Estimation of incidence from cross sectional data, risk factors for early sexual debut and cancer trends in a population where HIV is prevalent: the case of Malawi

Doctoral Dissertation by

Humphrey E Misiri

Submitted to the Faculty of Medicine at the
University of Oslo in partial fulfillment of the requirements for the degree of
Philosophiae Doctor (PhD)

30-12-2015

© Humphrey E Misiri, 2016

Series of dissertations submitted to the
Faculty of Medicine, University of Oslo

ISBN 978-82-8333-100-4

All rights reserved. No part of this publication may be reproduced or transmitted, in any form or by any means, without permission.

Cover: Hanne Baadsgaard Utigard
Printed in Norway: 07 Media AS – www.07.no

Acknowledgments

I am very grateful to Prof Odd O.Aalen for facilitating supplementary funding for my PhD project. I also thank my co-supervisors Prof Abdi Edriss, Prof Ottar N. Bjornstad and Prof Johanne Sundby for their academic and moral support for this cause.

I am very thankful to Prof Arnoldo Frigessi for accepting to be my thesis co-supervisor at the last minute. I appreciate his valuable and "insightful" guidance in the very short time period during which he co-supervised me.

Not to be forgotten are all co-authors and collaborators who have contributed to my work. These are Professor Ottar Bjornstadt, Dr Freddie Bray, Dr Donald Maxwell Parkin and Dr Frederik Andreas Dahl.

I am also grateful to the Head of the Biostatistics Department in Domus Medica, Prof Magne Thoresen, at the Institute of Basic Medical Sciences (IMB) of the Faculty of Medicine , University of Oslo (UiO) for allocating office space to me during 2008-2012.

I appreciate the happy moments I enjoyed whilst in Oslo. I also appreciate the kindness and friendliness of some few people I interacted with in Norway. I also appreciate the the sociability and helpfulness of quite a few people I came across in Norway.

Many things happened during the period (2009-2012) I was a PhD student and the period that followed 2012. Some of these events were good, some bad ,some disturbing, some very bad and very disturbing. Resilience has been the term all along.

I am sorry that I was not available at all times to all those who needed me and to whom I matter(mishpakhah(gezin en familie) v^e khaverim sheli(keniss en vrienden)- those who had to go through a lot during my absence in Norway as a student at UiO. But then, I can't always be available to all those who need my presence and my support. Such is life. In life, variety is the spice. Perhaps, my absence acted as spice in the lives of these people.

I dedicate this thesis to my father and the following who are beyond the curtain of time: my maternal and paternal grannies-Mrs Mapundi, Mr and Mrs Sagawa and my mother Mrs Esime Misiri, my aunt Mrs Roseby Ling'ande, my cousins John("Mukhwera") and Martha, my brother Hurneck and my sister Rhoda .

Lastly, Ecclesiastes 12:12-14: "And further, by these, my son be admonished :of making many books there is no end; and too much study is a weariness of the flesh. Let us hear the conclusion of the whole matter:Fear God and keep His commandments: for this is the whole duty of man. For God shall bring every work unto judgment , with every secret thing, whether it be good, or whether it be evil."

Humphrey Esmott Misiri
Blantyre, 30 December, 2015.

List of papers

The thesis is based on the following papers referred to by their Roman numerals:

I. Misiri HE, Edriss A, Aalen OO, Dahl FA. Estimation of HIV incidence in Malawi from cross-sectional population-based sero-prevalence data. Journal of the International AIDS Society 2012; 15(14) .

   Link: http://www.slideshare.net/misiri1/estimation-of-hiv-incidence-in-malawi-from-crosssectional-populationbased-seroprevalence-data

II. Misiri H. Risk factors of early sexual debut among men and women - a strong predictor of HIV and sexual risk in Malawi. Research 2014; 1(625).

   Link : http://www.labome.org/research/Risk-factors-of-early-sexual-debut-a-strong-predictor-of-HIV-and-sexual-risk-in-Malawi.html

III. Misiri H, Dzamalala C, Edriss A, Parkin DM, Bray F. Cancer incidence in Malawi: time trends in Blantyre 1996-2005 and predictions up to 2015. Global Advanced Research Journal of Medicine & Medical Sciences 2012; 1(6); 145-153.

   Link: http://garj.org/garjmms/content/2012/july.htm

Contents

Acknowledgments III

List of papers V

1 Introduction 1

1.1 HIV and AIDS 2

1.2 HIV prevalence and incidence 5

1.3 The need for estimation of HIV incidence from cross sectional sero-prevalence data 6

1.4 Cancer 8

1.5 Cancer in Malawi-a historical background 9

1.6 Mathematical and statistical models in HIV and cancer research 12

1.7 Rationale of the thesis 14

1.8 Logical connection of papers in this thesis 17

2 Aims 18

3 Materials and methods 19

3.1 Data sources 19

3.2 Methods 22

3.2.1 Estimation of incidence from cross-sectional sero-prevalence data-Paper I 22

3.2.1.1 Estimation of standard errors of parameter estimates 24

3.2.1.2 Sensitivity analysis 25

3.2.1.3 Graphs of the model-based and age-specific HIV prevalence estimates 26

3.2.2 Frailty and logistic regression models - Paper II 28

3.2.2.1 Frailty Models 28

3.2.2.2 The Weighted Logistic Regression Model 32

3.2.3 Prediction of the standardized incidence rates of cancer and cancer burden - Paper III 34
3.2.4 Projections of cancer burden 36
3.2.4.1 Estimation of standard errors of the predicted number of cases (burden) 38
3.2.5 Analyzing the change in the annual number of cases of cancer 38

4 Summary of results 40
4.1 Paper I: Estimation of HIV incidence in Malawi from cross-sectional population-based sero-prevalence data 40
4.2 Paper II: Risk factors of early sexual debut among men and women - a strong predictor of HIV and sexual risk in Malawi. 41
4.3 Paper III: Cancer incidence in Malawi: Time trends in Blantyre 1996-2005 and predictions up to 2015 44

5 Discussion 46
5.1 Methodological Considerations 46
5.1.1 Estimation of HIV incidence from sero-prevalence data-Paper I 46
5.1.1.1 Limitations in Paper I 52
5.1.2 Risk factors of early sexual debut among men and women - a strong predictor of HIV and sexual risk in Malawi.- Paper II 54
5.1.2.1 Limitations in Paper II 60
5.1.3 Projections of standardized incidence rates of cancer and cancer burden for 2015- Paper III 63
5.1.3.1 Limitations in Paper III 66
5.2 Discussion of results 70
5.2.1 Estimation of HIV incidence from cross-sectional sero-prevalence data- Paper I 70
5.2.2 Risk factors of early sexual debut among men and women - a strong predictor of HIV and sexual risk in Malawi-Paper II 74
5.2.3 Projections of standardized incidence rates of cancer and cancer burden for 2015- Paper III 79

6 Conclusion 88

Bibliography 90

Papers included in the thesis 108

1 Introduction

The human immunodeficiency virus (HIV) was discovered in the United States of America (USA) among gay young men over three decades ago. Since that time, HIV has spread to other parts of the world. Although it was once believed that Acquired Immune Deficiency Syndrome (AIDS) was a disease of the poor, HIV is now prevalent in many countries in all continents of the world. HIV infection causes immuno-suppression. Because of immuno-suppression, the life of an individual who is HIV positive is characterised by frequent illnesses which in many cases would have been warded off by the immune system had the patient been healthy and HIV negative [1]. These illnesses which are ushered in by immune-suppression due to HIV infection are called opportunistic infections.

Antiretroviral therapy (ART) is used to treat HIV patients. ART suppresses the manifestations of opportunistic infections. Besides, manifestations of HIV infection like weight loss, change in hair texture and anaemia disappear once a patient is put on ART. Antiretrovirals (ARVs) are now available in many countries of the world including Malawi. Antiretroviral therapy is now a source of hope for the sick since it has improved the survival of HIV patients. Without doubt, HIV is an unnecessary evil which is the curse and plague of the twentieth century and beyond.

In this thesis we introduce a mathematical model for estimating HIV incidence from cross-sectional sero-prevalence data. Besides, we describe the patterns of occurrence of Kaposi sarcoma, non-Hodgkin lymphoma, esophageal cancer, eye cancer, breast cancer, cervical cancer and all other cancers combined in the population of Blantyre district of Malawi during 1996 - 2005, and predict the cancer burden for 2015 in the area. Three papers constitute this thesis. In Paper I, a mathematical model for estimating HIV

incidence from HIV sero-prevalence data is introduced and applied to Malawi data. In Paper II, statistical models are used to determine the risk factors of early sexual debut among men and women in Malawi. In Paper III, a description of the patterns of occurrence of Kaposi sarcoma, oesophageal cancer, non-Hodgkin lymphoma, eye cancer, breast cancer, cervical cancer and all other cancers combined in Blantyre, Malawi during 1996-2005 is given. Besides, cancer incidence trends during the same period are analyzed and projections of cancer burden and standardized incidence rates of cancer for 2015 are made using statistical models. Subsections 1.1 through 1.6 are the background to the thesis. Subsection 1.7 is the rationale of the thesis. Subsection 1.8 explains the logical connection of papers in this thesis. Section 2 presents the aims of this thesis. Section 3 is a description of the methods used in each of the three papers of this thesis. Section 4 gives a summary of the results. Section 5 is the discussion of the results. Section 6 is the conclusion section of the thesis.

1.1 HIV and AIDS

AIDS is predominantly a sexually transmitted disease. Just over 10% of the world's population is in Sub-Saharan Africa [2]. However, about 69% of the 25.4 million people living with HIV in the world are in Sub-Saharan Africa [2, 3]. Southern Africa is the worst affected sub-region in the world [2]. Malawi is a country in sub-Saharan Africa especially in Southern Africa. Malawi is land-locked and shares boundaries with Tanzania to the north and north east, Zambia to the north-west and Mozambique to the east, south and south-west. The country's major exports which are are tobacco, tea, and sugar account for approximately 85 percent of Malawi's domestic exports [4]. Although Malawi has achieved progress in immunisation [5], communicable diseases are still a major problem. In particular, malaria and acute respiratory infections (ARI) are the leading causes of death especially among underfive children [6, 4].

The human immunodeficiency virus was discovered in Malawi in 1985 [7]. Since that time, HIV has spread gradually throughout Malawi over the years. Currently, AIDS has achieved prominence in Malawi as one of the leading causes of death [8]. In 2012, HIV/AIDS was the leading cause among the top ten causes of death in Malawi [8]. Ac-

cording to the report of the Malawi Demographic and Health Survey conducted in 2010 , the national prevalence of HIV which was 12 percent in 2004 dropped to 11 percent by 2010 [4]. In 2013, the HIV prevalance in Malawi was 10.5% [9]. By 2014, the prevalence of HIV had dropped to 10 % [10]. The number of people living with HIV in 2010 was about 1 million [10]. Although these changes in HIV prevalence from 12 percent in 2004 to 11 percent in 2010, to 10.5% in 2013 and then to 10 percent in 2014 are not big, it is a sign of hope that it is still possible for the HIV prevalence to drop even further. A possible reason why the change in prevalence is not substantial may be the effect of antiretroviral drugs [11] in prolonging survival of HIV positive persons [12]. Antiretroviral drugs were introduced in Malawi in 2000 [13]. For over a decade now, ARVs have been rolled out to many parts of Malawi. By 30 June 2009, 266,325 HIV positive persons were receiving ART in Malawi [14]. At the beginning of 2012, 67% of all adults and children in need of ART were receiving ART [15]. At the end of the first quarter of 2012, 480,409 HIV positive persons (37% males, 63% females) were on ART [16]. The majority of those on ART were adults (91%) and the rest were children aged 0-14 years. A small percentage of females on ART (9%) were pregnant and the majority (91%) were not pregnant. By 2014, 496,202 people were on ART [17]. Currently, ART coverage on a national scale is reasonable (50 percent) [18] but incomplete. Antiretroviral therapy has reduced mortality in Malawi [19, 20]. In fact, AIDS mortality which was initially at 22,000 deaths in 1985 rose to 76,000 deaths in 2005 after which it decreased to 43,000 in 2011. By 2014, the number of deaths due to HIV was 33,000 [10]. This is due to the scale up of the provision of ART from 3,000 people on ART in 2003 to 382,953 people on ART in 2011 [15]. In 2012, the HIV incidence was 48,000 per 100,000 person-years (95%CI :41,000-46,000) [10]. In 2013, according to the The Global Burden of Disease 2013 [21], the HIV incidence rate among males age 15-49 was 7,370 per 100,000 person-years (95% CI:6,519 - 8,387) . The HIV mortality rate [21] for the same age band was 4,849 per 100,000 person-years (95%:3,792 - 6,091). Among women, the HIV incidence rate [21] for the age range 15-49 was 6,189 per 100,000 person-years (95% CI:5,510 - 7,018). The HIV mortality rate (per 100,000 person years) for the same age band was 3,453 (95% CI:2,663 - 4,392). HIV has reversed the gains achieved through immunization in Malawi and has wreaked havoc in many countries of the world. AIDS continues to be a menace in many developing countries. It is therefore essential to discover cost-effective measures

for monitoring it.

In Malawi, just like in all sub-Saharan countries, HIV is spread mainly through heterosexual contact [3]. Early sexual debut which is the first sexual act before the age of 15 years is a risk factor of HIV infection in Malawi. Besides early sexual debut is also a risk factor of unwanted teenage pregnancies, dropping out of school and risky sexual practices like unprotected sex and abortion . There is low condom use in Malawi [4, 22]. The shortage of drugs for treating sick people is also a persistent problem in Malawi [23]. Malawi is therefore an environment of relatively high HIV prevalence and risky sexual practices. Early sexual debut, therefore, undoubtedly predisposes an individual to HIV infection in such an environment. Vertical transmission of HIV has not been completely eliminated in Malawi and other poor countries. HIV infection is treatable but incurable. Treatment for HIV is available in Malawi and some countries in the world. WHO recommends the use of ART for the prevention of HIV infection of young children , pregnant women and populations exposed to HIV risk [24]. Indeed ART is given to breastfeeding mothers to protect suckling children in some settings [25]. Prognosis on ART is reported to be generally variable in low-resource settings [26]. However, according to one study conducted in West Africa, prognosis is better among women than among men [27]. Another West African study reports that stigma is negatively associated with health outcomes among people on ART [28]. The well-known effect of ARVs in prolonging survival has also extended to children who were infected with HIV perinatally. There is evidence that children who were infected with HIV perinatally have successfully grown into adulthood because of the life-prolonging antiretrovirals [29]. Although there is no data from Malawi, it is highly probable that Malawi also has such children. If that is so, the percentage of such children is likely to be small currently but may grow to be another big significant fraction of the Malawi population in the future. The future challenge, therefore, will be the management and continued care of aging adults on ART and adolescents on ART who were infected perinatally or otherwise but are growing up with HIV [30] with the attendant complications of long-term ARV use. The complications arising from long term use of medication among living adolescents who were infected with HIV perinatally are chronic lung disease [31], neuro-development issues [32], cardiovascular problems [33], kidney disease [34], mental health problems [35], bone health

problems [36], adherence to medication and metabolic complications [37]. Treatment of metabolic complications originating from long-term ART use in perinatally HIV-infected children is possible [37]. The complications of ageing with HIV among adults are many and varied [38, 39, 40, 41, 42, 43, 44]. The fact that HIV positives can now live longer implies that it is bound to be difficult to observe a reasonable reduction in the national HIV prevalence. Therefore, it is important to monitor the HIV epidemic by following closely, new cases of HIV.

1.2 HIV prevalence and incidence

HIV prevalence is often utilized as a measure of disease occurrence instead of HIV incidence because it is easier to estimate from more accessible data. Besides, the data from which prevalence can be computed is common. HIV sero-prevalence data are obtainable from HIV sentinel clinics in many countries in Africa. These clinics provide data for monitoring the AIDS epidemic [45]. Malawi has an HIV sentinel surveillance system which was established in 1994 with an initial 19 sentinel clinics attended mainly by pregnant women [46]. By 2007, there were 54 sentinel clinics in selected districts in all the three provinces of Malawi [46].

As a measure of disease occurrence, prevalence has several disadvantages. Prevalence is computed by dividing the total number of people who are alive and with disease or infection by the total population count. Individuals who are infected with HIV but who die from AIDS or other causes are not included in the computation of prevalence. Therefore, for diseases with a higher rate of mortality than all-cause mortality, the probability of inclusion into the cross-sectional sample of a would-be prevalence survey is not the same for diseased and healthy individuals [47]. The reason is that many diseased people who die from the disease prior to the prevalence survey are not part of the computation for prevalence as they are dead by the time individuals are being sampled for such a survey [48]. This is called differential selection. A direct consequence of differential selection is that the estimate of prevalence is somehow biased.

Incidence, however, is a better measure of disease occurrence. In particular, the incidence rate of infection or disease is a good measure of disease occurrence as it is the rate of development of disease in a community in a certain period [48]. In the computation of incidence rate, the numerator is the number of people who contract a disease or are infected during a specific period. These include those who die from the infection or disease. Therefore, unlike HIV prevalence which does not include individuals who die from disease, the incidence rate gives a true and better picture of the magnitude of an epidemic than prevalence as it tells more about the past and future occurrence of disease in the population. This is important for monitoring epidemics and planning interventions. Despite the advantages of incidence measures, the difficulty of obtaining good estimates of the number of people who are infected during a specific period makes incidence hard to estimate. Prevalence is easier, thanks to sentinel counts, and is therefore used in practice in many countries.

1.3 The need for estimation of HIV incidence from cross sectional sero-prevalence data

HIV incidence can be easily computed from longitudinal data. In fact, this is the best way of estimating incidence of HIV in any population. A specific geographic area is demarcated and the inhabitants followed for a certain time period. At specific intervals, blood samples are taken from them and tested for HIV. The accuracy of the estimation of incidence from this type of data derives from the completeness of the data collected and accuracy in tracing study participants.

Due to logistical and financial challenges, longitudinal HIV studies are not often conducted even in the so-called rich countries. In the first place, a longitudinal study is very expensive to conduct as it has to be conducted over a long period of time. Any shortage of funds midway through the project will automatically lead to the cessation of the study and incomplete data. Furthermore, tracking people through time is a challenge. Migration may lead to loss to follow up. Some people may voluntarily drop out of the study and this may create the problem of missing data. It is also difficult to convince people to be undergoing HIV tests on a regular basis without proper motivation.

Longitudinal studies can be easily conducted by governments just like national demographic and health surveys. Unfortunately, many governments are often too busy with other activities to contemplate funding longitudinal studies just for the sole purpose of estimating incidence as this could be easily seen as a waste of resources. Because of these problems, in many countries cross-sectional sero-prevalence data of HIV are more common than longitudinal data.

In Africa, the commonest HIV/AIDS data available are cross sectional sero-prevalence data but longitudinal data are scarce [49, 50]. Most of the available sero-prevalence data are data collected at one point in time. HIV sero-prevalence data collected more than once - (also known as serial sero-prevalence data) are also scarce [51]. Although cross-sectional HIV sero-prevalence data are more readily available than longitudinal data, some varieties of cross-sectional data can yield biased estimates of prevalence and incidence. For example, in many countries in sub-Saharan Africa, the commonest sero-prevalence data available are from sentinel clinics. Sentinel clinics are health facilities set up to facilitate monitoring of the AIDS epidemic. More often than not, the target of these clinics are pregnant women who are in the 15-49 year age range. Unfortunately, these women are just a fraction of the population in any country and in any setting and they do not represent men. Places of antenatal care coverage are not a random sample and so these areas may not be representative of all other geographic areas and regions within a country. Besides, pregnant women are a group of women vulnerable to HIV infection since pregnancy results from unprotected sex. Since sentinel clinics recruit pregnant women, estimates of HIV prevalence and incidence derived from antenatal clinic data from sentinel clinics are often biased and unrepresentative [52].

HIV sero-prevalence data are now collected through cross-sectional population-based studies called demographic and health surveys (DHSs). ORC Macro (now called ICF Macro), is an organization based in the United Sates which has conducted DHSs in several countries in Africa. Recent DHSs have included an HIV module. Demographic and Health Surveys employ good sampling strategies and as a consequence DHS data are often representative of the country where the data were collected.

The relative ease with which HIV sero-prevalence data can be collected and the easy availability of HIV sero-prevalence data is a motivating factor to provoke research into devising methods for estimating incidence from cross-sectional sero-prevalence data. This is so because any method that utilizes sero-prevalence data could be a cost-effective means of computing HIV incidence when compared to the use of longitudinal data from expensive and difficult-to-conduct longitudinal studies. Any new method discovered can help greatly in monitoring the spread of AIDS especially in a poor country like Malawi. Estimation of the incidence of HIV helps in monitoring the course of the AIDS epidemic. Such projections of HIV incidence are crucial as they aid in the formulation of polices geared at controlling the spread of HIV and planning for future health care needs [53]. It is therefore important to develop methods that use cross sectional sero-prevalence data to estimate incidence. The advent of AIDS has made diseases which were considered rare to be no longer rare. Such diseases are referred to as opportunistic. There is evidence from research that by 2001, AIDS had contributed to the occurrence of cancer in Blantyre [54] in Malawi.

1.4 Cancer

Cancer is a progressive disease. Because of this, some cancers like breast and cervical cancer can be prevented if detected very early in their course of development. Human beings were earmarked for a limited life span [55]. Consequently, when people advance in years, many health problems follow them. One of these is cancer.

Currently, life expectancy has improved in many countries of the world, on average, due to advances in science, technology and medicine [56]. Many diseases which were once fatal are now easily detectable and curable. As a direct consequence of this long life, cancer is now a major health problem worldwide and its burden is huge. Cancer is one of the major causes of death in many countries [57].

Besides changed life expectancy, there has been a global change in lifestyle [56]. Western habits like smoking and sedentary lifestyles have been adopted in developing countries [56]. As a result, diseases like lung and colorectal cancer which were common only

in western affluent societies are now common even in low to middle income countries [56]. Improved life expectancy, population growth and ageing are expected [56] in the future.

1.5 Cancer in Malawi-a historical background

Cancer has been documented in Malawi since 1969 when McGlashan published a research article on alcoholic spirits and oesophageal cancer in several countries in Central Africa which include Malawi [58]. Cancer had been a rare disease before the discovery of the human immunodeficiency virus in Malawi in 1985 [59]. Even though this is so, there are several publications on cancer in Malawi dating back to the seventies documentng the occurrence of retinoblastoma [60], liver cancer [61] and Kaposi sarcoma [62] in Malawi. Currently, there is evidence of the connection between HIV and some cancers [63, 64, 65].

The Malawi Cancer Registry was established in Blantyre, Malawi in 1989[38], four years after HIV was discovered. This was the first attempt at organized cancer data collection. The registry's main source of data when it was first established was the central histopathology laboratory at Queen Elizabeth Central Hospital (QECH) in Blantyre. Additional sources of data were health facilities which were visited regularly by registry clerks for cancer data. The laboratory at QECH received samples from the whole Malawi. QECH is the main referral center for Malawi. In 1993, the registry focused its data collection activities on Blantyre district as a catchment area making it a population-based registry. Although the Malawi Cancer Registry collected data on a national scale from 1989 to 1993, the data were far from complete as coverage was not complete. It was difficult to collect national cancer data from Blantyre as the main source of data for the cancer registry from its inception were diagnoses from the histopathology laboratory at Queen Elizabeth Central Hospital.

From 1993 to June 2011, the source of data on cancer in Malawi has been this registry. The Malawi Cancer Registry has been the sole provider of cancer data on Malawi to the International Agency for Cooperation on Cancer (IARC) (globocan.iarc.fr) [66, 67]. Besides, for the first time in history, Malawi has contributed data to Cancer Incidence In Five Continents Volume 10 [68]. These data are for 2003-2007. Since 1989, there

have been improvements in the diagnosis of cancer. Besides, several new hospitals have been opened in Blantyre. These changes have also contributed to an increase in the volume of cancer data collected. What persist are problems related to funding and these also affect the volume of data collected. The registry in Blantyre has no premises of its own to operate from. In July 2011, another pathological laboratory was established in Lilongwe at Kamuzu Central Hospital (KCH) [69]. This hospital is the main referral center for the central districts of Malawi. There is a database at KCH which contains cases which present at KCH. However, there is no cancer registry.

Cancer registration is a problem in Malawi as there is no national registry. Parkin and Bray (2009) define completeness of registry data as the extent to which all of the incident cancers occurring in the population are included in the registry [70]. Currently, the main problems faced by the Malawi Cancer Registry are incomplete coverage, lack of funding and infrastructure (offices) to operate from. There is no vital registration system in Malawi. Because of this, it is difficult for the registry to achieve complete coverage of the Blantyre population as some cancer data which would have been captured by the vital registration system by way of cancer deaths recorded locally are missing. In addition to this, the registry relies on laboratory results from Blantyre. Therefore, only cases of cancer referred to this laboratory can be detected and recorded. Besides, only cases presenting at hospitals in Blantyre can be detected and recorded. Therefore those who do not present at hospitals are those who are missed as they would have been captured by a functioning vital registration system had there been one in existence. This implies that, not all cases of cancer in Blantyre are recorded by the registry.

The registry in Blantyre is not the only institution facing many problems like lack of funding and incomplete coverage of the catchment population. The quality of data from African registries is said to be variable [71]. Complete coverage of catchment populations in African settings is another big problem. It is estimated that registries cover only 11 % of the population [71]. Parkin et al (2008) reports that the reasons for slow growth of cancer registration in Africa are defining residents of a particular area, identifying patients with cancer and obtaining accurate diagnostic information for those with cancer [71]. In general, Parkin et al (2008) reports that cancer registration is more

difficult in Africa than in developed countries [71]. Therefore, as regards data quality in cancer registries of African and developed countries we can only speak in relative terms.

There is some level of cancer care in Malawi. Patients are given chemotherapy. Patients requiring radiotherapy are sent to Tanzania or Zimbabwe for such treatment. There are also attempts to palliate cancer [72]. The major problem is lack of resources. Therefore, it is important for the government to plan properly so that cancer awareness, treatment and palliation activities receive a fair share of the budget which is based on dwindling national financial coffers.

HIV has been in Malawi for over two decades. Besides, AIDS has to some extent also contributed to the occurrence of some cancers in Malawi. Successes in rolling out ART treatment in Malawi means that life expectancy of HIV positive persons will increase. With an increase in life expectancy, will be occurrence of cancer among the elderly HIV positive persons.

Computations of the burden of cancer are useful for planning for prevention and curative measures. Governments are able to plan for the future based on these estimates of future cancer burden. Analyses of trends of incidence rate of cancer are also important. The first report about cancer incidence in Blantyre came out in 2001 [54]. Since the inception of the Malawi Cancer Registry, no attempt has been made to estimate the future burden of cancer in Blantyre and to conduct an in depth analysis of the trends of standardized incidence rates of different cancers. In fact, Parkin et al (2008) reported that knowledge of cancer trends in Africa is inadequate [71]. According to the World Cancer Report 2014, cancer incidence is on the rise in many countries of the world [57]. This includes Malawi.

1.6 Mathematical and statistical models in HIV and cancer research

Research on Acquired Immune Deficiency Syndrome (AIDS) and cancer is ongoing in many countries in the world. Through research, the connection between HIV and cancer was unraveled. Other fruits of research are the development of the human papilloma virus (HPV) vaccine and the discovery of some risk factors of HIV and cancer. These discoveries have helped in devising some strategies to contain the spread of HIV and to control the occurrence of cancer.

Mathematicians and statisticians often use models in research as tools to solve scientific problems. Models are used to provide insight into the underlying data-generating mechanism of a scientific problem. In the 19th century and thereafter, scientists have formulated and applied models to solve biometric and medical problems. Models differ in the level of complexity and parameterisation. Depending on the context, some models are very complicated and have many parameters. In some cases, if a model is too complicated it might have parameters which do not make sense. On the contrary, some models are simple and have few parameters. Simple models are easy to understand and use. Models are not perfect but good models provide some useful insight into research problems and help answer questions. In fact, George E. Box once said, "Essentially, all models are wrong, but some are useful" [73]. Mathematical models have been used in the evaluation of vaccination policies in the United Kingdom [74]. Several models have been developed to help in the estimation of incidence from cross sectional seroprevalence data. These focus on the estimation of the rate of infection

The force of infection (FOI) is the rate of infection among susceptible members of a population. Hugo Muench first propounded a catalytic model for the estimation of the force of infection [75]. Griffiths (1974) [76] followed after Muench with his linear model for estimating the FOI of measles. Then came Grenfell and Anderson (1985) with their model [77]. Others have also formulated models about the force of infection in different contexts [78, 79, 80, 81]. Note that Muench, Griffiths (1974) and Grenfell and Anderson (1985) propounded catalytic models. Such models rely on the core assumption of

life-long immunity and negligible disease mortality. Because of this, catalytic models have to be modified to accommodate HIV infection which is irreversible and not immunizing. An overview of the estimation of the FOI from Hugo Muench in 1974 to 2009 is in a review by Hens et. al (2009) [82]. Grenfell and Anderson(1985)'s model was for estimating the FOI for measles using polynomials [77]. Shkedy et al (2006) extended the model to use fractional polynomials[8]. Keiding et al (1996) [83] introduced the use of link functions in the estimation of the age distribution at immunization. This paper helped to embed the estimation of the force of infection within the generalized linear models framework. Namata et al (2007) [84] propounded a model for estimating the force of infection using generalized linear mixed models following Keiding et al (1996). Finally, Hens et al (2012) [85] formalized the estimation of the FOI within statistics. Note that the FOI is often used in statistical models [86, 87, 88, 89, 90] for approximating the incidence rate of infection. Unlike the incidence rate which describes the rate of infection in the whole population in general, the force of infection approximates the the rate of infection only among susceptible members of the population.

Klaus Langohr gives a review of methods for estimating incidence from cross-sectional sero-prevalence data [51]. In addition to the methods described in his review, there are other approaches by Williams et al(2001) [91], Hallet et al (2008) [92] and Rajan and Sokal (2010) [93] which are also used for estimating incidence from cross-sectional sero-prevalence data. The methods in Langohr's review can be grouped into two: those that are based on the assumption of equal mortality among infected and healthy individuals and those that are based on the assumption of differential mortality. The assumption of equal mortality holds for some eye diseases like open angle glaucoma [94] but does not hold for HIV. HIV greatly affects mortality. Some estimation methods are also based on the assumption that the time to infection, time to death after infection and time to death of healthy individuals follow a certain probability distribution whereas some methods are nonparametric.

Langhor's review includes the approach by Leske et al (1981) for estimating the incidence of open angle glaucoma from cross-sectional data assuming equal mortality between healthy and infected persons [57]. Podgor et al (1986) developed a new method

for estimating incidence from cross-sectional prevalence data by extending Leske et al (1981)'s approach. Their method is for estimating incidence of disease from cross-sectional prevalence data assuming differential mortality among infected and healthy individuals for irreversible diseases. This method assumes that the force of infection is constant within age bands and incidence is estimated for these age bands.

Apart from mathematical models, statistical models are also useful in cancer research. Age and sex specific standardized incidence rates, for example, can be computed from empirical longitudinal data directly. Alternatively, mathematical and statistical models can be used to estimate age and sex specific standardized incidence rates. For example, compartmental models [95] can be used to estimate the incidence rates of HIV. Poisson regression models are also used for estimating the ratio of the rates of cancer occurrence. Time-to- event data can be analyzed using semi-parametric and parametric regression models. Random effect survival models are also used for analysing clustered time-to-event data. Rates and survival times can also be analyzed using loglinear models [96]. Prediction of cancer rates can also be performed using time-linear [97, 98] and age period cohort models [99]. Binary response variables can be analysed using logistic regression models. However, if a binary response and explanatory variables are from a multistage cluster sample, weighted logistic regression models are usually preferred to standard logistic regression models.

