5):631-8.

203. Gakunga R, Parkin DM, African Cancer Registry N. Cancer registries in Africa 2014: A survey of operational features and uses in cancer control planning. International journal of cancer Journal international du cancer. 2015;137(9):2045-52.

204. Ong TC, Mannino MV, Schilling LM, Kahn MG. Improving record linkage performance in the presence of missing linkage data. J Biomed Inform. 2014;52:43-54.

205. Achenbach CJ, Buchanan AL, Cole SR, Hou L, Mugavero MJ, Crane HM, et al. HIV viremia and incidence of non-Hodgkin lymphoma in patients successfully treated with antiretroviral therapy. Clinical infectious diseases : an official publication of the Infectious Diseases Society of America. 2014;58(11):1599-606.

206. Engels EA, Pfeiffer RM, Landgren O, Moore RD. Immunologic and virologic predictors of AIDS-related non-hodgkin lymphoma in the highly active antiretroviral therapy era. Journal of acquired immune deficiency syndromes (1999). 2010;54(1):78-84.