1.7 Rationale of the thesis

HIV sero-prevalence data are cheaper to collect than longitudinal data. Many countries rely on HIV prevalence to monitor the spread of HIV because they can not afford to monitor the spread of HIV by following populations prospectively in order to estimate HIV incidence. Malawi also uses the HIV prevalence estimated from sentinel clinics to monitor the HIV epidemic. Although Malawi has registered a decrease in HIV prevalence but the magnitude of the change implies that for every decade HIV prevalence is decreasing by 2%. If we continue at this pace, it might take 20 years for the national HIV prevalence to drop to 2%. The fact is that HIV prevalence alone does not provide details about the progress of an epidemic. HIV incidence provides a clearer and

detailed picture of the spread of HIV in a given setting. Any mathematical model that uses age-specific HIV sero-prevalence data to estimate HIV incidence could be handy. The motivation to devising methods for computing HIV incidence should be the easy availability of HIV sero-prevalence data. Any method that utilities sero-prevalence data could be a cost-effective means of computing HIV incidence when compared to the use of incidence data collected from longitudinal studies which are often logistically difficult to conduct. If a new method for estimating incidence that uses seroprevalence data is discovered, incidence estimates will be computed and used for monitoring the HIV epidemic in Malawi and other places where HIV is prevalent. This would be a new development since such an approach for estimating incidence is cost-effective. Any such method discovered can help greatly in monitoring the spread of AIDS especially in a poor country like Malawi. Incidence is better for monitoring the HIV epidemic [100]. Projections of HIV incidence are essential in the formulation of polices geared at controlling the spread of HIV and planning for future health care needs [53].

Early sexual debut which is a known to be risk factor of HIV infection is the first sexual act before the age of 15 years. In Malawi, early sexual debut is also a risk factor of HIV infection [101]. Besides, early sexual debut is also a risk factor of unwanted teenage pregnancies, dropping out of school and risky sexual practices like unprotected sex and abortion [102, 103]. In America, Asia and Europe, the effects of early sexual debut are reported to be HIV infection and the following: substance use, violent delinquency, unwanted pregnancies, sexually transmitted diseases, unprotected sex, abortion, cervical cancer, prostitution, multiple sex partners and paying for sex [104, 105, 106, 107, 108, 109]. In Malawi, HIV is endemic. Although the HIV prevalence is relatively high, there is the tendency for people to indulge in unprotected sex [22, 4]. Since HIV is treatable, ART is the source of hope to the HIV positives. However, there is the problem of persistent shortage of drugs for treating sick people [23]. Without doubt, early sexual debut, therefore exposes a young person to the risk of HIV infection in such an environment. Globally, it is currently known that it is possible for HIV positive children to grow into adulthood thanks to the effects of ART. This means that it is possible for sexual minors who are growing with HIV to transmit the virus to their peers during early sexual debut. Although ART is beneficial, long-term

use of ARVs is associated with unpleasant complications. Therefore a young person who is infected with HIV during early sexual debut has an unpleasant future of living with HIV chronically with the attendant complications. It is therefore, important for young people to delay sexual debut in order to protect themselves. Since HIV is spread through sexual contact, if sexual debut is delayed the spread of HIV will be checked. If the factors associated with early sexual debut are known, interventions tailored to delay sexual debut can be devised and implemented. Determination of the significant risk factors of early sexual debut can be performed using statistical models if appropriate data is available.

Some cancers are associated with HIV infection. These cancers include eye cancer, cervical cancer, Kaposi sarcoma and non-Hodgkin lymphoma. Different types of cancer occurring in Malawi have been documented. It is now known that cancer is not rare in Malawi. In the coming decades, non-communicable diseases like cancer will become an increasing burden on health services in Malawi as well as in other low and medium income countries [110]. The first publication about cancer incidence in Blantyre was published in 2001 (Banda et al (2001)) [54]. This report described the incidence of cancer in Blantyre between 1991 and 1998 inclusive. According to Banda et al (2001), the incidences of Kaposi sarcoma and cervical cancer were very high during 1994-1998. The incidences of squamous cell carcinoma of the conjuctiva (eye cancer), oesophageal cancer and breast cancer were also high during the same period. Non-Hodgkin lymphoma (NHL) was the third common cancer among males. The incidence of non-Hodgkin lymphoma among women was also reasonable (3.0 per 100,000 person-years). A description of types of cancer that occurred in Blantyre during 1996-2005 can also give a picture of the rates and trends of occurrence of cancer during that period. Furthermore, it is also informative to know the future number of cancer cases. Such information is essential for planning cancer services and making decisions concerning resource allocation to prevention, palliation and treatment of cancer. In order to increase the scope of analysis, it is necessary to the estimate cancer burden and to describe the patterns of occurrence of just few chosen types of cancers namely Kaposi sarcoma, oesophageal cancer, non-Hodgkin lymphoma, eye cancer, breast cancer,cervical cancer and all other cancers combined. The reasons why these cancers were chosen are three-fold. Firstly,

oesophageal cancer, cervical cancer, breast cancer, eye cancer and Kaposi sarcoma occurred with high incidence during 1994-1998. Secondly, NHL was the third common cancer among men during 1994-1998 and its incidence among women was reasonable. Lastly, cervical cancer,NHL, eye cancer and Kaposi sarcoma are AIDS -defining cancers.

1.8 Logical connection of papers in this thesis

The logical connection of the papers in this thesis is as follows: HIV is spread mainly through heterosexual contact in Malawi. Early sexual debut is a risk factor of HIV infection. In 2010 [4], 17 % of women whose sexual debut was at age 15 or younger were HIV positive. For that reason, in Paper II, statistical models are used to determine the risk factors of early sexual debut among men and women in Malawi. The HIV epidemic in Malawi is monitored using HIV prevalence but HIV incidence is the best measure for monitoring the epidemic. Therefore in Paper I, we introduce a mathematical model for estimating HIV incidence from cross-sectional sero-prevalence data. The frequency of occurrence and incidence of AIDS-associated cancers (Kaposi sarcoma, eye cancer, non-Hodgkin lymphoma and cervical cancer) and non-AIDS defining cancers (breast cancer and oesophageal cancer) has increased considerably in Malawi according to Banda et al (2001) [54]. Therefore in Paper III we use statistical models to estimate age standardized incidence rates and burden of AIDS-defining cancers (Kaposi sarcoma, eye cancer, non-Hodgkin lymphoma, cervical cancer) and non-AIDS-defining cancers(breast cancer,oesophageal cancer, and all other cancers combined).

2 Aims

The first aim of this thesis is to develop a new method for estimating HIV incidence from cross-sectional sero-prevalence data accounting for differential mortality and to apply the method on data from Malawi. The new method will give FOI and population incidence estimates for HIV in Malawi. Furthermore, through an intuitive conditional probability formulation of the mortality adjusted prevalence, we will compute the total likelihood function, which will enables us to estimate our model parameters simultaneously. This approach will enable us to estimate a model with few degrees of freedom. We will also apply a bootstrap technique to estimate confidence intervals and also run sensitivity analyses. The FOI estimates from our model will be converted to population incidence rates by compensating for the estimated age-dependent prevalence

Early sexual debut is a risk factor for HIV infection in Malawi. Therefore, the second aim of the thesis is to determine the risk factors of early sexual debut in Malawi. We will fit accelerated failure time models to obtain time ratios. We will also fit weighted logistic regression models in order to obtain odds ratios. From the lo-gistic regression models, significant risk factors of early sexual debut will be determined.

The third aim of this thesis is to describe the patterns of Kaposi sarcoma, non-Hodgkin lymphoma, esophageal cancer, eye cancer, breast cancer, cervical cancer and all other cancers combined in the population of Blantyre district in Malawi using cancer incidence data for 1996-2005. We will also predict the future incidence rates and burden of these cancers in the area on the basis of projected population changes.

-

3 Materials and methods

This section contains information on materials like data analysed for the papers in this thesis and the actual methods used for analyzing the data. Different data sets were analyzed for Papers I - III. The sources of data for each paper are explained in Subsection 3.1. Subsection 3.2 explains the methods used to analyze data for each of the three papers in this thesis.

Subsection 3.2.1 explains the new method that solves the problem of estimating HIV incidence from cross-sectional sero-prevalence data mentioned in Section 1. The method for determining the risk factors of early sexual debut in Malawi are explained in Subsection 3.2.2. Subsection 3.2.3 explains the models used for solving the problem of computing standardized incidence rates and cancer burden mentioned in Section 1.

3.1 Data sources

The new method in Paper I was tested using three data sets namely HIV sero-prevalence data, natural and HIV mortality rates. The natural mortalities used in Paper I were extracted from the Malawi Demographic Survey 2004 report. These mortalities were computed from the Malawi Demographic Survey 1992 by experts from ICF Macro, a US-based company which has conducted demographic and health surveys in many countries of the world. The methods used for computing the natural mortality rates are in literature [111].

For our model we need natural mortality rates which are not contaminated by HIV mortality. Natural mortality rates are contaminated by the HIV epidemic if deaths of persons who are HIV negative is in some way linked to the pressures of life and strain

brought about through the support of the HIV infecteds. The natural death of an HIV-negative person might be accelerated by the sickness or death of a person who is HIV positive but also connected to the incumbent in one way or another. In that way, HIV infection or AIDS is the component cause [112] of the natural death of the person who is HIV-negative. HIV-positive persons have a support network of guardians, relatives, friends and medical personnel. Those who bear the brunt of caring for HIV-positives are the guardians who are usually HIV-negative. Caring for an HIV positive person always puts some strain on the life of the guardian. This strain can trigger a sequence of events which might lead to the poor health and eventual death of the guardian. Therefore, in a population with a reasonable proportion of HIV-positives, the natural mortality rates are affected in that way. In that case, we say that the natural mortalities have been "contaminated" by the AIDS epidemic. The natural mortalities used in the model in Paper I are from a period when mortality in Malawi was not affected by the AIDS pandemic. HIV was discovered in 1985. By 2004, HIV had spread to many parts of Malawi. The adult natural mortality rates of 2004 were in some way affected by the HIV epidemic. That is why we used natural mortality rates from a period when the HIV prevalence was not very high.

The HIV sero-prevalence data were from the HIV component of the Malawi Demographic and Health Survey conducted in 2004 (MDHS2004). The MDHS2004 survey was a nationally representative cross-sectional study conducted in 2004. The rates of mortality due to natural causes were extracted from the MDHS2004 report. The rates of mortality due to HIV infection were extracted from a longitudinal research study conducted in rural Malawi by Crampin et al (2002) [113]. In their own words, Crampin et al (2002) reported that the HIV mortality rates were high [113]. The HIV mortality rates were consistent with extent of the HIV epidemic at that time and with estimates from other African countries [113]. Data from the Malawi Demographic and Health Survey 2010 (MDHS2010) were analyzed for Paper II. The MDHS2010 survey was also a nationally representative cross-sectional study conducted in Malawi in 2010.

Two data sets were analyzed for Paper III namely cancer incidence data for the 1996-2005 period from the Malawi Cancer Registry and population data of Blantyre district

for the same period. The cancer incidence data were anonymized to preserve confidentiality. The population data of Blantyre for the 1996-2005 period as well as for 2015 were obtained in several ways. The population data for 1996 and 1997 were estimated using linear interpolation between the census data for 1987 [114] and 1998 [115]. The population data for 1998 were of course, extracted from the 1998 population census. The population data for 1999 to 2005 were estimated by linear interpolation between the census data of 1998 and 2008 [116] within each sex and age group. The population data for 2015 were extracted from the official population projections. The National Statistical Office of Malawi (NSO) has population projections for each district in Malawi [117]. These projections are for years from 2009 to 2031 and are downloadable from the website of NSO.

The data analysed for Paper III are from an African registry. The IARC sets the standards of data quality for cancer registries. Western countries have the resources and the relevant systems in place to help them to collect cancer data of very high quality. Vital registration systems, for example, are found in western countries but not in many African countries. Considering the problems which the Malawi Cancer Registry faces and the measures taken to clean the data, the quality of the data are good. Recently, the data from the cancer registry in Blantyre were found to be of acceptable quality to be included in the publication called Cancer Incidence in Five Continents Volume X [68]. Malawi was one of four countries from sub-Saharan Africa whose data were found to be of acceptable quality for inclusion into this volume. The remaining three sub-Saharan countries are South Africa, Zimbabwe and Uganda. Other African countries whose data were accepted were Egypt, Libya and Algeria. These are North African countries. This means that from the whole of Africa, data from only seven countries are included in this publication [68]. This implies that data from the Malawi Cancer Registry are deemed to be of good quality internationally.

3.2 Methods

This section explains the methods used to analyze data for each of the three papers in this thesis. Subsection 3.2.1 explains the new method that solves the problem of estimating HIV incidence from cross-sectional sero-prevalence data mentioned in Section 1. The models for determining the risk factors of early sexual debut in Malawi are explained in Subsection 3.2.2. Subsection 3.2.3 explains the models used for solving the problem of computing standardized incidence rates and cancer burden mentioned in Section 1.

3.2.1 Estimation of incidence from cross-sectional sero-prevalence data-Paper I

Klaus Langohr (2010) gives an overview of methods for estimating incidence from cross-sectional sero-prevalence data [51]. In addition to the methods described in his review, there are other approaches by Williams et al (2001) [91], Rajan and Sokal (2010) [93] and Misiri (2014) [118] which are also used for estimating incidence from cross-sectional sero-prevalence data.

The methods in Langohr's review can be grouped into two: those that are based on the assumption of equal mortality among affected and healthy individuals and those that are based on the assumption of differential mortality. The assumption of equal mortality holds for some diseases like open angle glaucoma but does not hold for HIV. HIV greatly affects mortality. Some estimation methods in Langhor's review are also based on the assumption that the time to infection, time to death after infection and time to death of healthy individuals follow an assumed probability distribution whereas some methods are nonparametric.

Paper I describes a method for estimating incidence from cross-sectional sero-prevalence data [119]. The force of infection (FOI) is the rate of infection among the susceptible members of the population. The FOIs for age ranges 15-24, 25-34 and 35-54 among men and 15-24, 25-34 and 35-49 age ranges among women are estimated together with the prevalence at age 15 for males and females. The dimensions of measurement for the FOI are the number of infections per population size per unit time. In one year, the

FOI reduces to the number of infections divided by the population size which translates to the risk of infection among susceptible members of the population. In Paper I, the FOI is age-dependent and piece-wise constant. In other words, the FOI is constant within the age intervals 15-24, 25-34, 35-54 among men and 15-24, 25-34, 35-49 among women. For age range i, let N_i denote the number of people in age rage i and X_i denote a variable that takes the value 1 for an individual who is HIV positive and 0 for an individual who is HIV negative. Let H_i be the sum of X_i. Then, the number of individuals who are HIV negative in age range i is $N_i - H_i$ and the prevalence of HIV for age range i (i = 1, 2, 3), π_i, is $\pi_i = \frac{H_i}{N_i}$. Note that $\lambda_i.(N_i - H_i)$ gives the expected number of new HIV infections in range i. Consequently, $\frac{\lambda_i.(N_i - H_i)}{N_i} = \lambda_i.(1 - \pi_i)$ is the incidence rate of HIV in age range i (i=1, 2, 3) as given in the Methods Section of Paper I.

Note that by making the FOI piece-wise constant within defined age ranges, the time to infection in those age ranges is , by default, exponentially distributed. This assumption of an exponential distribution of the time to infection is the fundamental building block for the model in the paper by Podgor and Leske (1986) [120]. The assumption is implicit in the following relationship [49] between prevalence , incidence and duration of disease:

$CI = 1 - e^{-I.(\Delta t)}$

where

CI is the cumulative incidence or risk of disease

I is the incidence rate

Δt is the elapsed time between the beginning and ending of an interval

Podgor and Leske (1986)'s model is also implemented by Saidel et al (1996) [49]. In fact, Podgor and Leske start by defining three parameters λ_1, λ_2 and λ_3 where λ_1 is the death rate among those who are disease-free, λ_2 is the death rate among the diseased and λ_3 is the disease incidence rate. By defining age intervals (a, a+t), an exponential distribution of time to infection is assumed and special discrete equations of the number of infected and disease-free people at the beginning and the ending of each interval are formulated. The conditional probability of surviving given that a person was HIV positive at the beginning of an age interval is derived using λ_3. The conditional probability

of developing disease and surviving given that the individual was HIV negative at the beginning of an age interval is derived using λ_1, λ_2 and λ_3. The conditional probability of of remaining disease-free and surviving given that a person was HIV negative at the beginning of an age interval is derived using λ_1 and λ_3. These probabilities are carefully substituted into the first equations. The final result is an equation for incidence estimation. Note that estimation of incidence is done for one age interval at a time. Thus unlike us, Podgor and Leske (1984) uses a parametric approach on grouped data. Our approach though assuming the FOI to be piece-wise constant, uses data which is organized into single-year categories. Besides, our method is simpler since estimation of parameters is simultaneous and direct without having to use difficult -to-solve nonlinear equations. We only estimate a few parameters.

3.2.1.1 Estimation of standard errors of parameter estimates

The standard errors of the four parameters of our new model can be estimated in two different ways. Using the first approach, the standard errors of parameter estimates can be extracted from the variance-covariance matrix which is obtained by inverting the Hessian matrix produced by **nlm**. The square root of the diagonal elements of the variance-covariance matrix are the standard errors. The second approach is to estimate standard errors by a well known method called bootstrapping. Bootstrapping was invented in 1979 by Efron [121].

Suppose we want to find the standard errors of a parameter, θ by bootstrapping. Let X_1, X_2,..., X_n be a random sample from X. In what follows, I will explain the bootstrap procedure for estimating the standard error in the simple case where one parameter only is being estimated. Generalisation of the method to a multi-parameter case is straight forward. The bootstrap procedure works in what can be described as follows:

I. Consider the empirical distribution function $F_n(x)$ of X_1, X_2,...,X_n. The empirical distribution function is also known as the sample distribution function. It is based on the sample X_1, X_2,...,X_n. Define an indicator function $I\,(X \leq x)$ thus:
   $I\,(X \leq x) = 1$, if $X \leq x$
   $I\,(X \leq x) = 0$, if $X > x$

The sample or empirical distribution function is defined as $F_n(x) = \frac{\sum_{i=1}^{n} I(X_i \leq x)}{n}$.

II. Denote a sample drawn with replacement from the data thus :$\underline{X}_n^* = (X_1^*, X_2^*, \ldots, X_n^*)$, where each sample has distribution function F_n. Each sample drawn with replacement is called a resample of the sample data. For a given sample of size n, there are $\binom{2n-1}{n}$ different resamples that can be drawn from the data. For example if n=10, $\binom{19}{10}$= 92,378 resamples can be drawn. Generate B resamples and denote them by $\underline{X}_n^*(b)$ for b=1,2, ..., B.

III. Let $T_n = T(\underline{X}_n, F_n)$ be the estimator for θ. From each resample, $\underline{X}_n^*$, estimate θ using $T_n^* = T(\underline{X}_n^*, F_n)$. There will be B estimates of θ.

IV. Compute the mean of the B estimates of θ using the following formula: $\widehat{\theta} = \frac{1}{B}\sum_{b=1}^{B} T^*(\underline{X}_n^*(b), F_n)$. $\widehat{\theta}$ is the bootstrap estimate of θ.

V. Compute the bootstrap estimate of the variance of θ, $\widehat{Var(\theta)}$, using the following formula: $\widehat{Var(\theta)} = \frac{1}{B-1}\sum_{b=1}^{B}(T^*(\underline{X}_n^*(b), F_n) - \widehat{\theta})^2$. Note that the standard error of $\widehat{\theta}$ is $\sqrt{\widehat{Var(\theta)}}$.

In order to obtain 95% bootstrap intervals for the FOI, 10,000 resamples of the data were drawn. Estimates of the FOI were computed from each resample. 10,000 estimates of the FOI were computed. The lower and upper limits of the 95% bootstrap-based intervals for the FOI were the estimates of the 2.5th and 97.5th percentiles from the estimated distribution function of the FOI.

3.2.1.2 Sensitivity analysis

To examine the performance of the model and the behaviour of the parameter estimates , 12 scenarios were created. The first six scenarios were created by decreasing the HIV and natural mortality rates by 2.5%, 5%, 7%, 10%, 15% and 20%. The second six scenarios were created by increasing the HIV and natural mortality rates by the 2.5%, 5%, 7%, 10%, 15% and 20%. For each scenario, the performance of the method and the behaviour of the estimates were examined.

3.2.1.3 Graphs of the model-based and age-specific HIV prevalence estimates

The data in Paper I comprise single ages ranging from 15 to 54 years for men and from 15 to 49 years for women. At each age a, the number of people who tested HIV-positive and the number who tested HIV-negative are also given. Therefore, for each age a, the prevalence of HIV , p_a, can be computed from the data. The resulting estimates are empirical estimates of HIV prevalence. Note that the HIV prevalence at age a ,p_a, is a response variable and age a is the explanatory variable. A scatterplot of p_a and a will show the pattern or nature of the relationship between p_a and a. In Figures 1 and 2 of Paper I, HIV prevalence curves are superimposed on the scatterplots of age and HIV prevalence computed from the sample data for men and women respectively. The objective is to find out how close the model fits the data.

In general, for any response variable Y_i (i = 1,2,...,n) and an explanatory variable X_i (i=1,2,...,n), it is not always easy to determine the pattern of the relationship between the two variables from a scatterplot. It is advisable to graphically show the pattern of the relationship between the mean response and the predictor by a curve of the function which describes the relationship between the mean response and the explanatory variable. A regression curve is the graph which portrays the mean relationship between Y_i and X_i. Unlike the scatterplot, the regression curve clearly shows the nature or pattern of the relationship between the response and the explanatory variables. In the case of one response variable and one explanatory variable, the relationship between the response and the explanatory variable can be given by the following relation:

$Y_i = m\left(X_i\right) + \epsilon_i$

where

Y_i is the response variable

X_i is the explanatory variable

ϵ_i is the random error term. The error terms are assumed to be independent and identically distribted with mean 0 and variance σ^2

$m\left(X_i\right)$ is the regression function

The regression curve can be estimated by assuming that that the regression function takes some functional form. This is called the parametric approach. For example, for simple linear regression the error terms are assumed to follow a normal distribution with mean 0 and varance σ^2. Besides, the regression function is given as $m(X_i) = \beta_0 + \beta_1 X_i$.

Alternatively, the regression curve can also be estimated without assuming that the regression function takes a particular form and that the error terms have a certain probability distribution. This is called the nonparametric approach. Smoothing is the approximation of a regression curve without assuming a particular form for the regression function [122]. Nonparametrically, the regression curve can be estimated using the method of local averaging [123]. Besides, special functions called smoothers can also be used to estimate a regression curve. Examples of smoothers [123] are the Nadaraya-Watson estimator($\widehat{m}(X_i)^{N-W}$), the Gasser-Muller estimator and the Priestly-Chao estimator [123]. The Nadaraya-Watson estimator is given by the following formula:

$\widehat{m}_h^{N-W}(x) = \frac{\frac{1}{n.h_n}\sum_{i=1}^{n} K(\frac{x-X_i}{h_n})Y_i}{\left[\frac{1}{n.h_n}\sum_{i=1}^{n} K(\frac{x-X_i}{h_n})\right]}$.
where
K(.) is a function called a kernel
h_n is the smoothing parameter called the bandwidth
n is the sample size

Examples [123] of kernels are the rectangular kernel, the Epanechnikov kernel and the Gaussian kernel. The Gaussian kernel is given by $K(x) = \frac{1}{\sqrt{2\pi}}.exp\left(-\frac{x^2}{2}\right)$. Note that the bandwidth controls the smoothness of $\widehat{m}_h^{N-W}$. Increasing the bandwidth makes the regression curve smoother. The smoothed estimates of HIV prevalence shown in Figures 1 and 2 of Paper I were computed using the Nadaraya-Watson estimator. We used the Gaussian kernel. In R, smoothing using the Nadaraya-Watson estimator can be performed by the **ksmooth** function in the **stats** library.

The contribution of the present mathematical model to earlier research is two-fold namely the output and the estimation procedure itself. Concerning the output, our model gives the FOI and population incidence estimates of HIV for Malawi. Mind you,

this is novel.

As for the estimation procedure and approach, the assumptions of endemic equilibrium and age-dependent prevalence are not new. Other researchers [77, 94, 120] have used the same assumptions in their research. However, whereas a formal approach would have been equally good, we used an intuitive approach in the derivation of the conditional probability of being HIV positive given that one is alive. The result is a recurrence relation that gives an expression for the prevalence of HIV adjusted for both natural and HIV mortality. At each age a, the number of HIV positives from a group of both HIV negatives and HIV positives is assumed to follow a binomial distribution. A likelihood function is formulated using the binomal distribution. The FOI is assumed piecewise constant with each of the following age bands: 15-24 years, 25-34 years and 35-54 years for males or 35-49 years for females. The FOI need not be piecewise constant. However, the data are sparse. Therefore, this age categorisation helps us to have an estimable monotonic FOI as there is enough data to enable us to maximise the likelihood function in order to estimate parameters. Sensitivity analyses are performed and 95 % bootstrap intervals of parameter estimates are computed. Finally , the FOI are converted to population incidence estimates using the formula $PopulationIncidence = \lambda.(1 - \pi)$. This is novel too.

3.2.2 Frailty and logistic regression models - Paper II

In this subsection, I will explain the methods used for computing time and odds ratios for Paper II and the determination of risk factors of early sexual debut.

3.2.2.1 Frailty Models

Let T denote the survival time and z denote the vector of covariates, $z = (z_1, z_2, ..., z_k)$. A multiplicative hazards rate model is:

$$h(t) = h_0(t).c(\beta^T z)$$

where

h_0(t) is an arbitrary baseline or reference hazard - the hazard rate function under the

standard condition, z = 0
β^T is the fixed effects vector
$c(\beta^T z)$ is a known function

If $c(\beta^T z) = e^{\beta^T z}$, the model becomes: $h(t) = h_0(t).exp\left(\beta^T z\right)$. This is called the Cox proportional hazards model. In some scientific experiments, survival data are collected from study units which are in groups called clusters. In such cases, the Cox proportional hazards model becomes

$$h_{ij}(t) = h_0(t).exp\left(\beta^T z_{ij}\right)$$

where
$h_{ij}(t)$ is the hazard rate function for the jth individual from the ith group,
h_0(t) is an arbitrary baseline or reference hazard - the hazard rate function under the standard condition z = 0
β^T is the fixed effects vector and
z_{ij} is the vector of covariates.

If the baseline hazard is treated nonparametrically and a parametric form is assumed for the covariates, the model is said to be semi-parametric.

Clustered survival data are best modeled by using a random effect survival model. This is an extended Cox proportional hazard model. Klein and Moeschberger (1997) define frailty as being an unobservable random effect shared by subjects within a subgroup [124]. The Cox model is extended by adding what is called a frailty. Consequently, the frailty model is called the shared frailty model. Beard (1959) originated the idea of including random effects in survival models [125]. The shared frailty model is of the form :

$$h_{ij}(t) = h_0(t).exp\left(\beta^T z_{ij} + w_i\right)$$

where
w_i is the random effect for the ith group or cluster(i = 1, 2, 3,..., G).

Note that the w_i's are are independent and identically distributed. They are from some distribution with mean 0 and variance 1. The shared frailty model is also written as $h_{ij}(t) = h_0(t).u_i.exp\left(\beta^T z_{ij}\right)$ where the $u_i's(u_i = exp(w_i))$ are an independent and identically distributed sample from some distribution with mean 1 and an unknown variance θ . The u_i is the frailty for the ith group or cluster. Common distributions proposed for the random effect are the gamma distribution with one parameter, positive stable distribution and the inverse Gaussian distributions [125]. A one parameter gamma density is written as $g(u) = \frac{\theta^{-\frac{1}{\theta}}}{\Gamma(\frac{1}{\theta})}.u^{\frac{1}{\theta}-1}.exp^{-\frac{u}{\theta}}$. When the random effects are assumed to follow a one-parameter gamma distribution, the frailty model is called the gamma frailty model.

Approaches to modeling effects of covariates on survival include:

I. Modeling the hazard or log hazard rate

II. Modeling the failure time or log failure time directly.

The semi-parametric Cox proportional hazards model uses the first approach. The accelerated failure time (AFT) model uses the second approach.

For the AFT, let T denote the survival time. By making the natural log transformation $Y_T = ln(T)$ we proceed to model Y using a linear model. The linear model is

$Y_T = \mu + \beta^T z + \sigma W$
where
β^T is the vector of regression coefficients
z is the vector of covariates
W is the error distribution and
σ is the scale parameter

This is called the loglinear model representation for the AFT. The error distribution can be the standard normal distribution, extreme value distribution or a logistic distribution. The AFT model is called the accelerated failure time model because the

covariates have the effect of accelerating or decelerating the time scale.

For the AFT, the survival, density of T and hazard functions are represented as :
$S(t) = S_0(t.exp(-\beta^T z))$
$f(t) = exp(-\beta^T z).f_0(t.exp(-\beta^T z)))$
$h(t) = exp(-\beta^T z).h_0(t.exp(-\beta^T z))$

where
$exp(-\beta^T z)$ is called the accelerating factor.

The accelerated failure time model can also be extended to accommodate a random effect term to give a frailty model thus: $Y_{ij} = \mu + \beta^T Z_{ij} + b_i + \sigma W$.
where b_i is the random effect term.
On the original scale this translates to $T_{ij} = exp(\mu + \beta^T Z_{ij} + b_i + \sigma W)$.

There are three cases concerning the value that $exp(\beta)$ can take:

I. $exp(\beta)$ is less than 1 if β is less than 0.

II. $exp(\beta)$is equal to 1 if β is equal to 0.

III. $exp(\beta)$ is greater than 1 if β is greater than 0.

Note that the model formulation

$$Y_T = \mu + \beta^T z + \sigma W$$

corresponds to the accelerating factor $e^{-\beta^T z}$. Similarly, the model formulation

$$Y_T = \mu - \beta^T z + \sigma W$$

corresponds to the accelerating factor $e^{\beta^T z}$ since the coefficient vector for the loglinear AFT model representation

$$Y_T = \mu + \beta^T z + \sigma W$$

is $\beta^T = (\beta_0, \beta_1, \beta_2, ..., \beta_k)$ and it corresponds to a coefficient vector $-\beta^T$ for the accelerating factor.

The exponential of β, $exp(\beta)$ is called the time ratio in Stata parlance. This quantity tells how fast ($exp(\beta) < 1$) or how slow ($exp(\beta) > 1$) individuals experience the event of interest in comparison with the baseline category of a covariate Z.

The time ratios are interpreted as follows: For a given level of a categorical variable, a time ratio less than 1 implies that for that category the event of interest (sexual encounter in the case of Paper II) was initiated earlier (or faster) than for the baseline category. To the contrary, a time ratio greater than 1 implies that the event was initiated later (slower) than for the baseline category. In Paper II, we fitted random effect accelerated failure time models to men's and women's data to obtain time ratios. The time-to-event variable , T, had a lognormal distribution. The frailty distribution was gamma distribution. Maximum likelihood estimation of the regression coefficients was achieved using STATA 9 [126] software. Time ratios are optional output in STATA 9.

Note that frailty models help in accounting for heterogeneity due to omitted variables. If data are from a study where study subjects were in clusters, survival times for members of the same clusters are bound to be more correlated that event times from different clusters. This heterogeneity among members of different clusters ought to be accounted for in modelling. The traditional Cox regression model does not account for heterogeneity due to omitted variables. The MDHS2010 data were collected from clusters as the data were from a multi-stage cluster sampling scheme. To account for the clustering in the data, the random effect AFT model was used in modelling. The cluster variable was Enumeration Area.

3.2.2.2 The Weighted Logistic Regression Model

In Paper II, logistic regression models were fitted in addition to the accelerated failure time models. In this subsection, I will briefly explain the theory behind the models used. Consider a random sample of n human beings on whom data on a binary categorical response Y and several explanatory variables x_1, x_2, x_3, ... , x_n are collected. The responses Y_1, Y_2, ... , Y_n are a collection of independent and identically distributed (iid) random variables. Suppose that Y assumes either 0 or 1 and Pr(Y =1)

$=\pi$. Then Y has a Bernoulli distribution with parameter π. Since π depends on the explanatory variables, it is right to express π as $\pi(x)$. Therefore, the density of Y is $P(Y_i = y_i) = \pi(x)^{y_i}\left(1-\pi(x)\right)^{1-y_i}$, $y_i = 0, 1$.

Using least squares techniques a linear probability model can be fitted thus: $\mathrm{E}(Y_i) = \pi\left(x\right)=\beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_k x_k$ where β_i is the regression coefficient for variable x_i. A linear probability model has many known problems [96]. Consequently, a better alternative to the linear probability model is the logistic regression model which is expressed as $log\left(\frac{\pi(x)}{1-\pi(\mathbf{x})}\right) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_k x_k$. From this equation, $\pi\left(\mathbf{x}\right)$ can thus be expressed as:

$\pi\left(\mathbf{x_i}\right) = Pr(Y_i = 1|x)=\frac{exp(\beta_0+\beta_1 x_{i1}+\beta_2 x_{i2}+...+\beta_k x_{ik})}{(1+exp(\beta_0+\beta_1 x_{i1}+\beta_2 x_{i2}+...+\beta_k x_{ik}))}$ where $x_i = (x_{i1},\ x_{i2},\ ...,\ x_{ik})$ - a vector of covariates.

The maximum likelihood for the data is
$L(\beta) = \prod_{i=1}^{n} \pi(x)^{y_i}(1-\pi(x))^{1-y_i}$.

The corresponding loglikelihood is $l(\beta) = \sum_{i=1}^{n}[y_i.log\left(\pi(x)\right)]+\sum_{i=1}^{n}[(1-y_i)\,.log\left(1-\pi(x)\right)]$
The values of the regression coefficients, β_j which maximize the loglikelihood function are the maximum likelihood estimates. The regression coefficients are estimated using the Newton-Raphson procedure [96].

Standard logistic regression analysis of data from a random sample can be conducted using many computer packages as observations are assumed independent and so inference is uncomplicated. However, surveys use complex sampling strategies. In a multistage sampling procedure, for example, there is intra-cluster correlation. Consequently, observations from the same cluster are correlated. Therefore, the assumption of independence between observations which is the core assumption in standard regression analyses, does not hold. In such cases, maximum likelihood methods of estimating parameters shouldn't be used. Sampling weights are the inverses of the product of the conditional inclusion probabilities at each stage of sampling in a multistage sampling. These are often computed for each observation and included in a data set collected from

a multistage sample. Under multistage sampling, pseudo-maximum likelihood methods are recommended [127]. The pseudo-maximum likelihood function has the sampling weight incorporated in it.

Sampling weights are incorporated in the pseudo-maximum likelihood function as follows. Suppose that a sample of n study subjects are from m sampled clusters. Let i index cluster (i=1,2,..., m) and j index individuals within each cluster (j = 1, 2,..., n_i). Denote the number of subjects in sampled cluster i by n_i. Suppose that data on a binary response Y_{ij} and explanatory variables x_{ij} are collected from individuals in all sampled clusters. Let w_{ij} be the sampling weight for the observation from the jth subject from the ith cluster. A single observation contributes the following to the pseudo-maximum likelihood function [127]:

$$\pi(x_{ij})^{(w_{ij})y_{ij}} \ (1-\pi(x_{ij}))^{(w_{ij})(1-y_{ij})}.$$

The pseudo-maximum likelihood function is
$L(\beta)=\prod_{i=1}^{m}\prod_{j=1}^{n_j}(\pi(x_{ij}))^{w_{ij}.y_{ij}}(1-\pi(x_{ij}))^{(w_{ij}).(1-y_{ij})}$.

The regression coefficients are estimated by finding the maximum of the pseudo-likelihood function. Note that generalization of the derivation of the pseudo-maximum likelihood function of data from a cluster sampling strategy with three or more stages easily follows from the derivations above. Paper II reports results from the analysis of data collected from a multistage cluster sample [128]. The regression analysis conducted in Paper II adopted the above approach to estimate regression coefficients and corresponding standard errors using STATA 9 software [126].

3.2.3 Prediction of the standardized incidence rates of cancer and cancer burden - Paper III

Consider age-specific cancer incidence data where age has been categorised into k groups. The empirical estimates of the age-specific crude rate of cancer incidence for group i, M_i are obtained by dividing the number of cancer cases in age group i, c_i by the corresponding person years at risk for that age group , n_i thus $M_i = \frac{c_i}{n_i}$. If data are collected

over a period of several years, the formula for computing incidence is

$M_{it} = \frac{c_{it}}{n_{it}}$

where c_{it} is the number of cancer cases in age group i and period t

n_{it} is the number of person years at risk for age group i and period t

The age-standardized incidence rate (ASR) is computed using the following formula:

$$ASR = \frac{\sum_{i=1}^{k} w_i.M_i}{\sum_{i=1}^{k} w_i}$$

where

M_i is the crude incidence rate for age group i, i = 1, 2, $\ldots, k$

w_i is the weight for age group i, i = 1, 2, $\ldots, k$

Different weights can be used. However, for the analyses in Paper III, weights from the world standard population [129, 130, 131] were used. These weights are shown in the table below.

Table 1: Segi world standard population weights

i	Agegroup	Weight		i	Agegroup	Weight
1	0-4	12000		10	45-49	6000
2	5-9	10000		11	50-54	5000
3	10 -14	9000		12	55-59	4000
4	15-19	9000		13	60-64	4000
5	20-24	8000		14	65-69	3000
6	25-29	8000		15	70-74	2000
7	30-34	6000		16	75-79	1000
8	35-39	6000		17	80-84	500
9	40-44	6000		18	85+	500

Source:Bray et al (2002)

3.2.4 Projections of cancer burden

Future standardized incidence rates can be estimated from cancer data for a particular period by fitting statistical models to predict age-standardised incidence rates. The forecasts of standardized incidence rates are then multiplied by the estimated population at risk to yield the estimates of the number of cancer cases to occur in the future. These estimates are called the future burden of cancer [132]. Firstly ,age-standardised incidence rates can be estimated using using age period cohort models(APC) [133, 134, 135]. The age period cohort model [136] is denoted by the following relation:

$M_{ap} = exp\,(A_a + D.p + P_p + C_c)$

where

M_{ap} is the incidence rate in agegroup a and period p

A is the age component for agegroup a

D is the common drift parameter

P_p is the non-linear period component for period p

C_c is the non-linear cohort component for cohort c

This is a Poisson regression model which is known to give good fits when modelling cancer rates [136]. However, this model is known to have two problems [136]. The first problem is that it can give unrealistic predicted rates which grow exponentially. The second problem is that the model can only give good extrapolations for a period in the near future than for a period in the distant future.

Secondly age-standardized incidence rates can also be estimated from cancer data for a particular period by using time-linear models. Dyba and colleagues reviewed and proposed several time-linear models for predicting future incidence rates and predicting cancer burden [97, 98]. There are two main underlying assumptions for the time-linear models that are fitted for the purpose of projecting standardized incidence rates. The first assumption is that the counts, c_i or c_{it} have a Poisson distribution. The second assumption is that the incidence rates, M_i or M_{it} have a normal distribution.

The time-linear models for the extrapolation of incidence rates are:

$$E\left(\frac{c_{it}}{n_{it}}\right) = \alpha + \beta.t \tag{1}$$

$$E\left(\frac{c_{it}}{n_{it}}\right) = \alpha_i + \beta_i.t \tag{2}$$

$$E\left(\frac{c_{it}}{n_{it}}\right) = \alpha_i + \beta.t \tag{3}$$

$$E\left(\frac{c_{it}}{n_{it}}\right) = \alpha_i.\left(1 + \beta.t\right) \tag{4}$$

$$log\left[E\left(\frac{c_{it}}{n_{it}}\right)\right] = \alpha + \beta.t \tag{5}$$

$$log\left[E\left(\frac{c_{it}}{n_{it}}\right)\right] = \alpha_i + \beta_i.t \tag{6}$$

$$log\left[E\left(\frac{c_{it}}{n_{it}}\right)\right] = \alpha_i + \beta.t \tag{7}$$

where

c_{it} is the number of cases of cancer in age group i and period t

n_{it} is the number of person years at risk in age group i and period t

α_i is the intercept for age group i

α is the common intercept for all age groups

β_i is the slope parameter for age group i

β is the common slope parameter for all age groups

Models 1, 2 ,3 and 4 are used for making predictions for increasing trends. Models 5, 6 and 7 are used for making predictions for decreasing trends. Estimates of standardized incidence rates from these models are multiplied by the population forecasts to obtain the total number of cases of cancer expected in the future. The commonest models used for extrapolating trends are Models 2, 4, 6 and 7. For the projections in Paper III, Models 2 and 3 were fitted.

Note that the models in (1) to (7) above are also Poisson regression models just like the age period cohort model. However, the models for increasing trends can never produce explosive exponential predictions. Furthermore, the models for decreasing trends can never produce negative predictions of incidence rates. For these two reasons, the time-linear models are sometimes preferred to the APC.

3.2.4.1 Estimation of standard errors of the predicted number of cases (burden)

One of the underlying assumptions of the time-linear models that are fitted for the purpose of projecting standardized incidence rates is that the counts, c_i or c_{it} have a Poisson distribution. Let c_T denote the total number of cases- sum of the c_i or c_{it} for i=1,2, ..., k. Therefore, the underlying assumption for the estimation of the standard errors of predicted burden translates to the fact that the counts c_T have a Poisson distribution with parameter θ_T. Then the variance [97] of c_T is $var\left(c_T\right) = var\left(\theta_T\right) + E\left(\theta_T\right)$. Since θ_T is an unknown parameter it is estimated from the data.

3.2.5 Analyzing the change in the annual number of cases of cancer

When forecasts of cancer burden are made, it is often interesting to try to find out potential factors that will cause a future rise or fall of the burden of cancer. In any case, population change and changes in the risk of cancer over time are some of the factors that might affect the future cancer burden. Population change refers to the change in the population age structure and population size. It is also called demographic change. In order to analyze the contribution of population changes and the change in the risk of cancer on the mean annual number of cases for the future, the difference between the

projected number of cases of cancer and the observed mean annual number of cases for a known period is decomposed using an approach by Moller et al (2002) [136].

In the spirit of Moller et al (2002), let N_{ras} be the number of cases given a cancer risk r an age structure a and a population size s where r, a and s are levels of the period, *0* for which we have cancer incidence data (the base of prediction) or the future period f for which cancer incidence estimates are sought. In this context, N_{000} is the expected number of cases applying the rates, age structure and population size of period o. Similarly, N_{fff} is the predicted number of cases for period f. The quantity $\Delta_{tot} = N_{fff} - N_{000}$ is the total change in the annual number of cases and can be decomposed as follows:

$\Delta_{tot} = N_{fff} - N_{0ff} + N_{0ff} - N_{000}$

$= \Delta_{risk} + \Delta_{pop}$

where

N_{0ff} is the expected future number of cases when the present rates and age structure in period *0* are applied

Δ_{risk} is the change in the annual number of cases due to change in cancer risk

Δ_{pop} is the change in the annual number of cases due to changes in age structure and population size

The change in the mean number of cases due to changes in age structure and population size can also be decomposed into two components, one due to the change in age distribution, the second due to changes in population size:

$\Delta_{pop} = N_{0ff} - N_{000} = N_{0ff} - N_{00f} + N_{00f} - N_{000}$

$\Delta_{pop} = \Delta_{age} + \Delta_{size}$

where N_{00f} is the number of cases of cancer obtained by multiplying projected rates of future period f by the age structure and population size of period *0*.

4 Summary of results

4.1 Paper I: Estimation of HIV incidence in Malawi from cross-sectional population-based sero-prevalence data

The aim of the study was to develop a new method for estimating HIV incidence from cross-sectional sero-prevalence data.

We successfully devised a simple model in the form of a recurrence relation which aids in the estimation of HIV incidence for selected age ranges and prevalence for distinct ages. We fitted models for males and females separately. The Pearson Chi-square goodness of fit statistics of these separate analyses showed acceptable fit for both males ($p = 0.930$) and females ($p = 0.564$). Sensitivity analyses showed that for each of the eight sets of the mortality rates investigated in the sensitivity analysis, when the mortality rates were decreased or increased by a certain percentage, the FOI estimates also decreased or increased accordingly.

Our method is best suited for areas or countries where mortality data are available, but access to anti-retroviral therapy is a problem. The method can be easily extended to be used in a situation where anti-retroviral therapy (ART) is available.

The new method was applied to HIV sero-prevalence data from the Malawi Demographic and Health Survey 2010 (MDHS2010). Among males, the estimates of the FOI of HIV were 0.0063 (95% CI:0.0039-0.0089) for the 15-24 years age group, 0.0312 (95% CI:0.0225-0.0366) for the 25-34 years age group and 0.0158 (95% CI:0.0143-0.0281) for

the 35-54 years age group. Among females, the estimates of the FOI of HIV were 0.0225 (95% CI:0.0186-0.0263) for the 15-24 years age group, 0.0212 (95% CI:0.0140-0.0284) for the 25-34 years age group and 0.0211 (95% CI:0.0133-0.0289) for the 35-49 years age group.

The estimates of the prevalence of HIV at 15 years were 0.0035 and 0.0069 for males and females respectively.

Among males, the estimates of population incidence of HIV per 100,000 person-years were 616 (95% CI: 375- 870) for the 15-24 age group, 2705 (95% CI:1946-3172) for the 25-34 group and 1336 (95% CI:1209-2374) for the 35-54 age group. For females, the estimates were 2031 (95 % CI: 1686 -2377) for the 15-24 age group, 1712 (95% CI: 1131-2294) for the 25-34 age group and 1732 (95% CI: 1090-2375) for the 35-49 age group.

The above age-sex specific HIV incidence estimates are higher among women than among men in the 15-24 age range (2031 vs 616), lower among women than among men in the 25-34 age range(2705 vs 1712) and higher among women in the 35-49 age range than among men in the 35-54 age range(1336 vs 1732). The lowest incidence for men is for the 15-24 age range and for women for the 35-49 age range. The lowest prevalence at age 15 is for males and the highest at the same age is for females.

4.2 Paper II: Risk factors of early sexual debut among men and women - a strong predictor of HIV and sexual risk in Malawi.

The aim of this study was to determine the risk factors of early sexual debut.

The significant risk factors of early sexual debut among men were secondary education (AOR=0.577, 95 % CI : 0.394 - 0.846), Lomwe ethnicity (AOR = 1.671, 95% CI : 1.274 - 2.190), Islam (AOR = 1.560, 95% CI : 1.065 - 2.284), being married (AOR = 0.418, 95% CI : 0.324 - 0.538), being in a cohabiting relationship (AOR = 0.416, 95% CI : 0.294 - 0.589), being a widower (AOR = 0.104, 95% CI : 0.013 - 0.818) and the following birth cohorts :1965-1969 (AOR = 2.020, 95% CI : 1.273 - 3.301), 1980-1984

(AOR = 1.808, 95% CI : 1.134-2.883), 1985-1989 (AOR = 1.762, 95% CI : 1.089 - 2.852) and 1990-1995 (AOR = 4.757, 95% CI : 2.912 - 7.771).

There was a decreasing linear trend in the adjusted odds ratios with increasing level of education. The odds ratio for primary education was 1.151. For secondary education the odds ratio was 0.577.

There was an upward parabolic trend in the adjusted odds ratios for cohorts 1965-1969, 1980-1984, 1985-1989 and 1990-1995. The odds ratios were decreasing from birth cohort 1965-169 to 1975-1979. The odds ratios started to increase from birth cohort 1980-1984 to birth cohort 1990-1995.

All significant time ratios among men were less than 1. The following dummy variables had significant adjusted time ratios: central and southern province, primary education, Ngoni ethnicity, other Christian religions, Islam, other religions, all birth cohorts: (1965-1969, 1970-1974, 1975-199. 1980-1984, 1985-1989, 1990-1995) and being married, divorced or separated.

The following dummy variables had both significant time and odds ratios: Lomwe ethnicity, being a muslim, the following birth cohorts :1968-1969,1980-1984, 1985-1989, 1990-1995 and being married.

Among women, the significant risk factors of early sexual debut among women were primary education (AOR = 0.540, 95% CI : 0.435 - 0.670), secondary or higher education (AOR = 0.200, 95% CI : 0.135 - 0.297), Lomwe ethnicity (AOR = 1.700, 95% CI : 1.250 - 2.313), other ethnic groups (AOR = 1.445, 95% CI : 1.054 - 1.980), other Christian denominations (AOR = 1.588, 95% CI : 1.275 - 1.977), Islam (AOR = 1.591, 95% CI : 1.074 - 2.357), being married (AOR = 0.582. 95%CI : 0.418 - 0.809), being separated (AOR = 0.522, 95%CI : 0.330 - 0.827), and consensual sex (AOR = 0.833, 95% CI : 0.700 - 0.992).

There was a decreasing trend in the adjusted odds ratios with increasing level of education. The odds ratio for primary education was 0.540. For secondary education the odds ratio was 0.200. The last birth cohort had a borderline p-value(p=0.045).

Central province, primary and secondary education had significant time ratios which were greater than 1. However, the time ratios for the following variables were significant but less than 1: Lomwe ethnicity,Yao ethnicity, Ngoni ethnicity, Other ethnicity, other Christian religions, being a Muslim, birth cohorts 1985-1989 and 1990-1995, being married,cohabiting,widowed,divorced and being separated.

Primary and secondary education had significant odds ratios less than 1 but significant time ratios greater than 1. Islam, other Christian religions, Lomwe ethnicity and other ethnicities had significant odds ratios greater than one. However, the corresponding time ratios were less than 1 and significant. Being married and being separated had both significant odds and time ratios which were less than 1.

When other covariates were added to the univariate logistic regression models, the parameter estimates changed. For the models fitted to males' data, the parameter estimates for province, education, ethnicity, birth and marital status changed by 19% to 59%. The coefficients of place of residence and religion changed by 2% to 8%. For the models fitted to females data, the parameter estimates for Residence, province, religion, education, ethnicity, birth and marital status increased by 12% to 35%.

4.3 Paper III: Cancer incidence in Malawi: Time trends in Blantyre 1996-2005 and predictions up to 2015

The aims of this study were to describe the patterns of occurrence of cancer during 1996-2005 and to predict the incidence and total burden of cancer for 2015 in Blantyre in Malawi.

According to our analyses, the mean annual number of cases of cancer diagnosed in Blantyre during 1996-2005 was 792. Out of the 792 cases, 410 (51.8%) occurred in men and 382 (48.2%) in women. Almost half of the cancers among men were Kaposi sarcoma. The highest percentage (27.5%) of cases among women was cervical cancer. AIDS-defining cancers comprised 59.5% of the male cancer burden and 63.4% of the female cancer burden.

Among men, the age-standardized (world) incidence rates of cancer were 50.5 per 100,000 for Kaposi sarcoma, 22.3 per 100 000 for oesophageal cancer, 6.6 per 100 000 for non-Hodgkin lymphoma, 4.4 per 100 000 for eye cancer and 56.7 per 100,000 for all other cancers combined.

Among women, the age-standardized (world) incidence rates of cancer were 26.4 per 100 000 for Kaposi sarcoma, 49.3 per 100,000 for cervical cancer, 14.6 per 100 000 for oesophageal cancer, 5.3 per 100 000 for non-Hodgkin lymphoma, 5.4 per 100 000 for eye cancer, 11.9 per 100 000 for breast cancer and 42.7 per 100,000 for all other cancers combined.

The estimated total number of cases assuming that the rates observed in 1996-2005 would continue to 2015 is 1240. 618 (49.8%) are predicted to occur among women and 622 (50.2%) among men. The estimated burden of cancer for 2015 based on model-based trends is 2370 cases. Of these, 1147 (48.4%) will occur among men and 1223 (51.6%) among women. The numbers of non-Hodgkin lymphoma, eye cancer and other cancers combined among men are projected to have the greatest increase of over 200% between

1996-2005 and 2015.

Among women, the numbers of non-Hodgkin lymphoma, cervical cancer, eye and other cancers combined among men are projected to have the greatest increase of over 200% between 1996-2005 and 2015. Thus the lower limit of the projected cancer burden for 2015 is 1240 and the upper limit is 2370.

Age-incidence curves showed that for Kaposi sarcoma and cancer of the eye, incidence was greatest in the age group 35-44 in both sexes. For oesophageal cancer (men and women), breast (women) and cervical cancer (women) and other cancers (men and women), the incidence continued to rise with increasing age. The cancer-specific trends were generally increasing in both sexes between 1996 and 2005, although there was considerable year-on-year random variation evident in the trends.

In men, the age-standardized incidence rates of Kaposi sarcoma, non-Hodgkin lymphoma, oesophageal and eye cancer are predicted to increase. Similarly, in women the age-standardized incidence rates of non-Hodgkin lymphoma, oesophageal and eye cancer are predicted to increase. However, the age standardized incidence of Kaposi sarcoma in women is predicted to be constant up to 2015. In both sexes, the incidence rates of all other cancers are also predicted to increase. There were large year-to-year variations in the 1996-2005 trends of the age-standardized incidence rate of Kaposi sarcoma in men and other cancers but these are unlikely to cause similar changes in incidence.

5 Discussion

We aimed at developing a new method for estimating HIV incidence from cross-sectional sero-prevalence data accounting for differential mortality and to apply the method on data from Malawi, to determine the risk factors of early sexual debut, to describe the patterns of occurrence of Kaposi sarcoma, non-Hodgkin lymphoma, esophageal cancer, eye cancer, breast cancer, cervical cancer and all other cancers combined among the population of the Blantyre district of Malawi, and to predict the burden of cancer for 2015 on the basis of projected population changes. This section contains the methodological considerations and discussion of results of each of the papers included in this thesis.

5.1 Methodological Considerations

5.1.1 Estimation of HIV incidence from sero-prevalence data-Paper I

Klaus Langohr's review gives an overview of different methods for computing incidence from prevalence data [51]. Our model is also used for the same purpose but seems to overcome several of the limitations of previous approaches. Leske et al (1981)'s method assumes non-differential mortality, lifelong duration of the disease and stable disease in a stable population [94]. The model uses grouped prevalence data to estimate incidence. This model is not appropriate for AIDS because HIV infection affects mortality. The rates of mortality due to HIV are usually higher than natural mortality rates. Grenfell and Anderson (1985)'s model [77] also uses this same assumption which holds for measles but does not hold for HIV infection. Podgor and Leske (1986)'s model estimates incidence using grouped data and using the assumption of differential mortality [120].

This model was validated by Saidel et al (1996) [49]. The rate of HIV infection is assumed constant within age bands. The rates of natural and HIV mortality are also assumed constant within age bands. The time to death and time to infection within age bands are assumed to be exponentially distributed. One serious drawback is that the model can produce negative estimates of incidence if the prevalence at older ages is strikingly less than than the prevalence at younger ages [49]. Besides, while this method may be appropriate for estimating HIV incidence, given grouped age-specific sero-prevalence data, the incidence for the last age-group is never estimable directly. To illustrate this point, consider the data which Podgor and Leske (1986) used for testing their model. The data consist of grouped estimates of the prevalence of senile cataracts. Age is grouped into 53-57, 58-62, 63-67, 68-72, 73-77, 78-82. The data are reproduced in Table 2 below.

Table 2: Age-specific prevalence data of senile cataracts

Age group	N	Senile cataracts(%)
53-57	480	0.0123
58-62	583	0.0281
63-67	467	0.0469
68-72	316	0.1048
73-77	256	0.138
78-82	123	0.3242

Source: Podgor and Leske(1986)

Using their method, Podgor and Leske (1986) estimated the incidence of senile cataracts for all age groups except the last age group, 78-82. The fact is that structurally it is impossible to estimate the incidence of the last age group directly. Another drawback is that the first or last age group can be open-ended thereby making estimation of incidence difficult. Another extension of Podgor and Leske (1986)'s model is Misiri (2014)'s model [118]. This method also adjusts for differential mortality and uses grouped sero-prevalence data. The provision of ART is also adjusted for in the model. However, it has the same drawbacks as Podgor and Leske (1986)'s model. Rajan and Sokal

(2010)'s method is an extended version of Leske et al (1981)'s method. This method has the same drawback as Podgor and Leske (1986)'s method in that given grouped data, the incidence can never be computed for the last age group. It is also based on several assumptions some of which are about HIV and natural mortality rates. These assumptions about mortalities are imprecise and dubious. Another method by Gregson et al(1996) [137] adjusts for mortality from infection but natural mortality is never adjusted for. Besides extra assumptions meant to accommodate the incorporation of survival analysis techniques are needed. Their model is too mathematically complicated to be understood by the average reader. Our model adjusts for both natural and HIV mortalities and is simple.

Williams et al (2001) proposed a mathematical method for estimating incidence from prevalence data from a single cross-sectional survey combined with the growth rate of an HIV epidemic [91]. Our method is simpler than theirs. This approach cannot be used when growth rates are not available. We also note that when such estimates are available, they will carry additional uncertainty with them, which can render incidence prediction more uncertain. In our model, we use natural and HIV mortalities as well as age-specific sero-prevalence data. In summary, whilst other methods for estimating HIV incidence from sero-prevalence data exist, our approach allows us to estimate incidence and prevalence at the same time, makes less assumptions, uses data which are easier to obtain and has several advantages as mentioned above.

Our method can be used in areas where ART provision is non-existent or coverage is incomplete. This type of setting is not an artifact. There are many countries out there with very low ART coverage or no coverage at all [18]. In fact, by 2014, in about 80 % of the countries with data on provision of ART, coverage was below 50% [18]. This means that for very big countries with low ART coverage, it is still possible to find regions within such countries where ART provision is non-existent. In fact, as long as the world has poor countries, there will always be countries or regions where ART coverage is incomplete. Provision of ART is done within the framework of a functioning health system with an ideal number of health facilities and a good road network. Western countries are wealthier than most developing countries so they have functioning health

systems with good numbers of health facilities and the necessary resources to enable them to roll out the provision of ARVs to those who need them. In poor countries, the situation is different. Most countries like Malawi rely on donor support. Funds that countries are given are usually spent on essential services which do not always include the purchase of drugs for treating HIV-positives. In other words, because of the scarcity of resources, the priority in resource allocation may not be the purchase of medication for treating HIV patients. Therefore it is still possible to find countries with some regions with no ART provision.

In low resource settings, health facilities are rare and often widely scattered. The availability of skilled health workers is also a challenge. Coupled with lack of transport for reaching out to remote places without health facilities, distribution of ARVs to those in need can be a big problem. Consequently, coverage of the provision of ART in many low resource settings may be incomplete. In such places where ARVs are not provided to the sick, the model can still be used to estimate HIV incidence.

Our model can be extended to accommodate other scenarios like a situation where HIV incidence estimates are needed in an area where HIV positives are on ART. In fact, currently, we are working on extending our model to accommodate other scenarios like in the situation where HIV incidence estimates are needed in an area where HIV positives are on ART. A working document is currently available, Misiri (2014), which extends Podgor and Leske (1986) adjusting for provision of ART [118], and a paper will be submitted in the near future.

Our model is more realistic than the model of Grenfell and Anderson (1985) and all other catalytic models because it adjusts for both natural and HIV mortalities. HIV greatly increases mortality rates. HIV mortality rates are often higher than natural mortality rates. If HIV positive individual persons die, they no longer have the capacity to infect others. The proportion of HIV positive persons is reduced in the population. On the other hand deaths of HIV negative persons reduce the proportion of susceptibles. Both deaths affect a sexual network. HIV is spread within a sexual network comprising those who are HIV positive and those who are HIV negative.

The underlying assumptions on which our model is based are the steady state assumption(endemic equilibrium of HIV), time homogeneity of prevalence and incidence, piecewise constancy of the FOI, differential mortality and irreversibility of disease or infection.

The assumption of endemic equilibrium is tenable as HIV has been in the Malawi population since 1985. In fact, HIV is endemic in Malawi [138, 139]. However, it is not possible to test this assumption using data from a single cross-sectional sero-prevalence sample. The assumption of time homogeneity of prevalence and incidence is untestable from our data. However, this is a necessary simplifying assumption. Given data from a single cross sectional sample, it is difficult to separate the effects of age and time. In that case, the only option left is to assume either age-dependency or time-dependency but not both. Some have succeeded in formulating models for estimating age and time dependent incidence [140]. However, such examples are very few. For infectious diseases, the reality is that the rate of infection is usually both age and time dependent [140]. The assumption of time homogeneity is also implicit in the models of Podgor et al (1986), Leske et al (1981) and Rajan and Sokal (2010). The assumption of piecewise constancy of the FOI is ideal and helpful as it helps us to have a monotonic FOI. When we categorise age into age bands and assume the FOI to be constant within these age ranges, we have enough data in each age band to enable us to achieve convergence during maximisation of the likelihood function. This is important when the data are sparse. However, in real life situations, a very wide age range can comprise several cohorts with different rates of infection. In that case, the assumption of constancy of the FOI is unrealistic. We assumed the FOI of HIV to be constant within the age ranges 15-24, 25-34 and 35-54 (males) or 35-49 (females). Our choice of age categorisation was guided by facts. The age range 15-24 comprises adolescents who are usually sexually active. The age range 25-34 is associated with a very high HIV prevalence in Malawi. Having determined the first two age categories, the last category for each sex is easy to fix. Our age categorisation for the piecewise constancy of the FOI has a sound basis as it is based on knowledge of what happens in real life situations. The assumption of differential mortality is valid. In any setting where there are both HIV positive and HIV negative persons, mortality due to HIV is usually higher than natural mortality.

The implicit assumption of irreversible infection is valid. Once a person is infected with HIV, he or she can never be cured.

The assumptions of time homogeneity and piecewise constancy of mortalities were implicit in the presentations of the mortalities from the source documents. The natural and HIV mortalities were grouped. The age groups were 15-19, 20-24, 25-29, 30-34, 35-39, 40-44, 45-49 and 50-54. The assumptions are untestable. Both assumptions can be said to be valid because the widths of the age groups are narrow (5 years). In such a narrow width, rates can be expected to be constant and independent of time. The configuration of the mortality data is not a necessary requirement for fitting the model. Even if we had single age mortality rates, it would have been possible to fit our model.

In Paper I, we give estimates of standard errors of parameters obtained through bootstrapping. Although the model was fitted in R, it can be fitted on any platform apart from R where optimization is possible. Examples of such software are STATA, SAS and SPLUS. R has several optimizers one of which is the function **nlm**. This function has a routine for computing the Hessian matrix of parameters during maximization. The inverse of the Hessian matrix is the variance-covariance matrix of the parameter estimates. Therefore, on the diagonal of the inverted Hessian matrix are the standard errors of parameter estimates. However, in general, in other software the availability of inbuilt routines for computing the Hessian matrix for parameter estimates during optimization can never be guaranteed. The standard errors obtained from **nlm** are maximum likelihood estimates. Maximum likelihood estimates have well-known nice statistical properties. In the paper, we only present bootstrap standard errors because given the option of choosing between using bootstrapping and maximum likelihood estimation to estimate standard errors, the average data analyst would opt for maximum likelihood because it is the popular option. Our model was fitted in R by writing some code which was then run. In order to be independent of our R implementation, we decided to present in our paper the bootstrap standard errors only. In any case, the main objective of estimating standard errors is to provide an estimate of uncertainty and to provide an idea of the precision of estimation for a model. So if two methods give similar estimates, it's a confirmation that estimation of standard errors is being

done correctly. The estimates of standard errors of parameter estimates from maximum likelihood and bootstrapping were similar.

In Table 2 of Paper I, there is a misprint in relation to the way the columns containing the natural and HIV mortality rates are labelled. The third and fourth columns should be labelled correctly as "Natural mortality rates". The fifth column should be labelled correctly as "HIV mortality rates".

5.1.1.1 Limitations in Paper I

There are several limitations of the mathematical model in Paper I. The steady state assumption is one of them. The assumption restricts the use of our model to populations where the spread of HIV is in endemic equilibrium. Although HIV is prevalent in many parts of the world, the extent of the spread of HIV varies from country to country such that HIV is not necessarily endemic in all countries of the world where it is prevalent.

Another limitation which might have affected our HIV incidence estimates is the assumption of time-homogeneity. HIV is spread through hetero-sexual contact. The rate of contact between the infected and the HIV negatives is an important factor in the spread of HIV as it affects the FOI. The influence of time on the contact rate can never be completely ruled out. For instance, during religious, political, sports or arts festivals and some popular holidays (for instance New Year's Day) , the contact rate between HIV positive and HIV negative persons tends to be high. In such cases, the rate of HV infection can increase. It is worth noting that these festivities are scattered unevenly in the year. Therefore, the rate of HIV infection in a year may depend on the actual time of the year apart from the age of members of the population.

In a cross-sectional sero-prevalence sample, it is not possible to disentangle the effects of time and age when estimating incidence of infection from current status data [47]. Therefore, estimation of incidence is accomplished by assuming, plausibly, that time-dependence does not exist. This assumption is necessary. Note that it is still possible to formulate a mathematical model for estimating incidence from current status data assuming both time and age dependency. Ades et al (1993) devised a model for esti-

mating incidence of toxoplasmosis from cross sectional prevalence data assuming time and age dependency [140]. Using their model, Ades et al (1993) demonstrated that ignoring time-dependence in the estimation of the incidence of toxoplasmosis introduces bias in the estimates [140]. Apart from Ades et al (1993), Keiding et al (1996) also proposed a mathematical model for the relationship between age and time dependent incidence, prevalence and duration [83]. This model is a potential building block for a model for estimating age and time dependent incidence from cross-sectional data. The time-homogeneity assumption is popular and has been used by Podgor and Leske (1986), Grenfell and Anderson (1985) and Leske et al (1986). It is fairly easy to formulate a model for estimating incidence from cross sectional data by using the simplifying assumption of age-dependence only.

The piecewise nature of the FOI are also a limitation of our model. In a narrow age-band, it is plausible to assume the constancy of the rate of different events like HIV infection, natural and HIV-associated mortality because not much happens in a narrow time period. However, in a wide age band, constancy of the rate of occurrence of different events is not necessarily guaranteed because of the possibility of having mixed age-cohorts with different rates of HIV infection. Since sexual practices differ across cohorts, the rate of HIV infection can not be safely assumed to be the same for different age cohorts within the same age band. For instance, a 15-34 year age band includes the cohorts of teenagers aged 15-19 years, young adults aged 20-29 years, and adults aged 30-34 years all whose levels of sexual activity and sexual practices may not be similar. The rates of HIV infection may also not be similar. This implies that even though our categorisation of age into age groups in which the FOI was constant was based on facts, the widths of the age groups may have been a limitation.

Another limitation of the model is that it can only produce reliable estimates of HIV incidence given good sources of mortality rates. Therefore, to get the best results from the model, reliable estimates of natural and HIV mortality rates are needed.

5.1.2 Risk factors of early sexual debut among men and women - a strong predictor of HIV and sexual risk in Malawi.- Paper II

Exploratory analyses were performed for Paper II. The analyses included the model-building process. Using the likelihood ratio test, variables were added one at a time in the model. For each variable added to a model, the reduction in deviance was noted. Variables which when added to a model resulted in a significantly big reduction in deviance were retained in the model. At the end of the model-building process, for each gender, there were final survival and logistic regression models.

The explanatory variables for the final survival and logistic regression models in Paper II were all categorical. These are listed below with their respective categories or levels:

Ethnicity (1 = Chewa, 2 = Tumbuka, 3 = Lomwe, 4 = Yao, 5 = Sena, 6 = Ngoni, 7 = Other ethnic groups)

Region (1 = North, 2 = Centre, 3 = South)

Place of residence (1 = Urban, 2 = Rural),

Religion (1 = Catholic, 2 = CCAP, 3 = SDA, 4 = Other Christian, 5 = Moslems, 6 = Other)

Education level (1 = No education, 2 = Primary education, 3 = Secondary or higher education).

Whether first sex was consensual (1=No, 2=Yes)

Birth cohort
For males (1 = 1955-1959, 2 = 1960-1964, 3 = 1965-1969, 4 = 1970-1974, 5 = 1975-1979,

6 = 1980-1984, 7 = 1985- 1989, 8 = 1990-1995)

For females (1 = 1960-1964, 2 = 1965-1969, 3 = 1970-1974, 4 = 1975-1979, 6 = 1980-1984, 7 = 1985-1989, 8 = 1990-1995)

Marital status (1=Never married, 2=Married, 3=Cohabiting, 4=Widowed, 5=Divorced, 6=Separated)

The categorical variables have different reference levels. The reference levels were chosen for different reasons. The reference category for ethnicity is Chewa. The Chichewa ethnic group was chosen since it is well known in Malawi because of the Chichewa language and Chewa traditional dances. Chichewa is widely spoken in all provinces of Malawi. Besides, for many years, senior secondary school students were taught the culture of the Chewa ethnic group from a book titled "Kukula ndi Mwambo" as part of Chichewa literature. The material was examinable. Consequently the Chewa are a well-known ethnic group. It is better to choose a well-known ethnic group than to choose a lesser known ethnic group in Malawi as the reference. The Northern Province was chosen as the reference category for Province because it stands out due to its culture which differs substantially from the culture prevalent in the Central and Southern Provinces. Besides, it is the least developed of all provinces in Malawi. Urban residence was chosen as the reference category for place of residence because of its unique characteristics. Urban areas are associated with affluence, advanced levels of development and high economic status. The rural areas are relatively poor and less developed. Most people live in the rural areas. The reference category for religion is the Catholic church. This is the denomination which has the biggest number of members in Malawi. Besides, it is the most popular Christian denomination in Malawi. This is why it was chosen as the reference category for religion. The reference category for level of education was No education. This was chosen as the reference because in Malawi the literacy rate is low. Consequently, a reasonable proportion of the population is uneducated. In addition to this, people with some education are expected to know more about the world around them and to be easily teachable. By default, people have high expectations from the educated than from the uneducated. Therefore educated people are a distinct stratum

in society. The reference category for whether first sex was consensual or not was "No, first sex was not consensual". Consensual sex is a normal occurrence as two individuals have to agree to engage in a sexual act without using force or coercive tactics. On the other hand forced or coercive sex is considered unusual and by setting "No, first sex was not consensual" as the reference category, the odds of early sexual debut among those whose first sexual encounter was consensual can then be interpreted easily in comparison to the odds of early sexual debut among those whose first sexual encounter was not consensual. For marital status, the reference category was the Never married category. Marriage is an institution full of experiences. Married people go through many challenging and interesting situations which shape their understanding of the world around them, interpersonal skills and how to handle different situations. Marriage helps in honing people's abilities in handling different situations. For these reasons, married people or those who have at least been married before are a different stratum in society when compared to those who have never been married before. It therefore makes more sense to use the Never married category as the reference in order to facilitate comparisons of odds of early sexual debut. The first level for birth cohort among males was 1955-1959. Among females, the first level for birth cohort was 1960-1964. The choice of the first birth cohort was motivated by the amount of available data. For instance, the earliest year of birth among males was 1955. Among females the earliest year of birth was 1960. The first birth cohorts among males and females were chosen as reference categories in order to make comparisons of odds of early sexual debut easy and to be able to detect changes in sexual habits through cohorts. Government policies change with time. Culture and customs also change with time due to different reasons. Culture may be modified to suit modern understanding and events. For example, in many ethnic groups in Malawi, according to old customs when a married man died, his brother inherited his wife. However, nowadays because of the AIDS pandemic, many people have contracted HIV and eventually died of AIDS because of following this custom. Consequently, this custom has been discouraged and is being discontinued by many ethnic groups. Setting the first birth cohort of each gender as the reference category makes comparisons of odds of early sexual debut for different cohorts easy because the odds of early sexual debut for each birth cohort are compared to the odds of early sexual debut for the first cohort. In that way, changes in sexual habits through birth cohorts can be detected.

Following the choice of reference categories, the above explanatory variables were coded again to facilitate modeling in STATA. From each categorical variable, several indicator variables were constructed using dummy variable coding. To perform regression analyses in STATA with dichotomous or polytomous categorical explanatory variables so that the results are easy to interpret, the qualitative covariates must be changed into indicator variables. If a categorical variable is coded such that its categories are assigned codes in ascending or descending order then STATA will take the variable as quantitative. It is then used in fitting a model as it is. Thus for each qualitative variable, there is one corresponding parameter estimate. This makes interpretation of parameter estimates difficult. To recode a qualitative variable into indicator variables we proceed as follows. Consider a generic qualitative variable X with k categories. Using dummy coding, such a variable will be coded into a dummy vector with q= k-1 components. If the variable is X, the new components of the dummy vector after coding will be $X^{(1)}$, $X^{(2)}$, ..., $X^{(q)}$. [91]. If the values 0 and 1 are used in the coding then the jth component of the dummy vector will assume 0 or 1 as follows:

$X^{(j)} = 1$ if category j is observed
$X^{(j)} = 0$ if otherwise
for j = 1, 2, ... , q

The choice of reference category affects the way results from the analysis of data will be interpreted. If the kth category is the reference category then X will be a vector of zeros. For instance, consider Province which has three categories and is coded thus: 1 = North, 2 = Centre, 3 = South. If we set the North to be the reference category, then using dummy coding, Province becomes

$X^{(2)} = 1$ if category 2 (Centre) is observed
$X^{(2)} = 0$ if otherwise

$X^{(3)} = 1$ if category 3 (South) is observed
$X^{(3)} = 0$ if otherwise

When the first category (North) is observed, X becomes the zero vector because the first category is the reference category.

The explanatory variables in Paper II are a mixture of ordinal and nominal variables. Ordinal variables are education and birth cohort. Nominal variables are residence, province, marital status, religion, ethnicity and whether first sex was consensual or not. When an ordinal variable is included in a regression model as a covariate, it can be treated as nominal or ordinal. If it is treated as ordinal, scores are assigned to the various categories of the variable in ascending or descending order according to context. When the regression model is fitted, the effects of that ordinal variable are given as one parameter estimate. However, if the ordinal variable is treated as a nominal variable, it can be recoded into indicator variables and interpretation is usually based on odds ratios. In that case, interpretation is easy. In Paper II, we chose the second option because odds ratios are easy for readers to understand. That is why both birth cohort and education were treated as nominal. Nevertheless, the concept of trend can still be captured by the linearity of the odds ratios of categories bar the reference category.

Paper II has unadjusted and adjusted analyses resulting in many p-values. Therefore, there is the problem of multiple testing. Where there is multiple testing, there is the possibility of incorrectly rejecting the null hypothesis when there is no real effect. The false discovery rate (FDR) is the expected proportion of erroneous rejections among all rejections [141]. There is a method for correcting for multiple testing called the Benjamini-Hochberg (BH) procedure [141]. The Benjamini-Hochberg procedure corrects for the family wise detection error rate. The **stats** libraries of R-3.2.1 and R-3.2.3 have a function for performing corrections for multiple testing called **p.adjust**. The input to **p.adjust** are p-values from a multiple testing scenario and the output are p-values corrected for multiple testing.

When all the p-values in Paper II are corrected for multiple testing using a false discovery rate of 5%, all previously adjusted results (time ratios and odds ratios) which were not significant are still not significant. However, among men the following indicator

variables which were significant in the multivariable logistic regression model are now not significant: being a widower (corrected p-value = 0.061) and being a member of other Christian religions (corrected p-value = 0.049). Among women, after correcting for multiple testing, the following indicator variables which were significant in the multivariable logistic regression model are no longer significant: Whether first sex was consensual (corrected p-value = 0.078), birth cohort 1990-1995 (corrected p-value =0.083) and other Christian religions (corrected p-value = 0.051). For the crude analyses, the following indicator variables which were significant among men in the univariate logistic regression models are no longer significant: Sena ethnicity (corrected p-value = 0.078). Among women after correcting for multiple testing, the following unadjusted results in the univariate logistic regression modeling were significant but are now not significant: Central Province (corrected p-value = 0.052) and birth cohort 1985-1989 (corrected p-value = 0.083). Furthermore, the following indicator variables which were significant in the univariate accelerated failure time models are not significant: Tumbuka ethnicity (corrected p-value = 0.074) and 1985-1989 birth cohort (corrected p-value = 0.083). In summary, FDR , therefore, indicates that several of the findings could well be false discoveries, due to multiple testing. However, several of our findings remain significant also after FDR correction. In particular the most important effects which we found are still significant.

To control for confounding in the analysis, all covariates were included in the regression and survival analysis models. For the models fitted to males' data, the parameter estimates for province, education, ethnicity, birth and marital status have changed by about 19% to 59%. The coefficients of place of residence and religion have changed by about 2% to 8%. For the models fitted to females' data, the parameter estimates for Residence, province, religion, education, ethnicity, birth and marital status have changed by about 12% to 35%. The magnitude of these changes in parameter estimates point to presence of confounding effects which were eventually controlled for in the adjusted analyses.

The methods used in Paper II are conventional statistical approaches used by many in research. In the paper we give estimates of time ratios. Mkandawire et. al (2012) also used time ratios to analyse early sexual debut in the context of orphanhood status [102].

Whereas we analyzed nationally representative data, Mkandawire et. al (2012)'s data were from the Northern Province of Malawi.

The discussion of results of Paper II in Subsection 5.2.2 is based on the interpretation of adjusted odds ratios. We adopted the interpretation of adjusted odds ratios from logistic regression suggested by Hosmer and Lemeshow (1989) [142]. About this, on pages 41-42, Hosmer and Lemeshow (1989) say that "The odds ratio is a measure of association which has found wide use, especially in epidemiology, as it approximates how much more likely or unlikely it is for the outcome to be present among those with x = 1 than among those with x = 0."

5.1.2.1 Limitations in Paper II

There are several limitations in Paper II. The main limitation is the data analysed. Firstly, the data on first sex were collected from survey respondents through recall. The respondents were aged between 15 and 54 years at most. The wide age-range implies that the variations in ages of respondents is quite big. This means that some respondents were interviewed at relatively old ages. In such cases, there is the problem of memory lapse. Therefore, recall bias could be a potential limitation. Probably, because of recall bias, our estimates of odds and time ratios may have been biased. Furthermore, the data were collected after some persons had already had sexual debut. Had the data been collected prospectively, it would have been easy to know the influence of some modifiable factors on sexual debut. Religious affiliation, province and place of residence , for example, can change with time. Therefore, it is quite possible that by the time some respondents of the MDHS2010 survey were interviewed, they might have changed religious affiliation, province of residence and place of residence several times. In that case, the real effects of religion, province of residence and place of residence can not be estimated accurately.

Secondly, the time-to-first sex data are discrete-survival time data. However, we used survival analysis techniques (accelerated failure time (AFT) frailty models) for continuous time-to- event data to estimate time ratios. Our use of the AFT models was justified. Firstly, Atuhaire (2011) also used continuous time survival techniques to analyse current

status data on sexual debut which were discrete [143]. Secondly, the age-specific FOI can be computed from sero-prevalence data which is current status data. The ages are often discrete. Thirdly, on pages 41 and 81, Hens et al (2012) equate the age-specific FOI to the hazard function of continuous event-time survival analysis [85]. So the FOI estimated from current status data is the hazard function which we estimate from continuous time survival data. Fourthly, Sun (2006) presents and illustrates the analysis of discrete event times from current status data using proportional hazards models (on page 99), proportional odds models (on page 107) and additive hazards models (on page 112) [144]. All these statistical models are usually fitted to continuous event-time survival data. Lastly, Lindsey (1996) , on page 26, clearly states that "Empirically, any random variable, Y, can only take discrete values defined by the unit of measurement". This statement implies that when we take measurements of Y using some unit of measurement Δ and get y_i, the reality is that $y_i - \frac{\Delta}{2} \leq y_i \leq y_i + \frac{\Delta}{2}$. This is so because all empirical data are measured with finite precision. Apparently, there is no practical difference between discrete and continuous survival time data as far as empirical data are concerned. Therefore the use of continuous-time survival analysis techniques like the AFT frailty model to analyse discrete-time survival data is justified. Although this is so, there are approaches specially tailored for analysing discrete time survival data. According to these approaches, discrete survival data are analysed using logistic regression models. The problem is that when these approaches are used, time ratios are not estimable. This is one reason why we never used them in Paper II. Furthermore, to use the discrete-time survival analysis models, the data have to be expanded in a special way. This creates a very big problem if the event times are big. In that case, one small data set with a few records can translate to a big data set with tens of thousands or even a million of records after expansion. This is a big challenge when fitting models since many laptops have limited capacity to handle big data. This is the second reason why discrete-time survival models were not used to analyse sexual debut data. The problem of big data after data expansion will be illustrated using a hypothetical example. Other examples are in the Sabre R Users Manual [145] on page 138 and in Rabe-Hesketh and Skrondal (2008) [146] on page 337. Suppose we have hypothetical data of time to recovery after admission in hospital. The data comprise three variables namely the Patient Id, Time to recovery (in days) denoted as Time and Whether patient recovered or not

denoted by Status. Table 3 below contains the data.

Table 3: Hypothetical discrete event time data

Patient Id	Time	Status(1= Recovered, 0=Censored)
1	1	0
2	2	1
3	2	0

To expand these data, create a variable Y which stands for outcome of admission in hospital. Let Y=1 denote that a patient recovered and Y=0 denote that by the time data were collected the patient was censored. Each time to event, t, will be decomposed into 1, 2,3,...,t. For each patient, the decomposed event time is listed against the corresponding Patient Id. If the patient recovered, the numbers 1 , 2, ..., t-1 will have a corresponding Y value of 0 but the last number, t, will have a Y value of 1 corresponding to it. If the patient was censored, the numbers 1,2,3, ..., t will all have a corresponding Y value of 0. When the hypothetical data are expanded as explained, one gets the data in Table 4 below.

Table 4: Hypothetical discrete event time data after expansion

Patient Id	Time	Y
1	1	0
2	1	0
2	2	1
3	1	0
3	2	0

Details concerning fitting the logistic regression models on expanded data as explained above are in Rabe-Hesketh and Skrondal (2008) [146] on pages 331-372. The structure and parameterisation of discrete and continuous survival time models are different. Per-

haps, we would have obtained different time ratios, had it been possible to estimate time ratios using discrete time survival techniques. Perhaps our estimated time ratios are biased. In that sense, our use of continuous time analysis techniques on discrete time survival data could be a potential limitation.

Although there are limitations in our analyses in Paper II, there are some advantages associated with the methods in Paper II and the analyses. Our main objective in Paper II was achieved by using logistic regression models. Logistic regression models are standard approaches for modelling a binary response and several explanatory variables. Therefore, we used appropriate methods to determine risk factors. The only contraindication against using logistic regression models is when data have too many zero frequencies. There was no problem of zero frequencies since our sample sizes were very big. Our data were from a multistage cluster sample. Weighted logistic regression uses sampling weights to account for the sampling process [147] and clustering effects. This means that in our analyses for Paper II, clustering was adjusted for. Apart from clustering, logistic regression models also control for confounding. Consequently, we were able to control for confounding effects as evidenced by the sizes of the percentage changes in parameter estimates when covariates were added to univariate logistic regression models as explained above. Logistic regression models produce odds ratios which are easy to interpret and understand. We interpreted odds ratios in our analyses. Our data has discrete event times. Time ratios were computed using AFT models as secondary analyses. As explained above, there is no problem in using AFT models on discrete-event times as there is no difference between continuous event time and discrete event time data. Time ratios are easily obtainable as optional output from AFT models if models are fitted in STATA. The AFT models also controlled for clustering effects.

5.1.3 Projections of standardized incidence rates of cancer and cancer burden for 2015- Paper III

During preliminary analyses for Paper III, several models were fitted and compared. There are several methods for choosing the best model from several models based on the likelihood function. We will mention only two. One method is the likelihood ratio test where an assessment is made of the change in deviance as variables are added to the

model. An application of this approach in the context of finite mixtures is in Misiri and Edriss (2012) [148]. Another method is called the Akaike Information Criterion (AIC). The AIC is given as $AIC = Deviance + 2p$ where p is the number of parameters in the model. An alternative formula [124] for the AIC is $AIC = -2LogL + 2p$ where LogL is the logarithm of the likelihood function.

Some statistical packages produce AICs as output. Examples of software which produce AICs as output are the obsolete GLIM4, SAS and STATA. When choosing the best model using the AIC, several models are fitted and the AIC is computed for each one of them. The model with the smallest AIC is the best model. Note that even models which are not nested can still be compared using the AIC[100]. In Paper III, models of best fit were selected using the AIC. Except for Kaposi sarcoma among females, the final model fitted to cancer data was $E\left(\frac{c_{it}}{n_{it}}\right) = \alpha_i + \beta_i.t$. For Kaposi sarcoma among females the model fitted was $E\left(\frac{c_{it}}{n_{it}}\right) = \alpha_i + \beta.t$. Note that this model constrains all age groups to share the same slope, β.

There were missing values of age in the cancer data analyzed for Paper III. Methods for handling missing data are many but include multiple imputation and complete case analysis. Most of them are built in statistical software. Missing values are best handled using built-in routines. In R for example, multiple imputation is feasible using the **mice** library. In Paper III, corrections for missing values were performed only for the empirical estimates of age standardized rates for all ages combined by multiplying them by $\frac{T}{K}$, where T is the total number of sex specific cancer cases and K is the number for which age is not missing. Further handling of missing values when modeling was not possible since special STATA routines called ados were used for fitting time-linear models. These routines can never be modified by a user. Therefore, the extrapolation models could only be fitted as programmed leaving no room for modification. Consequently, implementation of other methods of handling missing values in the models was not possible.

Our predictions are not perfect but they can serve as a guideline to provide an impression of what cancer burden is in Blantyre, Malawi. The models used for predicting cancer bur-

den in this paper are fitted using STATA routines. The output gives standard errors of prediction and corresponding 95% prediction intervals. The prediction interval of each estimate is computed using the formula: $(Estimate - 1.96.SE, Estimate + 1.96.SE)$ where SE is the standard error of an estimate of cancer burden. In this way, prediction uncertainty is estimated and used in the presentation of results. To save space, the actual standard errors were not included in Table 2 of Paper III. Only prediction intervals based on those standard errors were included. The reader who wants to know the actual standard errors of each estimate of cancer burden for 2015 should work backward from the lower and upper limits of the intervals. In that case, the standard error of each estimate of cancer burden is equal to half of the width of an interval divided by 1.96.

The trends of Kaposi sarcoma, non-Hodgkin lymphoma, eye cancer, oesophageal cancer,cervical cancer ,breast cancer and all other cancers combined are described in Paper III. The burden of these cancers is also estimated. The cancer types in our article have very different incidence, prevalence, treatment, and prognosis. The analyses in Paper III are predominantly descriptive. Such descriptive analyses are common in the field of descriptive cancer epidemiology. In descriptive cancer epidemiology, it is possible to have a research article containing the analysis of data on several cancer types. Examples are Cvancarova et al (2009) [149], Cvancarova et al (2013) [150], Dudarev et al (2013) [151], Yang et al (2004) [152] and Moller et al (2002) [136].

Firstly, Moller et al (2002) is a massive research article with 96 pages published in a reputable European journal. In this article, cancer trends were analysed, future burden estimated and an analysis of the possible demographic effects on the annual number of cases in Scandinavian countries presented. Nineteen cancer types as well as all other cancers combined were analysed for Finland, Norway, Denmark and Sweden. Moller et al (2002)'s report presents results from the analyses of data from Scandinavian cancer registries. In their analyses, Moller et al (2002)'s used age-period-cohort models. Secondly, Dudarev et al (2013) is an analysis of cancer incidence and mortality in Chukotka, Russia. This paper presents results of the analysis of aggregated cancer data extracted from annual statistical reports of the P.A. Hertzen Research Institute of Oncology in Moscow [151]. Dudarev et al (2013) never used statistical models to estimate incidence

and mortality rates of 10 cancers. Thirdly, Cvancarova et al (2009) analysed data on at least three cancers and the results are in one article [149]. In another research article, Cvancarova et al (2013) analysed data on 23 cancers cancers and the results are published elsewhere [150]. These two research articles plus two more [153] are included in Milada Cvancarova's PhD thesis [154]. Lastly, Yang et al (2004) analysed data on many cancers from the cancer reporting system, CHIS, of China. Using time-linear models, they analysed cancer mortality in China. We analysed cancer registry data using simple time-linear extrapolation models. We also give estimates of cancer burden for 2015. Cvancarova et al (2009), Cvancarova et al (20130, Dudarev et al (2013) and Moller et al (2002) analysed data on more than one cancer type. Therefore, our analyses are similar in the sense that we all present results for several cancer types in one article. Our analyses are also similar to the analyses in Yang et al (2004) in that they also used the model, $E\left(\frac{c_{it}}{n_{it}}\right) = \alpha_i + \beta_i.t$ and analysed data on more than 9 cancers.

The time-linear models we have used for extrapolating cancer trends are not new. We are not the first people to apply them. However, our analysis is the first in Africa. In that sense, our use of the models is novel too.

5.1.3.1 Limitations in Paper III

The major limitation for Paper III is missing data since some ages were missing. Had it been that missing values of other variables like the occupation were the problem, that would not have affected modelling. However, age is an important covariate in our models. Therefore missing ages translate to reduced number of observations. Besides, missing ages correspond to missing information. About 10% of the data had missing ages. Probably, our estimates of incidence rates were affected by missing data.

Uncertainty in recorded ages [54] among the elderly in the data collected by the Malawi Cancer Registry as first reported by Banda et al (2001) is a limitation of our analyses. Since age is an important risk factor of cancer, accurate values of age are important in order to have good estimates of incidence rates. With the introduction of free primary education, this is expected to change in the future.

The base of prediction is the period during which cancer data were collected. The length of the base of prediction is another limitation of our analyses. The base of prediction for our analyses was 10 years because we wanted to have enough data for analysis. The underlying assumption of the time-linear models is that cancer trends are linear during the period when data were collected. Our preliminary analyses showed that the incidence trends for 1996-2005 were linear. In general, for small intervals, it might be fairly easy to observe linear trends of cancer incidence. However, for wide bases of prediction, some interventions and changes to the data collection and diagnostic practices affect the incidence rates of cancer. Consequently, the trends of cancer incidence rates can also be affected. Shorter periods are better as long as there is enough data to analyse. For instance, between 1996 and 2005, new hospitals sprouted in Blantyre. Furthermore, between 2000 and 2001, the Blantyre registry received huge funding resulting in sharp rises in incidence on the graphs of incidence trends of most sites. Additionally, during 1996-2005, there were improvements in the diagnostic practices at QECH which might have affected the volume of data collected. These changes during 1996-2005 might have affected our estimates of cancer burden and cancer incidence trends.

The omission of important covariates in the models used for predicting cancer incidence is also a limitation. The number of covariates in all the time-linear models is a naïve collection of time and age variables only. Indeed, the time-linear models were tested by using ex-post predictions of cancer incidence from Finnish data and the results showed that the models produced credible estimates [155]. However, the inclusion of only time and age variables in the models is an oversimplification of reality. Different cancers have different risk factors. Therefore, any cancer incidence model needs to include such factors as covariates. Unfortunately, many important risk factors are difficult to measure as in most cases people are diagnosed when they present with some symptoms. At that time of diagnosis, no measurements of the risk factor prior to or during diagnosis are available or possible. The end result is that cancer data do not contain the essential information on known risk factors of cancer etiology. Time is used in the models as a surrogate of the many overarching factors of cancer etiology. Since age is also a well-known risk factor of cancer it is also included in the time-linear models. It therefore possible that since some important covariates are omitted in the models, there

is some bias in the estimation of rates.

The structure of the model fitted to Kaposi sarcoma data for women is a limitation. This model assumes that all age categories have the same slope. This is wrong. Cancer incidence is known to vary by age. By constraining the slope to be constant for all ages, the incidence rate of Kaposi sarcoma for some categories was either overestimated or underestimated.

The categorisation of age into age groups for modelling is a limitation of the time-linear models. The age groups used for estimating cancer incidence rates depend on cancer site [155]. Depending on cancer site, it is possible for some age groups to have zero frequency. It is impossible to fit the time-linear models when at least one age group has zero frequency or a very small frequency. The remedial measure for zero frequencies or very small frequencies is to combine some age groups in order to have a reasonable number of observations in all age categories to facilitate model fitting. Different combinations of age groups lead to different estimates of incidence rate. Thus it is possible for different people to obtain different estimates from the same data if different age categories or combinations of them are used in modelling.

Incomplete case ascertainment is a limitation of our analyses. The results section of Paper III showed that only a modest percentage (at most 46%) of cases of oesophageal cancer and Kaposi sarcoma were verified morphologically. It means that for these sites, most of the cases might have been wrongly diagnosed as cancers. Therefore, given that some of the cancers were not morphologically verified, it is probable that most of the diagnosed cases are not cancers at all. These wrong diagnoses might have contributed to incidence trends.

Despite the above limitations, the time-linear models which we used for projecting incidence rates and cancer burden have several advantages. These models are simple. The parameterisation is also simple and easy to understand. The models do not impose unnecessary and unrealistic assumptions and parametric forms. It is believed that predictions made using simple models are bound to come true [132]. By using these

models, we were able to produce estimates of incidence rates of cancer and cancer burden together with corresponding prediction intervals. This is important in modeling as the precision of estimation can be inferred from these prediction intervals. The traditional approach for estimating incidence and cancer burden is the age period cohort model [99, 133, 136]. However, even for the APC, the main covariates are the age, period and cohort derived from the date of birth. Crucial risk factors of the development of cancer are not included in APCs. This means that the time-linear models are not that bad after all as far as the number of covariates is concerned. Furthermore, the time linear models do not produce explosive predictions and negative incidence estimates. Furthermore, time-linear models do not have the identifiability problems unlike the APC. In this respect, the simple extrapolation models are better than the APC. As regards precision of estimation, simple models are reputed to produce very precise estimates. Estimates computed using complex models tend to have decreased precision because the prediction intervals computed using complex models are often wide (page 7 of Dybas PhD thesis) [155].

5.2 Discussion of results

The discussion of results is divided into three parts. Each part is dedicated to a paper. Each part begins with a summary of results of the paper followed by the discussion of those results. I have done this in order to help the reader.

5.2.1 Estimation of HIV incidence from cross-sectional sero-prevalence data- Paper I

The first aim of this thesis was to develop a new method for estimating HIV incidence from cross-sectional sero-prevalence data accounting for differential mortality and to apply the method on data from Malawi. We successfully developed a mathematical model. Our model-analyses showed that the estimated HIV prevalence at age 15 years was small for both genders but the estimate for women was higher than for men. The age-sex specific HIV incidence rate estimates in Paper I were higher among women than among men in the 15-24 age range (2031 vs 616), lower among women than among men in the 25-34 age range (2705 vs 1712) and higher among women in the 35-49 age range than among men in the 35-54 age range (1336 vs 1732). The lowest incidence for men was for the 15-24 age range and for women for the 35-49 age range. The same pattern was observed in the estimates of the FOI. Sensitivity analyses showed that when mortality rates were decreased or increased by a small or big percentage, the model responded accordingly. The Chi-square goodness of fit test indicated that the model fitted men's and women's data adequately.

The sensitivity analyses showed that our method responds to changes in the sizes of mortality rates which is good. This implies that given data for two populations with similar sero-prevalence profiles but different mortality rates, our method will respond accordingly and yield different incidence estimates for each population commensurate with the mortality rates. This agrees with what happens in real life situations because rates of HIV mortality affect the sustenance of a sexual network. For an AIDS epidemic to persist, the sexual network has to be adequately sustained.

The small estimates of HIV prevalence at age 15 reflect reality in Malawi as by age 15 most of the young people are not sexually active. The estimates of HIV incidence rate from our model are consistent with the HIV epidemic in Tanzania and Zimbabwe, [156, 157, 158, 159, 160, 161, 162]. The estimated FOI values fluctuate around 0.02. Since in one year the FOI translates to the probability of HIV infection, we get from this that the probability of infection is 0.02 among men and women. Therefore using the fact that the cumulative incidence (CI) or risk of infection is $CI = 1 - e^{-I.(\Delta t)}$, the probability that a 15 year old person will be infected before the age of 50 is equal to $1 - exp\,(-0.02(35)) = 0.503$. Therefore a young person aged 15 has 50% chance of being infected with HIV before age 50.

The age-dependent incidence rate trends of HIV infection as estimated in this paper agree with the age-prevalence trends of HIV in Malawi in 2004. In particular, the high rates of HIV infection in women than in men in the 15-24 and 35-49 age ranges correspond to higher prevalence in women than in men in these age ranges in Malawi in 2004 [6]. Possible explanations for the high rate of infection in women in the 15-24 age range are the low age at first sex in 2004 [6], low age at first marriage in 2004, low condom use at first sex, and indulging in transactional (sex in exchange for some payment in form of money or in kind) or commercial sex [163, 164, 13]. Concerning low age at first sex in Malawi, by 2004 about 15% of women and 14% of men in the 15-24 age range, had their first sex at the age of 15 [6]. Obviously, this narrow margin of 1% implies that low age at sexual debut alone is not enough to explain the higher rate of infection in women. Condom use is another contributing factor. Condom use is low among adolescents [13]. It was even lower at first sex for women than for men in the 15-24 age range by 2004 [6]. Additionally, about 85% of women and 44% of men aged between 25 and 29 years had been married by age 22 by 2004 [6]. Hence, more women were married off earlier than men and so low sexual debut in women might have been the result of early marriages. Furthermore, transactional sex is common in sub-Saharan Africa [3] and there are suggestions that a significant amount of HIV transmissions in this region may be through heterosexual contact between HIV positive men and young girls [164] in relationships where the adolescent has no power to negotiate for condom use during sex because she is in a weaker position as she expects some sort of payment [165]. Prostitution is illegal

in Malawi but transactional or commercial sex is prevalent to the extent that unmarried women aged between 20 and 24 are 3 times more likely to engage in transactional sex than women aged at least 25 years [164]. Selling sex for living by young girls is common especially in urban Malawi [163]. Men often go for such young girls because they are believed to be free from HIV. In some cases, during commercial sex, men have to pay more to have unprotected sex with a young girl [163]. Not only young girls indulge in transactional or commercial sex, but food handlers, bar girls and cleaners offer sex in rest houses and bars also sell sex illegally [13]. In sub-Saharan Africa in general, female adolescents engage in heterosexual transactional sex with older partners [166].

The high incidence rate of infection in women in the 35-49 age range can be explained by low condom use in marriage in Malawi, having extramarital sex and being either divorced, separated or widowed [13]. In Malawi, the majority of women in the 30-49 year age range get married by the age of 25 years [165]. Wives and husbands dislike condom use in marriage [22] because sex is considered "sweet" and using a condom during sex denotes distrust and having unprotected sex with a woman shows mutual love and trust [167]. In general, whether married or not, women in sub-Saharan Africa do not have negotiating power in sexual relationships because of culture-based gender roles [3, 168]. So, they cannot negotiate for condom use during sex. This makes women vulnerable to HIV infection.

A possible reason why the incidence rate of HIV infection was low in men between 15 and 24 years may have been the relatively high age at first sex and first marriage as of 2004 [6]. Coupled with this is the fact that in Malawi, HIV is spread mainly through heterosexual contact and cross generational sex between male youths and older men and women is rare in Malawi. Therefore, unlike girls who can engage in sex at an earlier age, boys engage in sex for the first time relatively later.

Our model estimated an observed higher rate of infection among men aged 25-34 than among women in the same age range. Among men, this age group corresponds to a higher prevalence of HIV than among women as observed in 2004 [6]. This may be due to late age at marriage, engaging in extramarital sex and having concurrent multiple

partners as reported by Geubbels et al (2006) [13]. In Malawi, by 2004, about 70 % of men aged between 30 and 34 years had been married by the age of 25 (as compared to about 94% in women) [6]. In African settings, relationships and sexual encounters within relationships are initiated by men and the 25-34 age range is associated with a high level of economic activity. Since men marry later than women, the 25-34 age range comprises a reasonable percentage (of about 30%) of unmarried men with high economic potential and who are free to have multiple concurrent sexual partners. Besides, men are reported to be reluctant to change their risky sexual behaviors once they perceive themselves to be HIV positive [169]. Furthermore, most Malawians overestimate their probability of being HIV positive [170]. So once they think they are HIV positive they become reckless. Since only those who are economically viable (those aged 25-34) are in a position to pay for sex, this reckless attitude of risky sexual behavior after perceiving oneself to be HIV positive increases the risk of HIV infection in that age range.

The above discussion shows that the mathematical model of Paper I produced credible estimates of HIV incidence rate which were consistent with local data and observations from other sources. This shows that our model is reliable and can produce credible results if applied to a data set from another setting. Note, however, that there have been changes in trends of age at sexual debut and age at marriage among men and women since 2004. For example, by 2010, there was a decrease in the percentage of women age 15-19 who were married by age 15 from 6 percent in 2004 to 4 percent in 2010 [4] implying that there was an increase in age at marriage in Malawi among women between 2004 and 2010 although even in that case men still married considerably later than women [4]. The percentage of women aged 15-24 who are reported to have had early sexual debut is 14 % [4]. This is lower than the 15 % which was reported for the same cohort in 2004. The percentage of women age 15-24 who had early sexual debut in 2010 (14 %) is lower than the percentage of men (22 %) who had early sexual debut but from the same age cohort in the same year [4] implying that although a higher percentage of women had early sexual debut than men in 2004, by 2010 this had been reversed. However, as was observed in 2004, "the proportion initiating sexual activity early is higher among ever-married young women (20 percent) than among those who have not yet married (8 percent)" [4].

5.2.2 Risk factors of early sexual debut among men and women - a strong predictor of HIV and sexual risk in Malawi-Paper II

The aim of Paper II was to determine the risk factors of early sexual debut among men and women in Malawi. Our analysis has shown that for men, early sexual debut was significantly associated with secondary or higher education , Lomwe ethnicity , Islam , being married , being in a cohabiting relationship and the following birth cohorts :1965-1969 , 1980-1984, 1985-1989 and 1990-1995. Among women, early sexual debut was significantly associated with primary education , secondary or higher education, Lomwe ethnicity , other ethnic groups, other Christian denominations, Islam, being married and being separated. Province and the place of residence were not significantly related to early sexual debut for both genders.

Ethnicity seems to be a significant risk factor of early sexual debut since Lomwe men and women were twice more often to have early sexual debut than Chewa men and women. Besides, women of other ethnicities were also 1.4 times more likely to have indulged in early sexual debut than Chewa women. This finding may be explained by the fact that different ethnic groups have different cultural practices some of which predispose boys and girls to early sexual debut. Our findings agree with Traen and Samuelsen (2007) who reported that "Human sexuality is shaped by cultural context and historical period" from a study condcted in Norway [171].

The findings that Islamic men and women were twice more likely to have engaged in early sexual debut than Catholic men and women respectively agree with Fatusi and Bloom (2008) [172] and McGrath et al (2009) [173]. Fatusi and Bloom (2008) found that religion was associated with early sexual debut in Nigeria [172]. Besides, McGrath et al (2009) reported the similar findings from a study conducted in South Africa [173]. The association between religion and early sexual debut may be due to peer pressure within the social network of the religious community [173].

In Paper II, corresponding to married and cohabiting men were adjusted odds ratios of early sexual debut which were less than 1 implying that marriage is protective against early sexual debut. These findings do not in any way suggest that being married or being in a cohabiting relationship protects from early sexual debut. However, these findings agree with the findings of the MDHS2004 which say that Malawian men marry reasonably later than women [6] at a juncture in life when some women of the same age have already had first sex. Among women also, being married was significantly protective against early sexual debut. This finding is not consistent with the reality on the ground in Malawi. Malawi is one of the 20 countries in the world with the highest percentage of child marriages [174]. Furthermore, according to findings from the MDHS2004, women marry early in Malawi [6]. UNICEF also reported that 12% of marriages contracted in Malawi during 2005-2013 were child marriages where either the bride or the bridegroom was aged less than 15 years implying that for a good percentage of boys and girls marriage forced them into early sexual debut [175]. Besides, in 2015 alone, the Malawi Government annulled 600,000 child marriages [176]. The good news is that the Malawi Government passed a law in 2015 that bans child marriages [177]. This law has come at a right time. However, it will take a few years before the mindset of child marriages is completely eradicated.

Birth cohort was an important factor which was associated with early sexual debut especially among men when other factors were adjusted for. In particular, the following birth cohorts were significantly associated with early sexual debut:1965-1969,1980-1984, 1985-1989 and 1990-1995. In fact, there was a parabolic trend in the adjusted odds ratios (Paper II) from 1965-1969 to birth cohort 1990-1995. These results show that although those who were born during 1965-1969 were twice more likely to have engaged in early sexual debut than those born during 1955-1959, the multiple by which those who were born after the 1965-1969 cohort were more likely to engage in early sexual debut started to decrease from birth cohort 1970-1974 to 1975-1979. After birth cohort 1970-1974, there was an increasing trend in the odds ratios up to birth cohort 1990-1995. Individuals born during 1990-1995 were five times more likely to engage in early sexual debut that individuals born during 1954-1959. This shows that as regards early sexual debut those born during 1990-1995 were now practising the habits of indulging

in early sexual debut as those born during 1965-1969 but in a more pronounced way. This finding agrees with MDHS2010 and Zuma et al (2011) [4, 178] who reported birth cohort as one of the risk factors of early sexual debut in South Africa [178]. According to MDHS2010, on a national scale, there was a decrease in age at sexual debut among men during the same period [4]. This means that although age at first sex was notably increasing in the previous years up to 2004, between 2004 and 2010 there was a decrease in age at first sex. Many men engaged in first sex earlier than had been the trend in the past. Apparently, this shows that the trend in age at first sex can fluctuate with time. This observation of changes in age at first sex should leave no room for complacency as it is clear here that age at sexual debut in a population can change with time. In order to achieve low sexual debut among both genders, education must be used as an intervention.

Although birth cohort was significantly associated with early sexual debut among men, among women there was no significant association between early sexual debut and birth cohort. This finding contradicts the results of the MDHS2010. The MDHS2010 reported changes in the age at sexual debut among women between 2005 and 2010 [4]. According to MDHS2010, on a national scale, there was an increase in age at first sex among women from 2004 to 2010 [4]. These findings imply that by 2010 more women were now engaging in first sex later in life than was previously the case in 2004. The reasons for this inconsistency are unknown.

In Paper II, secondary education was a protective factor against early sexual debut among both men and women. This finding agrees with McGrath et al (2009) who also reported that school had the effect of delaying sexual debut among South African men and women [173]. Pettifor at al (2008) also reported that staying in school protected South African girls from the risk of HIV infection [179].

The fact that primary education was a protective factor against early sexual debut among girls can be explained by girls survival tactics to avoid getting pregnant so as to stay in school as pregnancy halts their schooling [180]. Being in school may, therefore, help both boys and girls to delay sexual debut since girls will try to avoid sexual contact

at all costs. If schoolgirls avoid sexual contacts, schoolboys will have nobody in school to have sex with. Therefore both boys and girls in school will have delayed sexual debut.

Evidence in support of our finding of the protective effect of education against early sexual debut is also supplied by MDHS2010. MDHS2010 reports that from a group of women age 15-24, 19.6% had early sexual debut within marriage and 7.9% had early sexual debut outside marriage. From the same cohort of women age 15-24 who had early sexual debut, 27.0 % had no education, 16.2% had primary education and 6.8% had secondary education and 4.5% had higher education. Therefore, the fact that from a cohort of women age 15-24 who had early sexual debut the biggest percentage were uneducated and smaller percentages educated supports our view that education can make a difference in people's lives as far as reducing age at sexual debut is concerned.

Among many possible interventions against early sexual debut, education is relatively easy to implement as an intervention. In particular, primary education can be exploited to control the age at sexual debut and sexual and reproductive health practices of school-going boys and girls since primary education is free in Malawi. In fact, sex and education interventions are said to be effective in delaying sexual debut in developing countries [181]. The main problem is implementation of education as an intervention. Implementation is hampered by lack of resources.

To implement education as an intervention for preventing early sexual debut in Malawi, special campaigns should be conducted to emphasize the importance of education in order to increase enrollment of pupils. Furthermore, sexual and reproductive principles aimed at encouraging pupils to remain in school should be integrated in the curriculum in accordance with community values as regards modesty in speech. Obscene language is strictly frowned upon in Malawi.

The potential effectiveness and sustainability of education as an intervention depends on improving the learning environment and providing some key basic necessities of school boys and girls. The learning environment consists of infrastructure, what happens in the school neighbourhood , pupils and their teachers. The learning environment has to

be made attractive to learners. Even if a campaign succeeds in increasing enrollment rates, the number of pupils can still decrease with time (attrition) as school children drop out of school because of poverty and the poor state of the learning environment. If the learning environment is attractive, pupils will want to go back to school to learn. The school environment will then become a second home. The community also needs to be conducive to learning.

Suggestions for improving the learning environment are many but we will give selected examples. Having well-ventilated and spacious classrooms lit by solar electricity can make classrooms attractive. In Malawi, most of the primary schools are in villages. The classrooms in these primary schools have no electricity. During the rainy season, visibility is poor in classrooms as the skies are cloudy most of the time. Solar electricity is relatively cheap. Therefore it can be used to provide light in primary schools classrooms.

Classrooms need to be furnished with good desks and chairs. In many places in Malawi, primary school children learn under a tree. In most schools, pupils sit on an earth floor, packed in a dimly lit, poorly ventilated classroom. When break time comes, pupils are relieved to have finally made it to the outside to get fresh air. Crammed classrooms are dangerous environments in which diseases can be spread very quickly.

Building good clean, roofed school toilets can also be attractive to both schoolboys and schoolgirls and can help improve school hygiene. Many rural schools have pit latrines in very bad condition. Pit latrines have to be demolished after several years and new ones dug.

Providing school children with important items they need to go to school can help them to remain in school. Supply of sanitary pads or tampons to schoolgirls aged at least 12 years, for instance, can help girls to stay in school. When they feel like they will have menses (i.e. pre-menstrual signs), schoolgirls can go to the secure toilets and get prepared using the sanitary pads or tampons distributed to them.

Regular supply of notebooks, schoolbooks, lead pencils and ballpoint pens can also help students to remain in school. Providing school children with bicycles can also motivate them to remain in school since in some areas schoolchildren have to walk long distances daily as they go to school. In such areas, access to school is a problem during the rainy season. Providing school children with school uniform can also be of great help to them. Some parents and guardians are too poor to afford school uniform for their children.

To make communities conducive to learning, community and religious leaders need to be sensitised to the dangers of early sexual debut, the influence of some bad religious and cultural practices in influencing early sexual debut and the benefits of education. Community leaders are custodians of cultural values and customs of ethnic groups they lead. Therefore, if a community has customs or values which dictate that youngsters should engage in ritual sex after initiation, these leaders can be convinced to influence members of their community to refrain from following such harmful practices by emphasizing the dangers of early sexual debut one of which is HIV infection. The same applies to religious leaders. The dangers of HIV are easy to appreciate since many people in Malawi have heard about or seen an AIDS patient [182].

5.2.3 Projections of standardized incidence rates of cancer and cancer burden for 2015- Paper III

The third aim of this thesis was to describe the pattern of occurrence of some chosen types of cancer and to estimate the burden of these cancers in 2015. The chosen types of cancer under consideration were Kaposi sarcoma, oesophageal cancer, non-Hodgkin lymphoma, eye cancer, breast cancer, cervical cancer and all other cancers combined.

The Malawi Cancer Registry was established in 1989 [54]. Since then, there have been improvements in the diagnosis of cancer. Besides, several new hospitals have been opened in Blantyre since 1989. This has also contributed to an increase in the volume of cancer data collected. What persist are problems related to funding and these affects the volume of data collected. In Paper III, there were sharp rises in the observed trends of all cancers in Figures 3 and 4. Specifically, the observed sharp rises in incidence for all cancers especially for 2000-2001 occurred just because the registry received huge

funding. Thus, improved diagnosis, increased number of hospitals and fluctuations in availability of adequate funding affects the volume of data collected. These are artefactual factors affecting trends in cancer incidence in Blantyre. Therefore, although the incidence rate trends for many cancers were increasing, these increases should be interpreted in the light of the fact that artefactual factors also contributed to the increased trends in cancer incidence rate for many cancer forms.

In Paper III, the temporal patterns of occurrence of Kaposi sarcoma, oesophageal cancer, eye cancer, breast cancer, cervical cancer, non-Hodgkin lymphoma and all other cancers combined as inferred from the incidence trends in Paper III were rather wild and dramatic. These patterns may be due to artefactual factors as already mentioned above. These are the potential sources of uncertainty which may have contributed to the big forecasts of cancer burden for 2015 and the dramatic rises in incidence rate trends between 1996-2005 and 2015. In view of the potential uncertainty due to these artefactual factors, we estimated the cancer burden for two scenarios. For Scenario 1, we assume that the rates of cancer occurrence observed in 1996-2005 will continue to 2015. Under this assumption, the projected mean annual number of cases for 2015 is the lower limit of the projected number of cases to be diagnosed in 2015 which is 1240 cases. This is higher than the mean annual number of cases for 1996-2005 which was 792. For Scenario 2, we assume increasing rates of occurrence of cancer as projected by the time-linear models. These are high rates of occurrence of cancer as can be seen from the incidence rate trends in Figures 3 and 4 of Paper III. The model-based projections of the mean annual number of cancer cases to be diagnosed in 2015 are the upper limit of the expected burden for 2015. These are 2370 cases which are three times bigger than the mean annual number of cases for 1996-2005.

Note that what we get from Scenario I is that even if the rates of occurrence of cancer remain constant from 1996-2005 to 2015 there will still be an increase in the annual number of cases in 2015. In that case, the only determinants of an increase in the mean annual number of cases will be population growth and population ageing.

AIDS-defining cancers formed 60% of the cancers in men in 1996-2005 and 63% of cancers in women during the same period (Paper III). The group designated as all other cancers combined does not contain any HIV-related malignancies. Although this is so, the yearly rate of increase in the incidence of all cancers combined was also big. Since most of the cancers were AIDS-defining cancers one would expect the HIV epidemic to be the only contributing factor to the increases in incidence of AIDS-defining cancers. However, from the discovery of HIV in Malawi in 1985, the HIV prevalence had been increasing up to 1999 when it reached its peak of 15%. Thereafter, the HIV prevalence has been decreasing at a slow rate from 15% in 1998 to 10% in 2014. Additionally, the rate of increase in the incidence of Kaposi sarcoma has been slower than any other malignancies considered although this cancer is also AIDS-defining. These three : the slow rate of increase in incidence of Kaposi sarcoma, the big rate of increase in the incidence of all other cancers combined and the rise and fall of the HIV prevalence trend from 1985 to 2005 imply that the increasing trends between 1996 and 2005 were not due to HIV alone. In fact, during 1996-2005 there were also increases in the incidence of non-AIDS defining cancers like oesophageal and breast cancer. Therefore, apart from HIV, additional contributing factors were population ageing and population growth.

Table 2 of Paper III provides the percentage changes in the mean annual number of cases expected between 1996-2005 and 2015 for two scenarios: assuming the rates of cancer observed between 1996 and 2005 are the same up to 2015 and assuming there are increases in the rates of different cancers. These percentage changes relative to the mean annual number of cases for 1996-2005 were high.

We have adapted Table 2 of Paper III to show the contributing factors to the change in the mean annual number of cases in 2015 relative to 1996-2005. The information is shown in Table 5 below. The difference between the observed number of cases for 2015 and the corresponding number observed during 1996-2005 expressed as a percentage is the percentage total change which is in the last column of Table 2 of Paper III. This was decomposed into the percentage change due to changes in population age structure and size (demographic change) and changes in the risk of cancer. The percentage changes due to changes in population age structure and size (demographic change) are

in Columns 2 and 5 of Table 5. The percentage changes due to changes in the risk of cancer 2015 are in Columns 3 and 6 of the same table. Note that Table 5 has been derived from Table 2 of Paper III using the method of Moller et al (2002) as explained in Subsection 3.2.5 of the methodology section of this thesis.

Table 5: Projected changes in the number of cases in 2015 due to percentage changes in population age structure and size and percentage changes in risk of different cancers in Blantyre, Malawi by 2015.

Men	-			**Women**	-	
Cancer	Population*	Risk++		Cancer	Population*	Risk++
Oesophagus	51.2	55.8		Oesophagus	62.5	95.8
Kaposi sarcoma	52.6	79.1		Kaposi sarcoma	62.2	63.3
Eye	50.0	272.2		Eye	65.0	305.0
NHL	53.3	330.0		NHL	57.1	285.7
-	-	-		Cervix	66.0	152.4
-	-	-		Breast	65.2	91.3
Other cancers	47.2	164.2		Other cancers	60.2	233.3
Total	50.7	129.0		Total	62.8	157.3

Adapted from Table 2 of Misiri et al(2012b)(Paper III)

* % change due to changes in the population age structure and size (also known as demographic change)

++ % change due to changes in the risk of cancer

The results in Table 5 clearly show that about 50 % and 60 % of the percentage changes in the number of cases among males and females respectively in 2015 will be due to demographic change. These contributions to the total percentage change in the number of cases in 2015 are modest. Additional factors which will also contribute to the increase in the mean annual number of cases by 2015 are increases in the risks of cancers. There are projected increases in the risks of Kaposi sarcoma, non-Hodgkin lymphoma, eye cancer, oesophageal cancer, breast cancer, cervical cancer and all other cancers combined as shown in Table 5. Excluding all other cancers combined, the following AIDS defining cancers have very big percentage increases in risk : eye cancer, NHL and cervical cancer.

The lowest percentage changes in risk among the AIDS-defining cancers are for Kaposi sarcoma. These low projections are reasonable considering that provision of ART which suppresses the manifestation of Kaposi sarcoma in HIV positive persons has been rolled out in Malawi since 2004.

The projected increases in the risk of Kaposi sarcoma, non-Hodgkin lymphoma, eye cancer, oesophageal cancer, breast cancer, cervical cancer and all other cancers combined (Table 5) are expected to arise from western lifestyle like sedentary behaviour, increase in tobacco smoking and HIV (Paper III). In fact, as regards smoking, in 2010, the prevalence estimates of tobacco use among men and women in Malawi were 17% and 1 % respectively [4]. As of 2010, among men, tobacco use was predominantly cigarette smoking and the prevalence estimates of tobacco smoking were 2.5, 12.8, 21.2, 22.8, 27.4, 25.5 and 31.6 % for male persons aged 15-19, 20-24, 25-29, 30-34, 35-39, 40-44 and 45-49 respectively. By 2015, the prevalence of tobacco smoking is expected to increase as people adopt western lifestyles like smoking as a result of changes in socioeconomic status and the influence of what they see on television.

Sedentary behaviour is a risk factor for colorectal cancer. Both lung cancer and colorectal cancer are included in the group designated as 'all other cancers combined'. In fact, according to GLOBOCAN2012, in terms of frequency of occurrence, colorectal cancer is among the 6 top cancers in 2012 in Uganda (ASR = 7.7 per 100,000), Zimbabwe (ASR = 8.1 per 100,000), Zambia (ASR = 5.3 per 100, 000), Malawi (ASR = 3.4 per 100, 000), Sub-Saharan Africa (ASR = 604 per 100, 000), and in the world (ASR = 20.6 per 100, 000) [66].

Among women, the constant model-based incidence rate trend of Kaposi sarcoma between 1996 and 2005 (Figure 4 of Paper III) is a paradox. Kaposi sarcoma was the second most frequent cancer among women during 1996-2005 (Paper III). Note that 1996-2005 is a period during which HIV was still endemic. The trends of all AIDS-defining cancers among both genders were increasing during the same period. Considering that Kaposi sarcoma is also AIDS-defining, one would have expected an increasing trend between 1996 and 2005. The estimates of percentage changes in risk of Kaposi sarcoma among

women in Table 3 and the percentage changes in the mean annual number of cases in Table 2 of Paper III do contradict the constant model-based trend for Kaposi in women in Figure 4 of Paper III. Under normal circumstances, one would have expected the three results to agree since the percentage changes in the table above (Table 3) are derived from the estimates of the total number of cases for 2015 under two scenarios computed from Model 2 of Paper III $\left(E\left(\frac{c_{it}}{n_{it}}\right) = \alpha_i + \beta t\right)$ and the mean annual number of cases for 1996-2005. Since the estimates for computing percentage changes in the risk of cancer and incidence rate trends are the product of the same time-linear extrapolation model, this constant incidence rate trend for Kaposi sarcoma among women is a strange finding which is difficult to explain. Perhaps, the most likely reasons for this finding are low morphological verification values for Kaposi sarcoma, the structure of the model fitted to Kaposi sarcoma data for women and the provision of ART. Low morphological verification values imply that not all cases of Kaposi sarcoma are morphologically verified perhaps due to heavy workload of laboratory staff or lack of resources. Concerning the structure or formulation of the model used for predicting the burden of Kaposi sarcoma among females, the model forces each age group to have the same constant slope. This may have resulted in the trend of Kaposi sarcoma among women to be constant. Dyba et al (2008) states that the prevention of death during the period when data were collected might lead to the failure of the time-linear extrapolation models [132]. ARVs were introduced in Malawi in 2000. Although the rolling out of ART in Malawi started in 2004, the provision of ART in Blantyre from 2000 to 2004 may have been substantial since Blantyre is a major urban centre with a big referral hospital. This provision of ART may have disturbed the smooth development [132] of the incidence rate of Kaposi sarcoma among women. Women are more likely to have their HIV status detected since almost all pregnant women who attend antenatal clinics are tested for HIV.

The results from the analyses in Paper III are taken to represent the pattern of occurrence of cancer in Malawi although the data were collected from Blantyre district only. This is in order. The Malawi Cancer Registry is the only institution that has conducted regular cancer data collection since 1989. As such, for the time being, the registry is the only most reliable source of data which can be analyzed to provide an insight into the problem of cancer in Malawi.

In 2012, Msyamboza et al (2012) analyzed data from a cross-sectional survey that collected cancer data from a cluster sample of health facilities in Malawi and the results are published elsewhere [183]. The survey was conducted by the Malawi Cancer Registry. This collection of data on a national scale is a commendable step in the right direction. However, the dependability of Msyamboza et al (2012) as a source of national cancer data in Malawi is a subject for debate.

Firstly, Msyamboza et al (2012) reports that data were collected from a cluster sample of 80 out of 84 health facilities providing cancer diagnosis, treatment or palliative care services in Malawi [183]. However, the reality on the ground is that as far back as 2002, Malawi had a total of 664 health facilities 101 (16.6 %) of which were hospitals and 401 (60.4 %) health centres (see Table 6).

Table 6: Number of health facilities in Malawi in 2002 by type of facility

i	Facility	No	%	i	Facility	No	%
1	Central Hospital	4	0.6	8	Maternity	16	2.4
2	Clinic	57	8.6	9	Mental Hospital	2	0.3
3	Dispensary	81	12.2	10	Rehabilitation Centre	2	0.3
4	District Hospital	22	3.3	11	Rural Hospital	37	5.6
5	Health Centre	401	60.4	12	Urban Health Centre	6	0.9
6	Health Post	4	0.6	13	Voluntary Counselling	3	0.5
7	Hospital	29	4.4	-	-	-	-
Total	-	-	-	-	-	664	100.0

Source: Health Facilities Database: National Health Facilities Inventory Survey. Malawi Government, 2002

By 2012, the total number of health facilities in Malawi was more likely to have been greater than 664. Secondly, as far as availability of diagnostic equipment is concerned, some cancers like cervical cancer can be diagnosed with relative ease (for example with acetic acid) in almost any type of health facilitiy. The commonest health facilities in Malawi which are in every district are health centers (60.4 %) which in most cases are

well equipped to handle simple diagnoses like the visual inspection with acetic acid for cervical cancer diagnosis. Therefore, it sounds incredible to claim that only 84 health facilities out of over 664 health facilities were offering cancer diagnosis, treatment or palliative care services in Malawi in 2012. Thirdly, the Msyamboza et al (2012) survey was conducted by the Malawi Cancer Registry- the same institution with the persistent problem of lack of infrastructure to operate from and funding for registry operations. The Malawi Cancer Registry is already an established entity covering a limited catchment area of one district with relatively few health facilities and even in such cases lack of funding and population coverage are a big problem. It is therefore difficult to obtain reliable and complete cancer data from a national survey. Furthermore, cancer data collection is a process which is supposed to be conducted in a meticulous way as it involves extracting information from health facility records. The data in the health facility records are collected by people who, in most cases, do not know the significance of what they are collecting and the data are recorded on forms not customized for research use. Therefore, a national survey of 80 health facilities is a very difficult undertaking which cannot yield good quality results given limited funding and a limited study period. In any case, a cross-sectional national cancer survey is not a good study design for collecting cancer incidence data in Malawi because of the above reasons. In contrast, Paper III contains results from in-depth analyses of trends and projections of cancer burden from the cancer incidence data in Blantyre. These data were collected on a regular basis and were subjected to thorough checks using IARC-CHECK software [184] in order to improve data quality unlike the cross-sectional data from the national survey which were collected hurriedly in a limited time span. Therefore, the results from the analysis in Paper III are of superior quality and reliability in relative terms as far as the provision of an insight into the pattern of occurrence of cancer in Malawi is concerned.

A good data collection strategy that can yield good quality and complete cancer data on a national scale would be to have a national network of cancer registration officers in all district hospitals whose duty will be to collect cancer data in each district. These data should be channeled into a centralized national cancer database. These officers should be headed by a national Cancer Registrar General. The data collection officers should visit health facilities in their district on a regular basis to extract cancer data.

In this way, all national cancer data can be captured in one national database. The network should then be supported by a vital registration system and increased literacy rate. Consequently, collection of cancer data on a national scale can be complete. However, the feasibility of this approach depends on the availability of adequate funding and resources.

6 Conclusion

Mathematical models are cost-effective and non-invasive tools in the quest for cheap methods for estimating HIV incidence. Although there are various models that can be used to estimate HIV incidence, models that require few inputs should be preferred. Some mathematical models require as input quantities that can only be measured using sophisticated equipment operated by well-trained personnel-inputs like CD4 count. In many low resource settings sophisticated equipment and corresponding expertise are rare or hard to find. Consequently, data collected using such sophisticated equipment are also rare. Good mathematical models should use input which can be collected easily in a low resource setting like Malawi. To monitor the HIV epidemic in Malawi effectively, estimates of HIV incidence are needed. Prevalence alone can never provide details about the extent of an epidemic. In the presence of ART provision, HIV prevalence is very imprecise in describing an epidemic since most of the HIV-positives live longer. Consequently, it is hard to observe a reasonable reduction in HIV prevalence. The change in HIV prevalence in Malawi in a decade is approximately 2%. Had the spread of HIV been monitored by using the HIV incidence, the truth of the extent of the HIV epidemic could have been known. Given the massive scale at which the Government of Malawi has, for over a decade, sensitized its citizens of HIV and the need for change in behaviour, the reduction in HIV prevalence should have been bigger. Apparently, HIV prevalence does not give information about new HIV infections in Malawi. We developed a simple but good mathematical model which when applied to Malawi data produced incidence estimates which were consistent with the available data on the epidemic in the region. This model was able to capture age and sex specific differences in HIV incidence and prevalence. The goodness of fit tests showed that the models fitted to men and womens data fitted adequately. The sensitivity analyses showed that the model responded to changes in the sizes of the natural and HIV mortality rates used in

the assessment of the model. The data on which our model was applied were collected in 2004 when rolling out of ART provision was started. By 2014, the ART coverage was 50 percent. This means that Malawi is divided into areas which have access to ART and areas with no access to ART. In those areas where ARVs are provided to HIV positives, this model cannot be used to estimate incidence. However, in the remaining areas with no ART coverage, our model can be used to estimate HIV incidence. Our model can be extended for use in HIV incidence estimation in areas with ART coverage not only in Malawi but in other countries as well. Extra data that may be required in that case are survival data of HIV patients on ART. Therefore, there is the need to continue collecting population-based HIV sero-prevalence data and time-to-event data of HIV patients on ART. The extension of the model , however, is the subject for future research.

HIV infection is treatable but incurable. Treatment for HIV is available in some countries. This treatment has resulted in the improvement of survival of children born with HIV, children infected behaviorally and infected adults. Since HIV infection is chronic, those who survive and live long with HIV have other problems following them: adherence to drugs, drug interactions, complications and resistance of HIV to ART. Although treating HIV is possible and beneficial, the horror of having to live with HIV and the attendant complications of long-term ARV use can be avoided if sexual debut can be delayed. We used statistical models to compute time and odds ratios. The significant risk factors of early sexual debut among men were secondary education , Lomwe ethnicity , Islam , being married , being in a cohabiting relationship and the following birth cohorts :1965-1969, 1980-1984, 1985-1989 and 1990-1995. The significant risk factors of early sexual debut among women were primary education, secondary or higher education, Lomwe ethnicity , other ethnic groups,Islam, being married and being separated. Of all these factors , education should be used to help delay early sexual debut among boys and girls in Malawi. If this can be achieved, the spread of HIV will also be controlled. Delaying sexual debut is important because it well-known that early sexual debut predisposes youngsters to risky sexual practices and HIV infection.

HIV has contributed to the increase in the incidence of many cancers in Blantyre, Malawi. Except for Kaposi sarcoma among women, the incidence rate trends of for

Kaposi sarcoma, non-Hodgkin lymphoma, breast, cervical ,eye, oesophageal and all cancers combined were all increasing during 1996-2005. These trends will continue to increase up to 2015. The observed increases in trend during 1996-2004 were partially artefactual due to improved diagnosis, increased number of hospitals and fluctuations in availability of adequate funding . Cancer is a currently a big problem in Blantyre,Malawi and will continue to be a public health problem in the coming decades due to population growth, population ageing, adoption of western lifestyles, increases in the risk factors of some cancers (eg increases in the prevalence of tobacco smoking) and the HIV epidemic. The mean annual number of cases of Kaposi sarcoma, non-Hodgkin lymphoma, breast, cervical, eye, oesophageal and all cancers combined in Blantyre will be big in 2015. This burden of cancer will be dominated by AIDS-defining cancers. Therefore, urgent consideration should be given to the orientation of preventive and curative services for cancer and other non-communicable diseases in the years ahead.

References

[1] WHO. HIV/AIDS. http://www.who.int/mediacentre/factsheets/fs360/en/.

[2] Barnett T and Whiteside A. *AIDS in the Twenty First Century−Disease and Globalisation*. Pallgrave Macmillan, New York, 2006.

[3] Ramjee G and Daniels B. Women and HIV in Sub-Saharan Africa. *AIDS Research and Therapy*, 10(30), 2013.

[4] National Statistical Office (NSO) and O. R. C. Macro. *Malawi Demographic and Health Survey 2010*. NSO and ORC Macro, Zomba,Malawi, 2010.

[5] Bowie C, Mathanga D, and Misiri H. Poverty, access and immunisation in Malawi-a descriptive study. *Malawi Medcol Journal*, 18(1):19–27, 2006.

[6] National Statistical Office (NSO) and O. R. C. Macro. *Malawi Demographic and Health Survey 2004*. NSO and ORC Macro, Zomba,Malawi, 2005.

[7] AVERT. HIV and AIDS in Malawi. http://www.avert.org/aids-malawi.htm.

[8] WHO. Malawi:WHO Statistical Profile. http://www.who.int/gho/countries/mwi.pdf?ua=1.

[9] WORLDBANK. Prevalence of HIV, total percentage of population ages 15-49. http://data.worldbank.org/indicator/SH.DYN.AIDS.ZS.

[10] National AIDS Commission. *Malawi AIDS Response Progress Report 2015*. Malawi Government, Lilongwe,Malawi, 2015.

[11] Blower S and Farmer P. Predicting the public health impact of antiretrovirals: Preventing HIV in developing countries. *AIDS Science*, 3(11):373–384, 2003.

[12] Nakagawa F, May M, and Phillips A. Life expectancy living with HIV: recent estimates and future implications. *Curr Opin Infect Dis*, 26:17–25, 2013.

[13] Geubbels E and Bowie C. Epidemiology of HIV/AIDS in adults in Malawi. *Malawi Medical Journal*, 18(3):99–121, 2006.

[14] HIV Unit. *ART Supervision Database*. Ministry of Health,Malawi Government, Lilongwe,Malawi, 2009.

[15] HIV Unit. *2012 Global AIDS Response Progress Report:Malawi Country Report for 2010 and 2011*. Ministry of Health,Malawi Government, Lilongwe,Malawi, 2011.

[16] HIV Unit. *Integrated HIV Program Report -January March 2012*. Ministry of Health,Malawi Government, Lilongwe,Malawi, 2012.

[17] The Henry Kaiser Family Foundation. Reported number of people receiving antiretroviral therapy. http://kff.org/global-indicator/arv-treatment.

[18] WORLDBANK. Antiretroviral therapy coverage(percentage of people living with HIV). http://data.worldbank.org/indicator/SH.HIV.ARTC.ZS.

[19] Jahn A, Floyd S, Crampin AC, Mwaungulu F, Mvula H, Munthali F, McGrath N, Mwafilaso J, Mwinuka V, Mangongo B, Pine PEM, and Glynn JR. Population-level effect of HIV on adult mortality and early evidence of reversal after introduction of antiretroviral therapy in Malawi. *Lancet*, 371:1603–1611, 2008.

[20] Floyd S, Molesworth A, Dube A, Banda E, Jahn A, Mwafulirwa C, Ngwira B, Branson K, Crampin AC, Zaba B, Glynn JR, and French N. Population-level reduction in adult mortality after extension of free Anti-Retroviral Therapy provision into rural areas in Northern Malawi. *PLoS ONE*, 5(10), 2010.

[21] Murray CJL, Ortblad KF, Guinovart C, Lim SS, Wolock TM, Roberts DA, Dansereau EA, Graetz N, Barber RM, Brown JC, Wang H, Duber HC, Naghavi M, and Dicker D. Global, regional, and national incidence and mortality for HIV, tuberculosis, and malaria during 19902013: A systematic analysis for the Global Burden of Disease Study 2013. *Lancet*, 384(9947):1005–1070, 2014.

[22] Chimbiri AM. The condom is an intruder in marriage: Evidence from rural Malawi. *Social Science and Medicine*, 64:1102–1115, 2007.

[23] Lufesi N, Andrew M, and Aursnes I. Deficient supplies of drugs for life threatening diseases in an African community. *BMC Health Serv. Res.*, 7(86), 2007.

[24] WHO. HIV/AIDS treatment and care. http://www.who.int/hiv/topics/treatment/en.

[25] White AB, Mirjahangir JF, Horvath H, Anglemyer A, and Read JS. Antiretroviral interventions for preventing breast milk transmission of HIV. *Cochrane Database of Systematic Reviews*, 10, 2014.

[26] Aghokeng AF, Monleau M, Peymard-Duvernay S, Dagnra A, Kania D, Ngo-Giang-Huong N, Toni TD, Tour-Kane C, Truong LXT, Delaporte E, Chaix ML, Peeters M, , and Ayouba A. Extraordinary heterogeneity of virological outcomes in patients receiving highly active antiretroviral therapy and monitored with the World Health Organization public health approach in Sub-Saharan Africa and Southeast Asia . *Clin Infect Dis*, 58(1):99–109, 2014.

[27] Jaquet A, Garanet F, Balestre E, Ekouevi DK, Azani JC, Bognounou R, Dah E, Kondombo JC, Dabis F, and Drabo J. Antiretroviral treatment and quality of life in africans living with hiv: 12-month follow-up in burkina faso. *Journal of the International AIDS Society*, 16(18867), 2013.

[28] Peitzmeier SM, Grosso A, Bowes A, Caesay N, and Baral SD. Associations of stigma with negative health outcomes for people living with HIV in the Gambia: implications for key populations. *J Acquir Immune Defic Syndr*, 68(Suppl 2):S146–153, 2015.

[29] Sohn AH and Hazra R. The changing epidemiology of the global paediatric HIV epidemic:Keeping track of perinatally HIV-infected adolescents. *Journal of the International AIDS Society*, 13(18555), 2013.

[30] Mofenson LM and Cotton MF. The challenges of success: adolescents with perinatal HIV infection. *Journal of the International AIDS Society*, 16(18650), 2013.

[31] Weber HC, Gie RP, and Cotton MF. The challenge of chronic lung disease in HIV-infected children and adolescents. *Journal of the International AIDS Society*, 16(18603), 2013.

[32] Cornell M Laughton B, Boivin M, and Van Rie A. Neurodevelopment in perinatally HIV-infected children: a concern for adolescence. *Journal of the International AIDS Society*, 16(18603), 2013.

[33] Lipshultz SL, Miller TL, Wilkinson JD, Scott GB, Somarrib G, Cochran TC, and Fisher SD. Cardiac effects in perinatally HIV-infected and HIV-exposed but uninfected children and adolescents: A view from the United States of America. *Journal of the International AIDS Society*, 16(18597), 2013.

[34] Bhimma R, Purswani MU, and Kala U. Kidney disease in children and adolescents with perinatal HIV-1 infection. *Journal of the International AIDS Society*, 16(18596), 2013.

[35] Mellins CA and Malee KM. Understanding the mental health of youth living with perinatal HIV infection: Lessons learned and current challenges. *Journal of the International AIDS Society*, 16(18593), 2013.

[36] Puthanakit T and Sibbery GK. Bone health in children and adolescents with perinatal HIV infection. *Journal of the International AIDS Society*, 16(18575), 2013.

[37] Barlow-Mosha L, Eckard AR, McComsey GA, and Musoke PM. Metabolic complications and treatment of perinatally HIV-infected children and adolescents. *Journal of the International AIDS Society*, 16(1600), 2013.

[38] The Aids Info Net. Older people and HIV. http://www.aidsinfonet.org/factsheets/view/616.

[39] O'Brien KK, Solomon P, Trentham B, MacLachlan D, MacDermid J, Tynan AM, Baxter L, Casey A, Chegwidden W, Robinson G, Tran T, Wu J, and Zack E. Evidence-informed recommendations for rehabilitation with older adults living with HIV: a knowledge synthesis. *BMJ Open*, 4, 2014.

[40] Finkelstein JL, Gala P, Rochford R, Glesby MJ, and Mehta S. HIV/AIDS and lipodystrophy: Implications for clinical management in resource-limited settings. *Journal of the International AIDS Society*, 18(19033), 2015.

[41] Maldonado-Martinez G, Fernndez-Santos DM, Rios-Olivares E, Mayor AM, and Hunter-Mellado RF. HIV/AIDS in the Puerto Rican elderly: Immunological changes between gender and Body Mass Index. *J Health Care Poor Underserved*, 24((40):94–105, 2013.

[42] Hopcroft L, Bester L, Clement D, Quigley A, Sachdeva M, and Rourke SB amd Nixon SA. "My body's a 50 year-old but my brain is definitely an 85 year-old": Exploring the experiences of men ageing with HIV-associated neurocognitive challenges. *Journal of the International AIDS Society*, 16(18506), 2013.

[43] Guaraldi G, Zona S, Orlando G, Squillace N, Stentarelli C, Nardini G, Beghetto B, Esposito R, and Palella FJ. Age-related co-morbidities in people living with HIV. *Journal of the International AIDS Society*, 11(Suppl 1):300, 2008.

[44] Coroiu C, Tollinchi F, Hittinger G, Mokhtari S, Granet P, Faucher−Zaegel O, Madrid A, Menard A, Mattei JL, Chadapaud S, and VISAGE study group. Drug interactions in the elderly HIV-infected patient. *Journal of the International AIDS Society*, 13(Suppl 4):108, 2010.

[45] Garcia-Calleja JM, Gouws E, and Ghys PD. National population based HIV prevalence surveys in sub-Saharan Africa: Results and implications for HIV and AIDS estimates. *Sex Transm Infect*, 82(Suppl III):iii64–iii70, 2006.

[46] Ministry of Health. *HIV and Syphilis Sero Survey and National HIV Prevalence and AIDS Estimates Report for 2007*. National AIDS Commision, Lilongwe, Malawi, 2008.

[47] Marschner I. A method for assessing age−time disease incidence using serial prevalence data. *Biometrics*, 53(4):1384–1398, 1997.

[48] Friis RH. *Epidemiology For Public Health Practice*. Jones and Bartlett Publishers, London,UK, 2008.

[49] Saidel T, Sokal D, Rice J, Buzingo T, and Hassig S. Validation of a method to estimate age-specific Human Immunodeficiency Virus (HIV) incidence rates in developing countries using population-based seroprevalence data. *American Journal of Epidemiology*, 144(3):214–223, 1996.

[50] Salomon JA and Murray CJL. Modelling HIV/AIDS epidemics in sub-Saharan Africa using seroprevalence data from antenatal clinics. *Bulletin of the World Health Organization*, 797:596–607, 2001.

[51] Langohr K. Estimation of the incidence of disease with the use of prevalence data. http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.31.1504rep=rep1type=pdf.

[52] Montana LS, Mishra V, and Hong R. Comparison of HIV prevalence and antenatal care surveillance and population-based surveys in sub-Saharan Africa. *Sex Transm Infect*, 84:i78–i84, 2008.

[53] Gail MH and Brookmeyer R. Predicting the incidence of HIV. *JAMA*, 263(11):1538–1539, 1990.

[54] Banda L T, Parkin DM, Dzamalala CP, and Liomba NG. Cancer incidence in Blantyre, Malawi 1994-1998. *Tropical Medicine & International Health*, 6(4):296–304, 2001.

[55] Breivik J. No solution to cancer: Have our genes evolved to turn against us? http://www.sciencedaily.com/releases/2007/04/070416160429.htm.

[56] Boyle P. *World Cancer Report 2008*. IARC, Lyon,France, 2008.

[57] Stewart BW and Wild CP. *World Cancer Report 2014*. IARC, Lyons,France, 2014.

[58] McGlashan ND. Oesophageal cancer and alcoholic spirits in Central Africa. *Gut*, 6:643–650, 1969.

[59] Reeve PA. Lung cancer in Malawi: Experience of bronchoscopy in Zomba General Hospital. *Cent Afr J Med*, 35(3):35–58, 1989.

[60] BenEzra D and Chirambo MC. Incidence of retinoblastoma in Malawi. *J Pediatr Ophthalmol*, 13(6):340–343, 1976.

[61] Norredam K. Primary carcinoma of the liver in Malawi: A review of 24 cases. *Trop Geogr Med*, 29(2):141–149, 1977.

[62] O'Connell KM. Kaposi's sarcoma: Histopathological study of 159 cases from Malawi. *J Clin Path*, 30:687–695, 1977.

[63] Gbabe OF, Okwundu CI, Dedicoat M, and Freeman EE. Treatment of severe or progressive Kaposi's sarcoma in HIV-infected adults. *Cochrane Database of Systematic Reviews*, 9, 2014.

[64] Anglemyer A, Agrawal AK, and Rutherford GW. Treatment of Kaposi sarcoma in children with HIV-1 infection. *Cochrane Database of Systematic Reviews*, 1, 2014.

[65] Gichuhi S and Irlam JH. Interventions for squamous cell carcinoma of the conjunctiva in HIV-infected individuals. *Cochrane Database of Systematic Reviews*, 2, 2013.

[66] IARC. Fact Sheets by Population. http://globocan.iarc.fr/Pages/fact-sheets-population.aspx.

[67] IARC. Fact sheets by population. http://globocan.iarc.fr/Pages/fact-sheets-population.aspx.

[68] IARC. *Cancer Incidence in 5 Continents Vol X*. IARC, Lyon,France, 2014.

[69] Gopal S, Krysiak R, Liomba NG, Horner MJ, Shores CG, Alide N, Kamiza S, Kampani C, Chimzimu C, Fedoriw Y, Dittmer DP, Hosseinipour MC, and Hoffman IF. Early experience after developing a pathology laboratory in Malawi, with emphasis on cancer diagnoses. *PLOS One*, 8:8, 2013.

[70] Parkin DM and Bray F. Evaluation of data quality in the cancer registry: Principles and methods Part II. Completeness. *European Journal of Cancer*, 45:756–764, 2009.

[71] Parkin DM, Sitas F, Chirenje M, Stein L, Abratt R, and Wabinga H. Part I: Cancer in indigenous Africans : Burden, distribution, and trends. *Lancet*, 9:683–692, 2008.

[72] WHO. Palliative care: for all ages and all diseases. http://www.who.int/features/2015/southafricapalliativecare/en/.

[73] Box GEP. Science and statistics. *Journal of the American Statistical Association*, 71(356):791–799, 1976.

[74] Nokes DJ and Anderson RM. The use of mathematical models in the epidemiological study of infectious diseases and in the design of mass immunization programmes. *Epidemiological Infections*, 101:1–20, 1988.

[75] Muench H. Derivation of rates from summation data by the catalytic curve. *Journal of the American Statistical Association*, 29:25–38, 1934.

[76] Griffiths DA. A catalytic model of infection for measles. *Applied Statistics*, 23(3):330, 1974.

[77] Grenfell BT and Anderson RM. The estimation of age-related rates of infection from case notifications and serological data. *The Journal of Hygiene*, 95(2):419–436, 1985.

[78] Long GH, Sinha D, Read AF, Pritt S, Kline B, Harvill ET, Hudson PJ, and Bjrnstad ON. Identifying the age cohort responsible for transmission in a natural outbreak of Bordetella bronchiseptica. *PLOS Pathogens*, 6(12), 2010.

[79] Massad E, Raimundo SM, and Silveira ASB. A continuous function model for the age-related force of infection. *Mathl Comput Modelling*, 13(4):101–112, 1990.

[80] Ades AE, Peckham CS, Dale GE, Best JM, and Jeansson S. Prevalence of antibodies to herpes simplex virus types 1 and 2 in pregnant women, and estimated rates of infection. *Journal of Epidemiology and Community Health*, 43:53–60, 1989.

[81] Nokes DJ and Anderson RM. Rubella sero-epidemiology in South East England. *J Hygiene Camb*, 96:291–304, 1986.

[82] Hens N, Aerts M, Faes C, Shkedy Z, Lejeune O, Van Damme P, and Beutels P. Seventy-five years of estimating the force of infection from current status data. *Epidemiol Infect*, 138(6):802–812, 2009.

[83] Keiding N, Begtrup K, Scheike T.H, and Hsibeder G. Estimation from current status data in continuous time. *Life Data Analysis*, 2:119–139, 1996.

[84] Namata H, Shkedy Z, Aerts M, Molenberghs G, Theeten H, Van Damme P, and Beutels P. Estimation of the force of infection from current status data using generalised linear mixed models. *Journal of Applied Statistics*, 34(8):923–939, 2007.

[85] Hens N, Shkedy Z, Aerts M, Faes C, Van Damme P, and Beutels P. *Modeling infectious disease parameters based on serological and social contact data:A Modern Statistical Perspective*. Springer, New York,USA, 2012.

[86] Brandgo-Fijo S, Campbell-Lendrum D, Davies DR, Brito MEF, Shaw JJ, and Davies CR. Epidemiological surveys confirm an increasing burden of cutaneous leishmaniasis in north-east Brazil. *Transactions of the Royal Society of Tropical Medicine and Hygiene*, 93:488–494, 1999.

[87] Garnett GP. An introduction to mathematical models in sexually transmitted disease epidemiology. *Sexually Transmitted Infections*, 78:7–12, 2002.

[88] Abebe A, Nokes DJ, Dejene A, Enquselassie F, Messele T, and Cutts FY. Seroepidemiology of Hepatitis B virus in Addis Ababa,Ethiopia:Transmission patterns and vaccine control. *Epidemiol Infect*, 131(1):757–770, 2003.

[89] Colugnati FAB, Staras SAS, Dollard SCD, and Cannon MJ. Incidence of cytomegalovirus infection among the general population and pregnant women in the United States. *BMC Infectious Diseases*, 7, 2007.

[90] Conn PB, Cooch EG, and Peter Caley P. Accounting for detection probability when estimating force-of-infection from animal encounter data. *Journal of Ornithology*, 2010.

[91] Williams B, Gouws E, Wilkinson W, and Karim SA. Estimating HIV incidence rates from age prevalence data in epidemic situations. *Statistics in Medicine*, 20:2003–2016, 2001.

[92] Hallett TB, Zaba B, Todd J, Lopman B, Mwita W, Biraro S, Gregson S, and Boerma JT. Estimating incidence from prevalence in generalised HIV epidemics:Methods and Validation. *PLOS Medicine*, 5(4), 2008.

[93] Rajan SS and Sokal D. A simpler tool for estimation of HIV incidence from cross-sectional, age-specific prevalence data. *Journal of Epidemiology and Community Health*, 2010.

[94] Leske MC, Ederer F, and Podgor M. Estimating incidence from age-specific prevalence in glaucoma. *American Journal of Epidemiology*, 113:606–613, 1981.

[95] Keeling MJ and Rohani P. *Modeling Infectious Diseases in Humans and Animals*. Princeton University Press, Princeton,USA, 2008.

[96] Agresti A. *Categorical Data Analysis*. John Wiley and Sons, New York, 1990.

[97] Dyba T, Hakulinen T, and Paivarinta L. A simple non-linear model incidence prediction. *Statistics in Medicine*, 16(20):2297–2309, 1997.

[98] Dyba T and Hakulinen T. Comparison of different approaches to incidence prediction based on simple interpolation techniques. *Statistics in Medicine*, 19(13):1741–1752, 2000.

[99] Clayton D and Schifflers E. Models for temporal variation in cancer rates I: Age, period and age-cohort models. *Statistics in Medicine*, 6:449–467, 1987.

[100] Boily MS, Pickles M, Alary M, Blanchard J, Moses S, Vickeman P, and Mishra S. What really is a concentrated HIV epidemic and what does it mean for West and Central Africa? Insights from mathematical modeling. *J Acquir Immune Defic Syndr*, 68(Suppl 2):S74–82, 2015.

[101] Stockl H, Karla N, Jacobi J, and Watts C. Is early sexual debut a risk factor for HIV infection among women in sub-Saharan Africa?: A systematic review. *Am J Reprod Immunol*, 69:27–40, 2013.

[102] Mkandawire P, Tenkorang E, and Luginaah IN. Orphan status and time to first sex among adolescents in Northern Malawi. *AIDS and Behavior*, 17(3):939–950, 2012.

[103] Levandowski B, PearsonE, Lunguzi J, and KatengezaH. Reproductive health characteristics of young Malawian women seeking post-abortion care. *AfrJReprodHealth*, 16:253–256, 2012.

[104] Baumgartner J, Waszak Geary C, Tucker H, and Wedderburn M. The influence of early sexual debut and sexual violence on adolescent pregnancy: A matched case-control study in Jamaica. *Int Perspect Sex Reprod Health*, 35:21–28, 2009.

[105] Ma Q, Ono-Kihara M, Cong L, Xu G, Pan X, and Zamani S. Early initiation of sexual activity: a risk factor for sexually transmitted diseases, HIV infection, and unwanted pregnancy among university students in China. *BMC Public Health*, 9(111), 2009.

[106] Kaestle C, Halpern C, Miller W, and Ford C. Young age at first sexual intercourse and sexually transmitted infections in adolescents and young adults. *Am J Epidemiol*, 161:774–780, 2005.

[107] Soares A. Early patterns of sexual activity: Age cohort differences in Australia. *International Journal of STD AIDS*, 14(11):745–752, 2007.

[108] Zhao FH, Tiggelaar SM, Hu SY, Xu LN, Hong Y, and Niyazi M. A multi-center survey of age of sexual debut and sexual behavior in Chinese women: Suggestions for optimal age of human papillomavirus vaccination in China. *Cancer Epidemiology*, 36(4):384–390, 2012.

[109] Stueve A. Early sexual initiation and subsequent sex-related risks among urban minority youth: The Reach for Health Study. *Fam Plann Perspect*, 33:268–75, 2001.

[110] Samb B, Desai N, Nishtar S, Wright A Mendis S, Bekedam H, Hsu J, Martiniuk A, Celletti F, Patel K, Adshead F, McKee M, Evans T, Alwan A, and Etienne C. Prevention and management of chronic disease: a litmus test for health-systems strengtheningin low-income and middle-income countries. *Lancet*, 376:1785–1797, 2010.

[111] Rutstein SO and Rojas G. *Guide to DHS Statistics:Demographic and Health Surveys Methodology.* Springer, New York, 2006.

[112] Rothman K. *Epidemiology:An Introduction.* Oxford University Press, Oxford,UK, 2002.

[113] Crampin AC, Floyd S, Glynn JR, Sibande F, Mulawa D, Nyondo A, Broadbent P, Bliss L, Ngwira B, and Fine PE. Long-term follow-up of HIV-positive and HIV negative individuals in rural Malawi. *AIDS*, 16:1545–1550, 2002.

[114] National Statistical Office (NSO). *Malawi Population and Housing Census 1987, Summary of Final Results*. NSO Malawi, Zomba,Malawi, 1991.

[115] NSO Malawi. 1998 census results. http://nso.malawi.net/data_on_line/demography/census_98/census_results.html, 1998.

[116] National Statistical Office (NSO). *Population and Housing Census:Main Report*. NSO Malawi, Zomba,Malawi, 2008.

[117] National Statistical Office of Malawi. Population projections for Malawi. http://nso.malawi.net.

[118] Misiri H. Estimation of HIV incidence from grouped cross-sectional sero-prevalence data in settings where anti-retroviral therapy is provided. Technical report, College of Medine,University of Malawi, Blantyre,Malawi, 2014.

[119] Misiri HE, Edriss A, Aalen OO, and Dahl FA. Estimation of HIV incidence in Malawi from cross-sectional population-based sero-prevalence data. *Journal of the Interna-tional AIDS Society*, 15(14), 2012.

[120] Podgor MJ and Leske MC. Estimating incidence from age-specific prevalence for irreversible diseases with differential mortality. *Statistics in Medicine*, 5:573–578, 1986.

[121] Efron B. Bootstrap Methods: Another look at the jackknife. *Annals of Statistics*, 7:1–26, 1979.

[122] Hardle W. *Applied Nonparametric Regression Vol 27*. Cambridge University Press, Cambridge, UK, 1990.

[123] Hart JD. *Nonparametric Smoothing and Lack -of- Fit Tests*. Springer, New York,USA, 1997.

[124] Klein JP and Moeschberger ML. *Survival Analysis:Techniques for Censored and Truncated Data*. Springer-Verlag, New York, 1997.

[125] Duchateau L and Janssen P. *The Frailty Model*. Springer, New York,USA, 2007.

[126] Stata Corp. *Statistical Data Analysis(STATA) Special Edition 9*. StatCorp, Texas, USA, 2005.

[127] Archer KL and Lemeshow S. Goodness-of-fit test for a logistic regression model fitted using survey sample data. *The Stata Journal*, 6(1):97–105, 2006.

[128] Misiri H. Risk factors of early sexual debut among men and women - a strong predictor of HIV and sexual risk in Malawi. *Research*, 1(625), 2014.

[129] Segi M and Kurihara M. *Cancer mortality for selected sites in 24 countries (1950-1957)*. Tohoku University of Medicine, Sendai, Japan, 1960.

[130] Doll R, Payne P, and Waterhouse J. *Cancer Incidence in Five Continents: A Technical Report 1966*. Springer, New York,USA, 1966.

[131] Bray F, Guilloux A, Sankila R, and Parkin DM. Practical implications of imposing a new World standard population. *Cancer Causes Control*, 13(2):175–182, 2002.

[132] Dyba T and Hakulinen T. Do cancer predictions work? *European Journal of Cancer*, 44(3):448–453, 2008.

[133] Holford TR. The estimation of age, period and cohort effects for vital rates. *Biometrics*, 39:311–324, 1983.

[134] Clayton D and Schifflers E. Models for temporal variation in cancer rates I: Age -period and age-cohort models. *Statistics in Medicine*, 6:449–467, 1987.

[135] Clayton D and Schifflers E. Models for temporal variation in cancer rates II: Age -period-cohort models. *Statistics in Medicine*, 6:469–481, 1987.

[136] Moller B, Fekjaer H, Hakulinen T, Tryggvadottir L, Storm HH, Talback M, and Haldorsen T. Prediction of cancer incidence in the Nordic countries up to the year 2020. *European Journal of Cancer*, 11:S1–S96, 2002.

[137] Gregson S, Donnely CA, Parker CG, and Anderson RM. Demographic approaches to the estimation of incidence of HIV-1 infection among adults from age-specific prevalence data in stable endemic conditions. *AIDS*, 10:1689–1697, 1996.

[138] Ministry of Health. *HIV and Syphilis Sero-Survey and National HIV Prevalence and AIDS Estimates Report for 2007*. National Aids Commission Malawi, Lilongwe,Malawi, 2007.

[139] Fergusson P, Chikaphupha K, Bongololo G, Makwiza I, Nyirenda L, Chinkhumba J, Aslam A, and Theobald S. Quality of care in nutritional rehabilitation in HIV-endemic Malawi: caregiver perspectives. *Matern Child Nutr*, 6:89–100, 2010.

[140] Ades AE and Nokes DJ. Model age and time specific incidence from seroprevalence:Toxoplasmosis. *American Journal of Epidemiology*, 137(9):1022–1034, 1993.

[141] Benjamini Y and Hochberg Y. Controlling the false discovery rate: A practical approach to multiple testing. *Royal Statistical Society:Series B(Methodology)*, 57(1):289–300, 1995.

[142] Hosmer Jr DW and Lemeshow S. *Applied Logistic Regression*. John Wiley and Sons, New York,USA, 1989.

[143] Atuhaire LK. Application of current-status survival analysis methodology to estimation of age at first sex: Uganda. *The African Statistical Journal*, 12:141–150, 2011.

[144] Sun J. *The Statistical Analysis Of Interval-Censored Failure Time Data*. Springer, New York, USA, 2006.

[145] Berridge DM and Crouchley R. *Multivariate Generalised Linear Models Using R*. CRC Press, Florida,USA, 2011.

[146] RabeHesketh S and Skrondal A. *Multilevel and Longitudinal Modelling Using Stata*. Stata Press, Texas,USA, 2008.

[147] Pfefferman D. The role of sampling weights when modelling survey data. *Internatonal Statistical Review*, 61(1):317–337, 1993.

[148] Misiri H and Edriss A. Age at cancer diagnosis in Malawi. *Global Advanced Journal of Medicine and Medical Sciences*, 1(6):154–162, 2012.

[149] Cvancarova M, Samuelsen SO, Magelssen H, and Foss SD. Reproduction rates after cancer treatment: Experience from the Norwegian Radium Hospital. *Journal of Clinical Oncology*, 27(3), 2009.

[150] Cvancarova M, Fossa SD, Lambert PC, Mller B, and Bray F. Proportion cured models applied to 23 cancer sites in Norway. *International Journal of Cancer*, 132:1700–1710, 2013.

[151] Dudarev AA, Chupakhin VS, and Odland JO. Cancer incidence and mortality in Chukotka, 1997-2010. *International Journal of Circumpolar Health*, 72, 2013.

[152] Yang L, Parkin DM, Li LD, Chen YD, and Bray F. Estimation and projection of the national profile of cancer mortality in China: 19912005. *British Journal of Cancer*, 90:2157–2166, 2004.

[153] Smastuen MC. A statistical approach to cancer survivorship :statistical methods applied on registry data for computing reproduction rates, cure proportion and survival. http://www.duo.uio.no/handle/10852/27996.

[154] Smastuen MC. *A statistical approach to cancer survivorship :statistical methods applied on registry data for computing reproduction rates, cure proportion and survival*. PhD thesis, University of Oslo, 2011.

[155] Dyba T. *Precision of cancer incidence predictions based on Poisson distributed observations*. PhD thesis, University of Helsinki, 2000.

[156] Pallangyo K, Bakari M, and Lyamuya E. Incidence of HIV-1 Infection and associated risk factors in a cohort of police officers in Dar es Salaam, Tanzania. *12 th International AIDS Conference*, 1998.

[157] Bakari M, Lyamuya E, and Mugusi F. The prevalence and incidence of HIV-1 Infection and syphilis in a cohort of police officers in Dar es Salaam, Tanzania. *AIDS*, 14:313–320, 2000.

[158] Boerma JT, Urassa M, and Senkoro K. Spread of HIV infection in a rural area of Tanzania. *AIDS*, 13:1233–1240, 1999.

[159] Borgdorff MW, Barongo LR, and Klokke AH. HIV-1 incidence and HIV-1 associated mortality in a cohort of urban factory workers in Tanzania. *Genitourinary Med*, 71:212–215, 1995.

[160] Killewo JZJ, Sandstrom A, and Bredberg U. Incidence of HIV-1 Infection among adults in the Kagera Region of Tanzania. *Int J Epidemiol*, 22, 1993.

[161] Mbizvo MT, Machekano R, McFarland W, Mashayamombe S, Ray S, Basset M, and Katzenstein D. HIV seroincidence and correlates of seroconversion in a cohort of male factory workers in Harare, Zimbabwe. *AIDS*, 10:895–901, 1996.

[162] Lopman BA, Garnett GP, Mason PR, and Gregson S. Individual level injection history: A lack of association with HIV incidence in rural Zimbabwe. *PLoS Med*, 2:142–146, 2005.

[163] Kalanda BF. Life skills and reproductive health education changes behaviour in students and teachers: Evidence from Malawi. *Educational Research and Reviews*, 5(4):169–174, 2010.

[164] Chatterji M, Murray NJ, London DR, and Anglewicz PA. The social and cultural determinants of exchange of sex for money or gifts among young women in Kenya and Malawi. *International Conference on AIDS*, 14, 2002.

[165] Luke N. Age and economic asymmetries in the sexual relationships of adolescent girls in Sub-Saharan Africa. *Studies in Family Planning*, 34(2):67–86, 2003.

[166] Luke N and Kurz KM. Cross-generational and transactional sexual relations in sub-saharan africa:prevalence of behavior and implications for negotiating safer sexual practices. http://fr.allafrica.com/download/resource/main/main/id/00010860.pdf, 2010.

[167] Tavory I and Swidler A. Condom semiotics: Meaning and condom use in rural Malawi. *American Sociological Review*, 74:171–179, 2009.

[168] Wodi BE. Gender issues in HIV/AIDS epidemiology in Sub-Saharan Africa. *Gender Issues in HIV/AIDS*, 2:1–10, 2005.

[169] Kaler A. "My girlfriends could fill a Yanu-Yanu Bus":Rural Malawian men's claims about their own serostatus. http://www.demographic-research.org/special/1/11/S1-11.pdf.

[170] Anglewicz P and Kohler HP. Overestimating HIV infection: The construction and accuracy of subjective probabilities of HIV infection in rural Malawi. *DemogrRes*, 20(6):65–96, 2009.

[171] Traeen B and Samuelsen SO. Sweet 16 and never been kissed? Experiences from a longitudinal Norwegian study. *Electronic Journal of Human Sexuality*, 10(6), 2007.

[172] Fatusi A and Blum R. Predictors of early sexual initiation among a nationally representative sample of Nigerian adolescents. *BMC Public Health*, 8(136), 2008.

[173] McGrath N, Nyirenda M, Hosegood V, and Newell M. Age at first sex in rural South Africa. *Sex Transm Infect.*, 85:i49–i55, 2009.

[174] Ntonya G. Ending child marriages in malawi. http://www.safeguardyoungpeople.org/endingchildmarriagesinmalawi/.

[175] UNICEF. The state of the worlds children 2015: Executive summary. http://www.unicef.org/publications/files/SOWC_2015_Summary_and_Tables.pdf.

[176] Chimjeka R. Snatched from the jaws:Malawi abolishes 600,000 child marriages. *Weekend Nation*, 19(52):1–4, 2 January 2016.

[177] Human Rights Watch. Malawi: New marriage law can change lives. https://www.hrw.org/news/2015/04/17/malawinewmarriagelawcanchangelives.

[178] Zuma K, Mzolo T, and Makonko E. Determinants of age at sexual debut and associated risks among South African youths. *African Journal of AIDS Research*, 10:189–94, 2011.

[179] Pettifor A, Levandowski B, MacPhail C, Padian N, Cohen M, and Rees H. Keep them in school: The importance of education as a protective factor against HIV infection among young South African women. *Int J Epidemiol.*, 37:1266–1273, 2008.

[180] Moore A, Awusabo-Asare K, Madise N, John-Langba J, and Kumi-Kyereme A. Coerced first sex among adolescent girls in sub-Saharan Africa: Prevalence and context. *Afr J Reprod Health*, 11:62–82, 2007.

[181] UNAIDS. Global report of the aids epidemic. www.unaids.org.

[182] Misiri H and Muula AS. Attitudes towards premarital testing of Human Immunodeficiency Virus among Malawians. *Croat Med*, 45(1):84, 2004.

[183] Msyamboza KP, Dzamalala C, Mdokwe C, Kamiza S, Lemerani M, Dzowela T, and Kathyola D. Burden of cancer in Malawi; common types,incidence and trends: National population-based cancer registry. *BMC Research Notes*, 5(149), 2012.

[184] Cooke A. *CANREG 3 Manual*. International Agency for Research on Cancer (IARC), Lyon, France, 1998.

Paper I

Journal of the International AIDS Society

BioMed Central
The Open Access Publisher

This Provisional PDF corresponds to the article as it appeared upon acceptance. Fully formatted PDF and full text (HTML) versions will be made available soon.

Estimation of HIV incidence in Malawi from cross-sectional population-based sero-prevalence data

Journal of the International AIDS Society 2012, **15**:14 doi:10.1186/1758-2652-15-14

Humphrey E Misiri (hmisiri@gmail.com)
Abdi Edriss (abdikhalil@yahoo.com)
Odd O Aalen (o.o.aalen@medisin.uio.no)
Fredrik A Dahl (fredrik.a.dahl@ahus.no)

ISSN	1758-2652
Article type	Research
Submission date	22 June 2011
Acceptance date	14 March 2012
Publication date	14 March 2012
Article URL	http://www.jiasociety.org/content/15/1/14

This peer-reviewed article was published immediately upon acceptance. It can be downloaded, printed and distributed freely for any purposes (see copyright notice below).

Articles in *Journal of the International AIDS Society* are listed in PubMed and archived at PubMed Central.

For information about publishing your research in *Journal of the International AIDS Society* or any BioMed Central journal, go to

http://www.jiasociety.org/info/instructions/

For information about other BioMed Central publications go to

http://www.biomedcentral.com/

© 2012 Misiri *et al.* ; licensee BioMed Central Ltd.
This is an open access article distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by/2.0), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Estimation of HIV incidence in Malawi from cross-sectional population-based sero-prevalence data

Humphrey Misiri[1*,2]
[*]Corresponding author
Email: hmisiri@gmail.com

Abdi Edriss[4]
Email: abdikhalil@yahoo.com

Odd O Aalen[1]
Email: o.o.aalen@medisin.uio.no

Fredrik A Dahl[1,3]
Email: fredrik.a.dahl@ahus.no

[1] Department of Biostatistics, University of Oslo, Sognsveien 9, Oslo, Norway

[2] Community Health Department, College of Medicine, University of Malawi, Blantyre, Malawi

[3] HØKH, Akershus University Hospital, Lørenskog, Norway

[4] Bunda College, University of Malawi, Lilongwe, Malawi

Abstract

Background

Incidence is a better measure than prevalence for monitoring AIDS, but it is not often used because longitudinal HIV data from which incidence can be computed is scarce. Our objective was to estimate the force of infection and incidence of HIV in Malawi using cross-sectional HIV sero-prevalence data from the Malawi Demographic and Health Survey conducted in 2004.

Methods

We formulated a recurrence relation of population prevalence as a function of a piecewise-constant force of HIV infection. The relation adjusts for natural and HIV-induced mortality. The parameters of the recurrence relation were estimated using maximum likelihood, and confidence intervals of parameter estimates were constructed by bootstrapping. We assessed the fit of the model using the Pearson Chi-square goodness of fit test. We estimated population incidence from the force of infection by accounting for the prevalence, as the force of infection applies only to the HIV-negative part of the population.

Results

The estimated HIV population incidence per 100,000 person-years among men is 610 for the 15–24 year age range, 2700 for the 25–34 group and 1320 for 35–49 year olds. For females, the estimates are 2030 for 15–24 year olds, 1710 for 25–34 year olds and 1730 for 35–49 year olds.

Conclusions

Our method provides a simple way of simultaneously estimating the incidence rate of HIV and the age-specific population prevalence for single ages using population-based cross-sectional sero-prevalence data. The estimated incidence rates depend on the HIV and natural mortalities used in the estimation process.

Background

AIDS has had a devastating impact on communities in sub-Saharan Africa. In Malawi, the HIV prevalence is approximately 11% [1] for the whole population. Many developing countries (including Malawi) use prevalence from sentinel sero-surveys for monitoring the AIDS epidemic [2], although for such purposes, the HIV incidence is considered a more relevant measure [3]. The force of infection (FOI) is often used in statistical models [4-8] for estimating incidence rate of infection [4], including the present analysis. The FOI gives the probability that an HIV-negative person contracts the virus during a year, and is slightly higher than population-level incidence, which evaluates the number of new cases relative to the total population size.

The present analysis uses data from a single Malawian demographic and health survey (DHS). If the age-dependent prevalence had been measured several times, it would facilitate incidence estimation [9]. Our situation, with cross-sectional current status data sampled at a single point in time, is common in African countries, and makes incidence estimation more problematic. The key assumption needed to estimate incidence in this situation is that the prevalence be stationary (time homogeneous) in each age group.

The contribution of the present article has two parts. First, the results give FOI and population incidence estimates for HIV in Malawi, which is novel. Second, while our modelling assumptions of endemic equilibrium and age-dependent prevalence are not new, our analysis provides some novelties compared with earlier research: through an intuitive conditional probability formulation of the mortality adjusted prevalence, we compute the total likelihood function, which enables us to estimate our model parameters simultaneously. This enables us to estimate a model with few degrees of freedom, which is desirable because the prevalence data is too sparse for estimating the FOI on a fine grid. We apply a bootstrap technique to estimate confidence intervals and also run sensitivity analyses. Finally, we convert FOI estimates to population incidence rates by compensating for the estimated age-dependent prevalence.

Methods

Data sources

The HIV sero-prevalence data for this paper are from the Malawi Demographic and Health Survey 2004 (MDHS2004). This was a cross-sectional country-wide study conducted in Malawi in 2004 [10]. The DHS collected data on different topics, including HIV, from a multistage sample that spanned the whole country. From this multistage cluster sample, a group of women aged 15 to 49 years and men aged 15 to 54 were tested for HIV. The HIV status of individuals was determined using Enzygnost, Vironostika and Western Blot tests. The results of the HIV testing are shown in Table 1.

Table 1 Age- and sex-specific HIV test results for men aged 15–54 years and women aged 15–49 years, Malawi

	Males			Females			All genders		
Age (years)	*HIV+*	*HIV-*	*Total*	*HIV+*	*HIV-*	*Total*	*HIV+*	*HIV-*	*Total*
15	1	76	77	1	95	96	2	171	173
16	0	91	91	4	116	120	4	207	211
17	1	90	91	0	81	81	1	171	172
18	0	109	109	8	115	123	8	224	232
19	0	87	87	7	118	125	7	205	212
20	2	76	78	20	136	156	22	212	234
21	5	99	104	21	98	119	26	197	223
22	4	80	84	17	165	182	21	245	266
23	5	68	73	18	95	113	23	163	186
24	1	85	86	28	93	121	29	178	207
25	8	95	103	21	110	131	29	205	234
26	5	84	89	17	85	102	22	169	191
27	18	75	93	21	82	103	39	157	196
28	16	71	87	19	89	108	35	160	195
29	5	81	86	13	70	83	18	151	169
30	17	78	95	19	84	103	36	162	198
31	13	57	70	13	48	61	26	105	131
32	14	71	85	11	66	77	25	137	162
33	11	31	42	12	56	68	23	87	110
34	8	55	63	22	57	79	30	112	142
35	8	46	54	14	60	74	22	106	128
36	12	57	69	13	49	62	25	106	131
37	7	20	27	8	30	38	15	50	65
38	6	37	43	8	47	55	14	84	98
39	8	22	30	12	36	48	20	58	78
40	10	45	55	12	60	72	22	105	127
41	7	38	45	10	33	43	17	71	88
42	6	48	54	9	40	49	15	88	103

43	4	18	22	5	32	37	9	50	59
44	11	41	52	8	36	44	19	77	96
45	2	26	28	7	47	54	9	73	82
46	7	26	33	9	31	40	16	57	73
47	3	20	23	5	27	32	8	47	55
48	2	29	31	4	29	33	6	58	64
49	1	12	13	5	27	32	6	39	45
50	3	24	27				3	24	27
51	3	30	33				3	30	33
52	6	30	36				6	30	36
53	3	17	20				3	17	20
54	0	16	16				0	16	16
Total	*243*	*2161*	*2,404*	*421*	*2443*	*2,864*	*664*	*4604*	*5268*

The natural mortality rates were extracted from the MDHS2004 report [10]. These estimates were computed from the Malawi Demographic and Health Survey 1992 (MDHS1992) data [11] using methods explained in detail elsewhere [12]. The mortality estimates for 1992 estimates are representative of natural mortality rates because by 1992, HIV prevalence was very low in Malawi and mortality was mainly due to other causes. About 80% of the Malawi population live in rural areas [13]. Furthermore, the HIV mortality rates used in this paper are from a longitudinal study conducted in rural Malawi by Crampin *et al.* (2002) [14]. These are estimates of rates of mortality among HIV-positive individuals. Since these rates were estimated from data from typical rural Malawi, where the majority of the country's people live, the rates are assumed to be representative of national HIV mortality rates. The natural mortality rates and HIV mortality rates are given for age groups with a width of five years thus: 15–19, 20–24,25–29,30–34,35–39,40–44,45–49 (Table 2). The natural mortality rate for males aged 50–54 years was not available from the source, and has been estimated through linear regression on age.

Table 2 Age- and sex-specific natural and AIDS mortality rates for Malawi

Age group index		Natural mortality rates (per 100,000 person-years)*		HIV mortality rates (per 100,000 person-years)#
(k)	Age group	Men	Women	for men and women
1	15–19	0.0038	0.0053	0.0471
2	20–24	0.0041	0.0036	0.0593
3	25–29	0.0068	0.0068	0.0675
4	30–34	0.0084	0.0072	0.1354
5	35–39	0.0076	0.009	0.1354
6	40–44	0.0101	0.0089	0.1427
7	45–49	0.0097	0.0096	0.1427
8	50+	0.0097	0.0096	0.2339

Source: * Malawi Demographic and Health Survey 1992 # Crampin *et al.* (2002)

Model formulation

Our model is based on the following assumptions: the spread of HIV is in endemic equilibrium, so that the prevalence in each age group is stable over time; mortality of HIV-positive people depends only on age; and an infected person remains HIV positive until death. We also assume the FOI to be constant within each age group, and use three groups; 15–24, 25–34 and 35+, associated with FOI values $\lambda_{15-24}\lambda_{25-34}\lambda 35+$. We use this relatively coarse discretization in order to keep the number of parameters low, which facilitates robust parameter estimation. Let μ_a and φ_a denote the natural (HIV-negative) and HIV-positive mortality rates at age *a*, respectively. Let π_a denote the population prevalence at age *a*, and let *A* be the maximum age in the prevalence data (54 for males and 49 for females). Also, let λ_a refer to the λ that corresponds to the age *a*.

Furthermore, assuming that the underlying population is so large that a deterministic equation can describe the dynamic of the population prevalence, the population prevalence at age a + 1 years is given by the recurrence relation:

$$\pi_{a+1} = \frac{\pi_a(1-\varphi_a)+(1-\pi_a)(1-\mu_a)\lambda_a}{\pi_a(1-\varphi_a)+(1-\pi_a)(1-\mu_a)}, \qquad (1)$$

Where a = 15,16,…,A-1.

This relation gives the population prevalence of HIV at any age from 15 up to *A* years as a function of the four-dimensional parameter vector $\theta = (\pi_{15},\lambda_{15-24},\lambda_{25-34},\lambda_{35+})$.

Equation 1 is easiest explained as an application of conditional probability computations: consider a random person born *a* years ago (i.e. his age is *a*, assuming he is alive), so that the following represents his conditional probability of being positive at age *a + 1*, given that he is alive at that age:

$$P\left(positive_{a+1} \middle| alive_{a+1}\right) = \frac{P\left(positive_{a+1} \cap alive_{a+1}\right)}{P\left(alive_{a+1}\right)}$$

The event in the numerator (being positive and alive) can be split in two disjoint events of surviving after being alive and positive at age *a*, and surviving after being alive and negative at age *a* while contracting HIV during that year:

$$P(positive_{a+1} \cap alive_{a+1}) = P(positive_a \cap alive_a).(1-\varphi_a) + P(negative_a \cap alive_a).(1-\mu_a).\lambda_a$$

A similar decomposition of the denominator gives:

$$P(positive_{a+1} \mid alive_{a+1}) = \frac{P(positive_a \cap alive_a).(1-\varphi_a) + P(negative_a \cap alive_a).(1-\mu_a).\lambda_a}{P(positive_a \cap alive_a).(1-\varphi_a) + P(negative_a \cap alive_a).(1-\mu_a)}$$

When we divide by $P(alive_a)$ in the numerator and denominator, we recognize this as

$$P(positive_{a+1} \mid alive_{a+1}) = \frac{P(positive_a \mid alive_a).(1-\varphi_a) + P(negative_a \mid alive_a).(1-\mu_a).\lambda_a}{P(positive_a \mid alive_a).(1-\varphi_a) + P(negative_a \mid alive_a).(1-\mu_a)}$$

This is Equation 1, because population prevalence in an age group is the probability of a random person of that age being HIV positive, given that he is alive: $\pi_a = P(positive_a \mid alive_a)$. The denominators can throughout the computation be assumed to be non-zero, as otherwise the total population would be dead.

We will denote the sample size at age *a* by n_a, and the number of HIV positives at age *a* by y_a At each age *a*, a sampled individual has the probability π_a of being HIV positive. The sample sizes n_a are considered deterministic, and the events of different individuals being HIV positive are assumed to be independent. So, the number of HIV positives at age *a* years follows a binomial distribution with parameters n_a and π_a. The corresponding probability mass function, *f,* is given by the equation:

$$f(y_a) = \binom{n_a}{y_a} \pi_a^{y_a} (1-\pi_a)^{n_a - y_a}, y_a = 0,1,2,\ldots n_a.$$

The log-likelihood function is thus:

$$\ell(\theta \mid y_a) = K + \sum_{a=15}^{A} y_a.\log(\pi_a) + \sum_{a=15}^{A} (n_a - y_a)\log(1-\pi_a)$$

where K is a constant independent of θ.

The maximum likelihood estimate of θ was computed using the **nlm** function in R [15]. This function uses a version of the Newton–Raphson optimization algorithm. In order to test the reliability of the **nlm** function, we also implemented a random search optimization based on incremental random steps of variable length. It converged toward the same solution as the **nlm** function.

Smoothed estimates of age-specific prevalence were computed using the Nadaraya-Watson estimator. These nonparametric estimates were plotted on the same graph with empirical and model-based estimates.

Goodness of fit

The goodness of fit of our estimation was assessed using the Pearson Chi-square statistic [16]. Let p_a denote the estimated HIV prevalence at age *a* and *t* denote the number of parameters in the model. Let $e_a = (y_a - n_a p_a)$ be the residual at age *a* years. Then $\chi^2 = \sum_{a=15}^{A} \frac{e^2}{[n_a p_a (1-p_a)]}$ gives the Pearson Chi-square statistic with A–14–t = A–18 degrees of freedom.

Confidence intervals and sensitivity analysis

The R-function nlm() gives the Hessian matrix of the log-likelihood. We estimate the standard error of the parameters as the square root of the diagonal elements of the inverse of the Hessian. We estimated the 95% confidence intervals as the point estimates + −1.96 times the standard errors, which is the standard procedure. We also implemented a bootstrap scheme, where we resampled 10,000 data sets with replacement, which produced virtually identical confidence intervals as the Hessian-based method.

We also conducted sensitivity analyses: we increased both mortality rates simultaneously by 2.5%, 5%, 7% 10%, 15% and 20% and so we had six sets (sets of two rates each: HIV and non-HIV mortality rates) of increased mortality rates. We also decreased both mortality rates by 2.5%, 5%, 7%, 10%, 15% and 20%, and had six sets (each of two: HIV and non-HIV rates) of decreased mortality rates. We therefore had a total of 12 of mortality rates. Using each of the 12 sets of mortality rates, the FOI and corresponding 95% bootstrap intervals were estimated.

Force of infection and incidence

The FOI (λ) is the probability that an uninfected person gets infected during a 12-month period, which may be the most relevant incidence parameter for an individual. However, the age-wise incidence rate for the total population (including HIV-positive people) may be more useful for healthcare planning purposes. We therefore give corresponding population incidence (I) estimates based on the relation: $I = \lambda(1-\pi)$. For an age group G, we therefore need to compute the estimated group prevalence π_G as a weighted average: $\pi_G = \frac{\sum_{a \in G} \pi_a N_a}{\sum_{a \in G} N_a}$,

where N_a is the size of the Malawian population at age a. In this way, we computed the incidence estimates: $I_{15-24} = \lambda_{15-24}(1-\pi_{25-34})$, $I_{25-34} = \lambda_{25-34}(1-\pi_{25-34})$ and $I_{35+} = \lambda_{35+}(1-\pi_{35+})$ for males and females, expressed as the expected number of new cases a year, per 100,000 inhabitants.

Results

We fitted models for males and females separately. The Pearson Chi-square goodness of fit statistics of these separate analyses showed acceptable fit for both males (p = 0.930) and females (p = 0.564). The FOI and population incidence rate estimates with confidence intervals for males and females are given in Table 3.

Table 3 Estimates of FOI and incidence (per 100,000 person-years), by age and sex

	Age group	Estimated FOI	95% confidence interval		Estimated incidence	95% confidence interval	
Gender			Lower	Upper		Lower	Upper
Males	15–24	0.0063	0.0039	0.0089	616	375	870
	25–34	0.0312	0.0225	0.0366	2705	1946	3172

	35–54	0.0158	0.0143	0.0281	1336	1209	2374
Females	15–24	0.0225	0.0186	0.0263	2031	1686	2377
	25–34	0.0212	0.0140	0.0284	1712	1131	2294
	35–49	0.0211	0.0133	0.0289	1732	1090	2375

The estimate of the model parameter of prevalence at age 15 was $\pi_{15} = 0.0035$ for males and $\pi_{15} = 0.0069$ for females. The estimated prevalence functions are shown in Figure 1, plotted together with empirical point estimates from the data. In the figure, we also included a Nadaraya-Watson non-parametric estimate of the prevalence (stapled curve), which closely resembles the model curve.

Figure 1 Estimated age-specific sero-prevalence curves for men and women

For each of the eight sets of the mortality rates investigated in the sensitivity analysis, different estimates of the FOI and prevalence were obtained (Additional file 1: Table S1 and Table S2). When the mortality rates were decreased or increased by a certain percentage, the FOI estimates also decreased or increased accordingly.

Discussion

A key assumption in our model is that of an endemic equilibrium of HIV, giving a stationary prevalence in each age and sex group. It is necessary because we have access to prevalence data from only one point in time, and for the same reason, the assumption is not testable from our data. It is reasonable, however, because the virus has been present in Malawi for more than 20 years [17]. In fact, HIV is endemic in Malawi [17, 18]. Notably, our model does not assume that the age distribution in the population is stationary, which would not have been reasonable. We assume that the mortality of HIV-positive people depends only on age, and so we do not model the progression of AIDS. This is a simplification. However, the average lifespan of HIV-infected people is reasonably modelled. Also, an HIV-positive person's time since infection will correlate with his age, which is included in the model.

Our approach has similarities with several direct approaches for estimating the incidence of HIV from cross-sectional sero-prevalence data [19]. In particular, our assumptions closely match those of Saidel *et al.* (1996) [2], which is based on the method of Podgor *et al.* (1983) [20]. Their method relies on estimating the incidence rate for one age interval at a time, using estimated prevalence at the beginning and end of the interval in the computation. In a situation with large prevalence samples, this is possible. With our relatively sparse prevalence data, we need to estimate fewer incidence parameters.

Our method estimates the parameter values simultaneously, and achieves more efficient estimation by utilizing the prevalence observations in each age cohort. The model described by Gregson *et al.* (1996) [21] has similarities with ours, and they also model the incidence with few degrees of freedom. However, we find their model less intuitive and steeped in mathematical language that is less accessible to epidemiologists. Also, they appear to disregard the natural mortality of HIV-negative persons. Note that other approaches for estimating the FOI based on Muench's catalytic model and its modifications [22-24] are unreasonable due to their assumption of negligible disease-induced mortality. The fact that our models achieve adequate Chi-square fit gives some indication that they are sound.

The estimated FOI for the different sex and age groups varies by around 0.02, which means that an HIV-negative adult will have around 2% risk of contracting the virus within a year. This is a high number, which shows that a 15-year-old sero-negative person has an approximately 50% chance of contracting HIV before the age of 50 years. Our estimates are consistent with the observation of Hallet *et al.* (2008) that "Even in settings where HIV has reached high endemic levels, incidence rates are typically less than ~4%" [3]. Compared with estimates of incidence from longitudinal Malawi data collected from Blantyre pregnant women and Nchalo workers in 1995 and 1997 [25], our estimates are smaller on average. This is not strange considering that both the Blantyre pregnant women and Nchalo workers were cohorts of high-risk men and women and so not representative, unlike the MDHS2004 sample from which the data for this paper comes.

Corresponding estimates of incidence from Zimbabwe are 2.34% for women in the 15–24 year age group [26], 2.3% to 2.8% for men and 1.51% for women in the 25–34 age group [26, 27] and 1.5% to 2.2% for men aged at least 35 years [26]. From Tanzania, corresponding estimates of HIV incidence are 0.47–0.68% for men and 2.38–2.5% for women in the 15–24 age group [28-30], 2.6–3.09% for men and 1.53–2.5% for women in the 25–34 age group [28, 31][29,32] and 1.09–1.75% for men in the 35–54 age group [29, 31, 32]. Both estimates from Zimbabwe and Tanzania are within the 95% bootstrap intervals of our estimates for each age group. Our estimates are therefore comparable to those of neighbouring countries. For Zimbabwe, estimates for the 25–49 and 15–24 age ranges among men and women respectively agree with our estimates. Our estimates are therefore comparable to those of neighbouring countries.

The estimated FOI differs considerably for males and females in that males on average contract the virus at an older age. This is compatible with the fact that females on average start sexual activities earlier than males in Malawi. The model also gives very low HIV prevalence for 15 year olds. This is reasonable as few are sexually active before this age, and those born with HIV are mostly diseased before the age of 15 years in the absence of antiretrovirals (ARVs). The estimates for population incidence are similar to the FOI estimates. The difference is proportional to the prevalence, so it is smaller for the lowest age group.

Conclusions

Our analysis is based on population-based HIV sero-prevalence data from 2004, and since then, provision of antiretroviral drugs has been rolled out to many parts of Malawi. For Malawian sub-populations that currently have access to ARVs, our estimates of incidence rate are likely to be less accurate. However, for areas in Malawi where, for some reason, people still do not have access to antiretroviral drugs, our estimates are likely to be valid today.

Our method is best suited for areas or countries where mortality data is available, but access to antiretroviral therapy is a problem. However, for areas where HIV patients have access to antiretroviral therapy, our method can be easily extended to adjust for provision of ARVs as long as survival data of patients on ARVs is available. There is therefore a need for continued collection of not only population-based HIV sero-prevalence data, but also survival data of HIV patients on ARVs. .

Competing interests

The authors do not have any competing interests.

Authors' contributions

OA conceived the study and participated in its coordination. HM obtained the data, performed data analysis and drafted the manuscript. FAD suggested the mathematical model. OA, AE and FAD critically revised the manuscript. All authors read and approved the final manuscript.

Acknowledgements

We are very grateful to Macro International for granting us permission to use the MDHS2004 data. Humphrey Misiri acknowledges support from the Norwegian Programme for Development, Research and Education. The study was also partially supported by funds from the Research Council of Norway through contract/grant number: 177401/V50.

References

1. National Statistical Office (NSO), O R C Macro: *Malawi Demographic and Health Survey 2010*. Zomba: NSO and ORC Macro; 2010.

2. Saidel T, Sokal D, Rice J, Buzingo T, Hassig S: **Validation of a method to estimate age-specific Human Immunodeficiency Virus (HIV) incidence rates in developing countries using population-based seroprevalence data.** *Am J Epidemiol* 1996, **144:**214–223.

3. Hallett TB, Zaba B, Todd J, Lopman B, Mwita W, Biraro S, Gregson S, Boerma T: **Estimating Incidence from prevalence in generalised HIV epidemics:methods and validation.** *PLoS Med* 2008, **5(4):e80**.

4. Garnett GP: **An introduction to mathematical models in sexually transmitted disease epidemiology.** *Sex Transm Infect* 2002, **78:**7–12.

5. Brandgo-Fijo S, Campbell-Lendrum D, Davies DR, Brito MEF, Shaw JJ, Davies CR: **Epidemiological surveys confirm an increasing burden of cutaneous leishmaniasis in north-east.** *Trans Roy Soc Trop Med Hygiene* 1999, **93:**488–494.

6. Colugnati FAB, Staras SAS, Dollard SCD, Cannon MJ: **Incidence of cytomegalovirus infection among the general population and pregnant women in the United States.** *BMC Infectious Diseases* 2007, **7:71**.

7. Conn PB, Cooch EG, Caley P: **Accounting for detection probability when estimating force-of-infection from animal encounter data.** *J Ornithol* 2010.

8. Abebe A, Nokes DJ, Dejene A, Enquselassie F, Messele T, Cutts FY: **Seroepidemiology of hepatitis B virus in Addis Ababa, Ethiopia: transmission patterns and vaccine control.** *Epidemiol Infect* 2003, **131**:757–770.

9. Marschner IC: **A method for assessing age-time disease incidence using serial prevalence data.** *Biometrics* 1997, **53:**1384–1398.

10. National Statistical Office (NSO), O R C Macro: *Malawi Demographic and Health Survey 2004*. Zomba: NSO and ORC Macro; 2005.

11. National Statistical Office (NSO), O R C Macro: *Malawi Demographic and Health Survey 1992*. Zomba: Malawi Government; 1992.

12. **Online Guide to DHS Statistics** [www.measuredhs.com/help/Datasets/whtdhtml.html]

13. National Statistical Office of Malawi NSO: *2008 Population and Housing Census:Main Report.* Zomba,Malawi.: National Statistical Office of Malawi; 2008.

14. Crampin AC, Floyd S, Glynn JR, Sibande F, Mulawa D, Nyondo A, Broadbent P, Bliss L, Ngwira B, Fine PE: **Long-term follow-up of HIV-positive and HIVnegative individuals in rural Malawi.** *AIDS* 2002, **16:**1545–1550.

15. **The R Project for Statistical Computing** [http://www.R-project.org/]

16. Whitaker HJ, Farrington CP: **Estimation of infectious disease parameters from serological survey data:the impact of regular epidemics.** *Stat Med* 2004, **23**:2429–2443.

17. Ministry of Health: *HIV and Syphilis Sero-Survey and National HIV Prevalence and AIDS Estimates Report for 2007.* Lilongwe,Malawi.: National Aids Commission Malawi; 2008.

18. Fergusson P, Chikaphupha K, Bongololo G, Makwiza I, Nyirenda L, Chinkhumba J, Aslam A, Theobald S: **Quality of care in nutritional rehabilitation in HIV-endemic Malawi: caregiver perspectives.** *Matern Child Nutr* 2010, **6:**89–100.

19. **Estimation of the Incidence of Disease with the Use of Prevalence Data** [https://eldorado.tu-dortmund.de/bitstream/2003/4949/1/99_12.pdf]

20. Podgor MJ, Leske MC, Ederer E: **Incidence estimates for lens changes, macular changes, open-angle glaucoma and diabetic retinopathy.** *Am J Epidemiol* 1983, **118:**206–212.

21. Gregson S, Donnelly CA, Parker CG, Anderson RM: **Demographic approaches to the estimation of incidence of HIV-1 infection among adults from age-specific prevalence data in stable endemic conditions.** *AIDS* 1996, **10:**1689–1697.

22. Shkedy Z, Aerts M, Molenberghs G, Beutels P, Van Damme P: **Modelling age-dependent force of infection from prevalence data using fractional polynomials.** *Stat Med* 2006, **25:**1577–1591.

23. Hens N, Aerts M, Faes C, Shkedy Z, Lejeune O, Van Damme P, Beutels P: **Seventy-five years of estimating the force of infection from current status data.** *Epidemiol Infect* 2010, **138:**802–812.

24. Grenfell BT, Anderson RM: **The estimation of age-related rates of infection from case notifications and serological data.** *J Hyg* 1985, **95:**419–436.

25. Kumwenda NI, Taha TE, Hoover DR, Markakis D, Liomba GN, Chiphangwi JD, Celentano DD: **HIV-1 incidence among male workers at a sugar estate in rural Malawi.** *J Acquir Immune Defic Syndr* 2001, **27:**202–208.

26. Lopman BA, Garnett GP, Mason PR, Gregson S: **Individual level injection history: a Lack of association with HIV incidence in rural Zimbabwe.** *PLoS Med* 2005, **2:**142–146.

27. Mbizvo MT, Machekano R, McFarland W, Mashayamombe S, Ray S, Basset M, Katzenstein D: **HIV seroincidence and correlates of seroconversion in a cohort of male factory workers in Harare, Zimbabwe.** *AIDS* 1996, **10:**895–901.

28. Killewo JZJ, Sandstrom A, Bredberg U: **1993 Incidence of HIV-1 Infection among adults in the Kagera Region of Tanzania.** *Int J Epidemiol* 1993, **22:**528–536.

29. Borgdorff MW, Barongo LR, Klokke AH: **HIV-1 Incidence and HIV-1 Associated Mortality in a Cohort of Urban Factory Workers in Tanzania.** *Genitourinary Med* 1995, **71:**212–215.

30. Boerma JT, Urassa M, Senkoro K: **Spread of HIV infection in a rural area of Tanzania**. *AIDS* 1999, **13:**1233–1240.

31. Bakari M, Lyamuya E, Mugusi F: **The prevalence and incidence of HIV-1 Infection and syphilis in a cohort of police officers in Dar es Salaam, Tanzania.** *AIDS* 2000, **14:**313–320.

32. Pallangyo K, Baraki M, Lyamuya E: **Incidence of HIV-1 Infection and Associated Risk Factors in a Cohort of Police Officers in Dar es Salaam, Tanzania.** *12th International AIDS Conference; Geneva 1998*. 1998.

Additional files

Additional_file_1 as DOC
Additional file 1 : Table S1. Results of sensitivity analysis for men. Table S2. Results of sensitivity analysis for women.

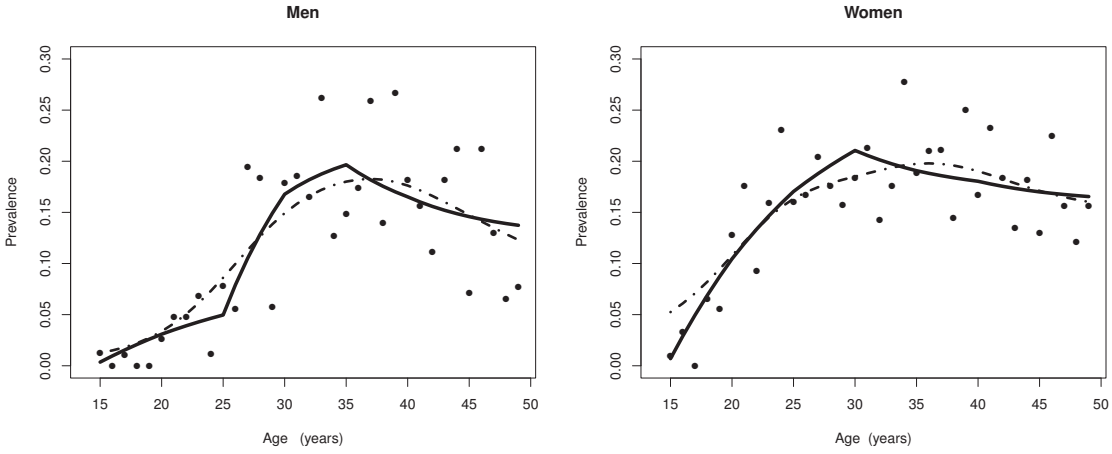
Men
Women
Prevalence
0.00
0.05
0.10
0.15
0.20
0.25
0.30
15
20
25
30
35
40
45
50
Age (years)

Figure 1

Additional files provided with this submission:

Additional file 1: Misiri Suppl file 1_AcceptedChanges.doc, 123K
http://www.jiasociety.org/imedia/4419772956895264/supp1.doc

Additional Files

Table 4. Results of sensitivity analysis for men

% change in mortality rates	Age(years)	Estimated FOI	95% confidence interval	
			Lower limit	Upper limit
2.5% decrease	15-24	0.0063	0.0038	0.0088
	25-34	0.0274	0.0207	0.0342
	35-54	0.0157	0.0095	0.0219
5% decrease	15-24	0.0063	0.0038	0.0088
	25-34	0.0272	0.0205	0.0339
	35-54	0.0151	0.0089	0.0212
7% decrease	15-24	0.0063	0.0038	0.0088
	25-34	0.0270	0.0203	0.0337
	35-54	0.0146	0.0085	0.0207
10% decrease	15-24	0.0063	0.0038	0.0088
	25-34	0.0267	0.0201	0.0334
	35-54	0.0139	0.0079	0.0199
15% decrease	15-24	0.0063	0.0038	0.0087
	25-34	0.0262	0.0197	0.0328
	35-54	0.0126	0.0068	0.0185
20% decrease	15-24	0.0064	0.0038	0.0087
	25-34	0.0295	0.0192	0.0323
	35-54	0.0212	0.0057	0.0172
No change in mortality rates	15-24	0.0063	0.0039	0.0089
	25-34	0.0312	0.0225	0.0366
	35-54	0.0158	0.0143	0.0281
2.5% increase	15-24	0.0063	0.0038	0.0088
	25-34	0.0279	0.0211	0.0347
	35-54	0.0170	0.0106	0.0234
5% increase	15-24	0.0063	0.0038	0.0089
	25-34	0.0281	0.0213	0.0350
	35-54	0.0175	0.0111	0.0240
7% increase	15-24	0.0063	0.0038	0.0089

	25-34	0.0283	0.0214	0.0352
	35-54	0.0180	0.0115	0.0245
10% increase	15-24	0.0064	0.0038	0.0089
	25-34	0.0286	0.0217	0.0355
	35-54	0.0188	0.0122	0.0254
15% increase	15-24	0.0062	0.0037	0.0086
	25-34	0.0317	0.0245	0.0388
	35-54	0.0118	0.0066	0.0170
20% increase	15-24	0.0064	0.0039	0.0089
	25-34	0.0295	0.0225	0.0366
	35-54	0.0212	0.0143	0.0281

Table 5. Results of sensitivity analysis for women

% change in mortality rates	Age(years)	Estimated FOI	95% confidence interval	
			Lower limit	Upper limit
2.5% decrease	15-24	0.0219	0.0181	0.0256
	25-34	0.0152	0.0081	0.0223
	35-54	0.0213	0.0135	0.0291
5% decrease	15-24	0.0218	0.0180	0.0255
	25-34	0.0149	0.0078	0.0219
	35-54	0.0206	0.0129	0.0284
7% decrease	15-24	0.0217	0.0180	0.0255
	25-34	0.0146	0.0076	0.0216
	35-54	0.0201	0.0124	0.0278
10% decrease	15-24	0.0217	0.0179	0.0254
	25-34	0.0142	0.0072	0.0212
	35-54	0.0193	0.0117	0.0270
15% decrease	15-24	0.0215	0.0178	0.0252
	25-34	0.0136	0.0067	0.0205
	35-54	0.0180	0.0105	0.0255
20% decrease	15-24	0.0214	0.0129	0.0167
	25-34	0.0177	0.0061	0.0092

	35-54	0.0251	0.0198	0.0241
No change in mortality rates	15-24 25-34 35-54	0.0225 0.0212 0.0211	0.0186 0.0140 0.0133	0.0263 0.0284 0.0289
2.5% increase	15-24 25-34 35-54	0.0220 0.0158 0.0226	0.0182 0.0087 0.0147	0.0258 0.0230 0.0305
5% increase	15-24 25-34 35-54	0.0221 0.0183 0.0258	0.0162 0.0090 0.0233	0.0233 0.0153 0.0312
7% increase	15-24 25-34 35-54	0.0221 0.0184 0.0259	0.0164 0.0092 0.0236	0.0238 0.0158 0.0318
10% increase	15-24 25-34 35-54	0.0222 0.0184 0.0260	0.0168 0.0096 0.0240	0.0246 0.0165 0.0327
15% increase	15-24 25-34 35-54	0.0224 0.0174 0.0259	0.0186 0.0102 0.0177	0.0262 0.0247 0.0341
20% increase	15-24 25-34 35-54	0.0225 0.0187 0.0263	0.0179 0.0106 0.0252	0.0272 0.0189 0.0354

Paper II

Research

Risk factors of early sexual debut among men and women - a strong predictor of HIV and sexual risk in Malawi

Humphrey Misiri[1, 2] (hmisiri at gmail dot com)
[1] Department of Biostatistics, Institute of Basic Medical Sciences, University of Oslo, Oslo, Norway. [2] Department of Community Health, College of Medicine, Private Bag 360, Chichiri. Blantyre 3. Malawi

DOI http://dx.doi.org/10.13070/rs.en.1.625
Date 2014-03-25
Cite as Research 2014;1:625
License CC-BY

Abstract

Introduction: In Malawi, just like in other sub-Saharan countries, HIV and other sexually transmitted infections are spread by hetero-sexual contact. Early sexual debut is associated with a higher risk of sexually transmitted infections and HIV. The aim of this paper is to determine the risk factors of early sexual debut in Malawi. **Methods**: Data on 6512 men and 7398 women extracted from the Malawi Demographic and Health Survey 2010 data were analyzed. Separate parametric frailty and logistic regression models were fitted to data for men and women. **Results**: The significant risk factors of early sexual debut among men were secondary education (males: AOR=0.577, 95% CI=0.394, 0.846), Lomwe ethnicity (AOR=1.671, 95% CI =1.274-2.190), Islam (AOR = 1.560, 95% CI=1.065-2.284), being married (AOR=0.418, 95% CI= 0.324-0.538) being in a cohabiting relationship (AOR = 0.416, 95% CI = 0.294 - 0.589), being a widower (AOR = 0.104, 95% CI = 0.013 - 0.818) and the following birth cohorts :1965-1969 (AOR = 2.020, 95% CI=1.273 - 3.301), 1980-1984 (AOR =1.808, 95% CI= 1.134-2.883), 1980-1984 (AOR = 1.808, 95%CI = 1.134 - 2.883) , 1985-1989 (AOR = 1.762, 95% CI = 1.089 - 2.852) and 1990-1995(AOR = 4.757, 95% CI = 2.912 -7.771). The significant risk factors of early sexual debut among women were (AOR = 0.200, 95% CI = 0.135 - 0.297), Lomwe ethnicity (AOR = 1.700, 95% CI = 1.250 - 2.313), Islam AOR = 1.591, 95% CI = 1.074 - 2.357), being married (AOR = 0.582. 95%CI = 0.418 - 0.809), primary education (AOR= 0.540, 95% CI = 0..435 - 0.670), other ethnic groups (AOR = 1.445, 95% CI = 1.054 - 1.980), other Christian denominations (AOR = 1.588, 95% CI = 1.275 - 1.977), being separated (AOR = 0.522, 95% CI = 0.330 - 0.827), and consensual sex (AOR = 0.833, 95% CI = 0.700 - 0.992). **Conclusions**: Education is a protective factor against early sexual debut and so should be used as an intervention for delaying sexual debut in Malawi.

Introduction

Early sexual debut is a risk factor of HIV infection [1] [2]. In Malawi, just like in all sub-Saharan countries [3], HIV is spread through hetero-sexual contact. The prevalence of HIV is 11% [4] and HIV incidence is higher among women than among men [5]. Early sexual debut, therefore, undoubtedly predisposes an individual to HIV infection in such an environment of relatively high HIV prevalence.

Although, antiretroviral therapy (ART) helps to suppress some manifestations of AIDS-defining cancers, coverage of the provision of antiretroviral therapy (ART) is not complete in Malawi. Moreover, there is persistent shortage of drugs [6] for treating sick people. Consequently, HIV continues to be the leading cause of death in this era of antiretroviral therapy. Since the first sexual act can result in infection with HIV, an individual who contracts HIV or STI's during the first sexual encounter at a very young age is at a very big disadvantage because in the absence of any medication his future is bleak. It is therefore important to delay sexual debut in order to avoid this.

In Malawi, sexual activity is risky experience because there is low condom use [4] [7]. Early sexual debut is a known risk factor of unprotected sex [8]. Unprotected sex leads to disease or unwanted teenage pregnancies. A study focusing on orphanhood status and an audit of port-abortion patients in Malawi have shown that early sexual debut is a risk factor of abortion, teenage pregnancy and dropping out of school [9] [10]. Therefore, delaying sexual debut may help to reduce risky sexually practices like unprotected sex.

Culture plays an important role in the lives of most people in Malawi most of whom (about 80%) live in the rural areas. Malawi is a multi-ethnic [11] and male-dominated society. The country is divided into the northern, central and southern provinces which differ by ethnic composition [11] and level of urbanization. The cultural traditions of the ethnic groups in the northern provinces of Malawi are similar but they differ from those of the central and southern provinces which are similar [12] [13]. Several major ethnic groups namely the Yao, Lomwe and Chewa

encourage ritual sex after initiation at puberty where in many cases condoms are not used [14] thereby making first sex risky.

Early sexual debut influences young Malawian women to engage in risky sexual practices in order to fend for themselves. Poverty is reported to be a risk factor of risky behaviour [15]. Since Malawi is a very poor country, some women migrate to urban areas looking for financial opportunities where they end up in prostitution [16]. Some of these are young girls aged less than 15 years [17]. Many of these leave for town to practise prostitution after they have already engaged in early first sex [18].

Studies conducted in America, Europe and Asia have reported the effects of early sexual debut to be substance use, violent delinquency, unwanted pregnancies, sexually transmitted diseases, HIV infection, unprotected sex, abortion, prostitution, paying for sex, multiple partners, cervical cancer [19] [20] [21] [22] [23] [24]. There are also some studies conducted in countries outside Africa which have examined sexual debut [25] [26] [27] [28]. Studies conducted in several sub-Saharan African countries have documented trends in age at first sex [29] [30] [31], early sexual debut [25] [26] [27] [28] and risk factors of age at first sex [32].

There have been no generalisable research results in Malawi on a national scale to examine the risk factors of early sexual debut which is a well-known predictor of HIV. Because of this, there is a knowledge gap. There are a few studies [10] [15] [33] [34] which focused on different topics but mentioned early sexual debut in passing. These studies provide sketchy information about early sexual debut. The aim of this study is to determine the risk factors of early sexual debut in order to fill the gap in knowledge about the risk factors of early sexual debut by analysing nationally representative data from a demographic and health survey conducted in Malawi in 2010. The results of this study will be generalisable because the data is representative nationally.

Methods

Data

The data analyzed for this paper is from the Malawi Demographic and Health Survey 2010 (MDHS2010). The MDHS (2010) data were freely downloaded from www.measuredhs.com after obtaining permission from ICF Macro International. The MDHS2010, was a cross-sectional nationally representative survey conducted in Malawi from June to November 2010 [35]. The survey collected data on different topics from a national multistage cluster sample of 27,340 households.

Each district was demarcated into enumeration areas (EAs). Sampling for the DHS was conducted at the district and EA levels. In each district, the primary sampling units (psus) were the EAs and in each sampled EA, the secondary sampling units (ssus) were households. A total of 849 EAs and 27,340 households were sampled for the survey. In each sampled household, only men aged 15-54 years and women aged 15-49 years were deemed eligible for the survey.

Structured questionnaires with several modules which included an HIV module were administered to eligible members of sampled households. These questionnaires have been used before in 2000 and 2004 in similar surveys. Ethical clearance of the protocol for the MDHS2010 was obtained from the Malawi Health Sciences Research Committee, the Institutional Review Board of ICF Macro, and the Centre for Disease Control and Prevention (CDC) in Atlanta ,USA [35].

For the HIV module, a sub-sample of a third of the total sample was chosen. This translated to a total of 14,407 men and women aged 15-54 and 15-49 years respectively. Apart from taking the HIV test, these sampled individuals were asked to state their age at first sex. Additionally, data on socio-demographic characteristics of the survey participants were also collected. These socio-demographic variables included age, marital status, ethnicity, region or province, place of residence, religion, education level, whether first sex was forced or consensual and date of birth.

The total numbers of explanatory variables for men and women's data are different since data on whether first sexual was consensual or not were collected from women only.

The end point of the current analysis is first sex. The response variable is age at sexual debut which is a time-to-event variable. Survey participants who never had sex are right–censored. For this paper early sexual debut is first sex before the age of 15 years. This is derived from the age at first sex.

Models fitted

Parametric survival and logistic regression models were fitted to the data. Age at first sex was the outcome for

survival analysis. Early sexual debut (Yes, No) was the outcome variable for logistic regression. Survival analysis techniques are appropriate for analyzing this data since age at early sexual debut is a time-to-event variable. Since the data are hierarchical with two levels namely enumeration area (EA) and the household, the parametric accelerated failure time model (with frailty) was used to analyze the data. The random effect variable was enumeration area.

Preliminary analyses were conducted to determine the distribution which best fitted the survival time data. Several distributions which included the lognormal and Weibull distributions were assumed for the age at sexual debut and a corresponding survival model fitted. For each chosen distribution, a corresponding hazard function was graphed. Besides, the selected distribution was also fitted using Lindsey and Mersch(1992)'s approach [36] if the distribution was a member of the exponential family. A survival model was fitted after preliminary analyses.

Time ratios, which are part of the optional output produced by Stata software, were extracted. The interpretation of time ratios is straight forward. The reference category has by default a time ratio equal to 1. For a given level of a categorical variable, a time ratio less than 1 implies that for that category sex was initiated earlier (or faster) than for the baseline category. To the contrary, a time ratio greater than 1 implies that sex was initiated later (slower) than for the baseline category.

Logistic regression models were used for determining the risk factors of early sexual debut. Univariate and multiple logistic regression models were fitted to the data for each gender. Primary sampling units (EAs) were controlled for in the logistic regression analyses. Sampling weights were also incorporated in the regression analyses. For all analyses, STATA 10 was used. All p-values less than 0.05 were indicative of a significant result.

Data for men and women were analyzed separately because men and women are different since there are obvious differences in physiological and biological factors, social and behavioral factors between men and women which make the sexual experience of men different from that of women The effect of gender was not analyzed because the number of variables for men's and women's data were different [37].

The MDHS2010 survey questionnaires had been used in previous surveys of a similar nature. Additionally, these questionnaires were translated into Chichewa – the language widely spoken all over Malawi and Chitumbuka - the language spoken mainly in the northern province so bias was controlled [35].

Results

The number of respondents who answered the question about age at sexual debut was 13910 out of a total of 14407 respondents. Information about the remaining 497(4%) is missing. Of the 13,910 who answered the question about sexual debut, 6512(47%) were men and 7398(53%) were women (Table 1).

About 53% of men were married or had been married before and about 38% were single. 73% of women were either married or had at least been married before and 8% were in cohabitation and 19% had never been married before. Over 80% of men and women were from rural areas. Over 90% of the men's sample and over 80% of the women's sample had primary or higher education.

From the men's sample, 915(14%) were censored, 43% had first sex by age 15 and 82% had first sex by age 20. The median age at sexual debut for men computed using survival analysis techniques was 18 years. As for women,

		Men		Women	
Variable	**Category**	**Total(N=6512)**	**Had sex**	**Total(N=73398)**	**Had sex**
Region	Northern	1170(18.0)	937(16.7)	1301(17.6)	1125(17.6)
	Central	2410(37.0)	2110(37.7)	2576(34.8)	2184(34.1)
	Southern	2932(45.0)	2550(45.6)	3521(47.6)	3093(48.3)
	Unknown	-	-		-
Religion	Roman Catholic	1425(21.9)	1207(21.6)	1555(21)	1316(20.6)
	CCAP	1069(16.4)	869(15.5)	1176(15.9)	956(14.9)
	SDA	481(7.4)	397(7.1)	549(7.4)	465(7.3)
	Other Christian	2670(41.0)	2341(41.8)	3272(44.2)	2904(45.4)
	Muslim	624(9.6)	563(10.1)	775(10.5)	695(10.9)
	Other	242(3.7)	219(3.9)	66(0.9)	61(1)
	Unknown	1(0.0)	1(0.0)	5(0.0)	5(0.1)
Ethnicity	Chewa	1942(29.8)	1683(30.1)	2230(30.1)	1893(29.6)
	Tumbuka	644(9.9)	511(9.1)	751(10.2)	640(10)
	Lomwe	1152(17.7)	1013(18.1)	1211(16.4)	1049(16.4)
	Yao	656(10.1)	581(10.4)	747(10.1)	673(10.5)
	Sena	359(5.5)	314(5.6)	418(5.7)	368(5.7)
	Ngoni	860(13.2)	746(13.3)	1037(14)	893(13.9)

only 996 (13%) were still virgins at the time of the survey, 25% had first sex by age 15 and 75% had sex by age 18. The median age at sexual debut for women computed using survival analysis techniques was 17 years.(fig1).

The probability of remaining a virgin decreases rapidly from 10 to 25 years for both genders after which it reduces to almost zero (Fig 1). The lognormal distribution provides a good fit to the age at first sex (Fig 2) and the hazard functions plotted for each gender (Fig 3) resemble lognormal hazard functions depicted in literature [38] (fig2) (fig3).

The adjusted results from the survival models (Table 2) show that among men central province, southern province , primary education , Lomwe ethnicity , other ethnicities, other Christian religions, other religions, birth cohorts 1965-1969 to 1990-1995, being single, divorced or separated have significant time ratios which are less than 1. The adjusted results from the survival models (Table 3) show that central province, primary education and secondary or higher education has significant time ratios greater than 1. All other significant variables significant time ratios which are less than 1.

Men who have secondary or higher education were less likely to engage in early sexual debut than uneducated men (AOR= 0.577, 95% CI = 0.394-0.846)(Table4). The odds of early sexual debut were 67% higher for Lomwe men than for Chewa men. Islamic men were 56% more likely to engage in early sexual debut than Catholic men. (AOR =0.560 , 95% CI= 1.065 - 2.284). Men who were

	Other	898(13.8)	748(13.4)	1001(13.5)	883(13.8)
	Unknown	1(0.0)	1(0.0)	3(0.04)	-
Education	None	404(6.2)	387(6.9)	1080(14.6)	1061(16.6)
	Primary	4209(64.6)	3554(63.5)	4962(67.1)	4251(66.4)
	≥Secondary	1899(29.2)	1656(29.6)	1356(18.3)	1090(17)
	Unknown	-	-		-
Residence	Urban	902(13.9)	754(13.5)	995(13.4)	837(13.1)
	Rural	5610(86.1)	4843(86.5)	6403(86.6)	5565(86.9)
	Unknown	-	-		-
Marital status	Never married	2477(38.0)	1562(27.9)	1481(20)	485(7.6)
	Married	3253(50.0)	3253(58.1)	4301(58.1)	4301(67.2)
	Cohabiting	560(8.6)	560(10.0)	633(8.6)	633(9.9)
	Widowed	28(0.4)	28(0.5)	280(3.8)	280(4.4)
	Divorced	109(1.7)	109(1.9)	382(5.2)	382(6)
	Separated	85(1.3)	85(1.5)	321(4.3)	321(5)
	Unknown	-	-		-
Age group	15-19	1619(24.9)	880(15.7)	1660(22.4)	746(11.7)
	20-24	1092(16.8)	949(17)	1400(18.9)	1328(20.7)
	25-29	963(14.8)	945(16.9)	1412(19.1)	1403(21.9)
	30-34	842(12.9)	835(14.9)	1003(13.6)	1002(15.7)
	35-39	711(10.9)	707(12.6)	815(11)	815(12.7)
	40-44	505(7.8)	504(9)	580(7.8)	580(9.1)
	45-49	447(6.9)	445(8)	528(7.1)	528(8.2)
	50-54	333(5.1)	332(5.9)	1660(22.4)	746(11.7)
	Unknown	-	-	-	-

Table 1.Socio-demographic characteristics of the participants of the national Demographic and Health Survey, Malawi 2010.

[enlarge]

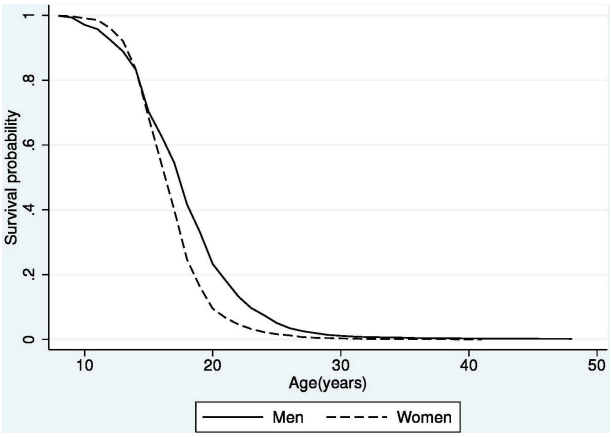

Figure 1. Survival probability for men and women.

[enlarge]

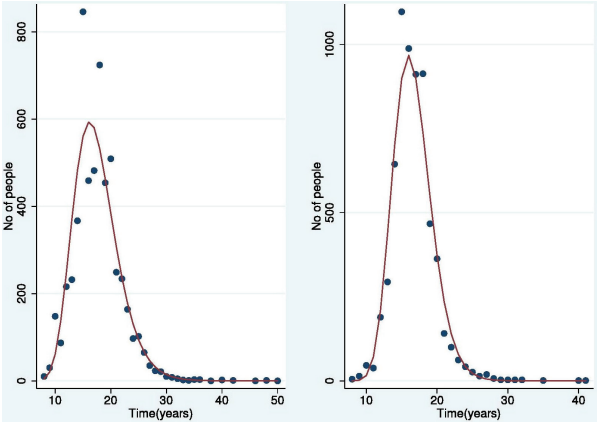

Figure 2. Number of people vs. age of sexual debut for men (left) and women (right).

Variable	Category	Crude Time Ratios			
		Time Ratio(95%CI)	p-value	Time Ratio(95%CI)	p-value
Province	Northern	Reference	-	Reference	-
	Central	0.961(0.943,0.98)	p<0.0001	0.971(0.947,0.995)	0.018
	Southern	0.933(0.916,0.951)	p<0.0001	0.968(0.945,0.991)	0.007
Residence	Urban	Reference	-	Reference	-
	Rural	0.982(0.964,1.001)	0.061	0.992(0.975,1.010)	0.408
Education	None	Reference	-	Reference	-
	Primary	0.947(0.925,0.97)	p<0.0001	0.965(0.943,0.988)	0.003
	Secondary or higher	1.004(0.979,1.030)	0.749	1.012(0.986,1.038)	0.368
Ethnicity	Chewa	Reference	-	Reference	-
	Tumbuka	1.037(1.014,1.062)	0.002	1.001(0.974,1.030)	0.918
	Lomwe	0.95(0.933,0.968)	p<0.0001	0.947(0.927,0.968)	p<0.0001
	Yao	0.943(0.923,0.964)	p<0.0001	0.98(0.951,1.011)	0.198
	Sena	0.965(0.937,0.993)	0.015	0.975(0.945,1.005)	0.097
	Ngoni	0.975(0.956,0.995)	0.014	0.968(0.949,0.987)	0.001
	Other	0.996(0.976,1.017)	0.735	0.977(0.954,1.000)	0.051
Religion	Roman Catholic	Reference	-	Reference	-
	Presbyterian	1.011(0.992,1.031)	0.260	1.006(0.987,1.024)	0.556
	Seventh Day Adventist	1.001(0.976,1.027)	0.924	1.009(0.986,1.034)	0.441
	Other Christian religions	0.982(0.967,0.998)	0.023	0.983(0.969,0.998)	0.025
	Muslim	0.933(0.911,0.954)	p<0.0001	0.939(0.912,0.968)	p<0.0001
	Other religions	0.969(0.937,1.001)	0.055	0.964(0.934,0.994)	0.020
Birth cohort	1955-1959	Reference	-	Reference	-
	1960-1964	0.994(0.961,1.028)	0.739	0.991(0.959,1.025)	0.615
	1965-1969	0.960(0.928,0.993)	0.017	0.953(0.922,0.985)	0.005
	1970-1974	0.952(0.923,0.983)	0.003	0.949(0.919,0.979)	0.001
	1975-1979	0.958(0.929,0.988)	0.006	0.945(0.917,0.974)	p<0.0001
	1980-1984	0.939(0.911,0.968)	p<0.0001	0.925(0.898,0.953)	p<0.0001
	1985-1989	0.924(0.896,0.952)	p<0.0001	0.893(0.865,0.922)	p<0.0001

born in the 1965-1969 birth cohort were 102% more likely to engage in early sexual debut :1965-1969 (AOR = 2.020, 95% CI=1.273 - 3.301). The odds of early sexual debut were 13% higher for men born in the 1980-1984 period than for men born during 1955-1959 (AOR = 1.808, 95%CI = 1.134 - 2.883). Malawian men born during the 1985-1989 period were 76% morelikely to engage in early sexual debut than those who were born in the 1995-1959 period (AOR = 1.762, 95% CI = 1.089 - 2.852) and 1990-1995(AOR = 4.757, 95% CI = 2.912 -7.771). Married men were 58% less likely to have engaged in early sexual debut than men who were single (AOR = 0.418, 95% CI = 0.324 - 0.538). The odds of early sexual debut were 58 % lower for men in a cohabiting relationship than for men who were single (AOR = 0.416, 95% CI = 0.294 - 0.589). Widowers were 90% less likely to have engaged in early sexual debut than single men (AOR = 0.104, 95% CI = 0.013 - 0.818).

The adjusted results from regression models for women(Table 5) show that women who had primary education were less likely to engage in early sexual debut than uneducated women (AOR= 0.540, 95% CI = 0..435 - 0.670). The odds of early sexual debut were 80% lower for women who had secondary or higher education than for women who had no education (AOR = 0.200, 95% CI = 0.135 - 0.297).

Lomwe women were 70% more likely to engage in early sexual debut than Chewa women (AOR = 1.700, 95% CI = 1.250 - 2.313). The odds of early sexual debut for women of other ethnicities were 45% higher than for Chewa women (AOR = 1.445, 95% CI = 1.054

	1990-1995	0.864(0.839,0.889)	p<0.0001	0.833(0.805,0.863)	p<0.0001
Marital status	Never married	Reference	-	Reference	-
	Married	1.062(1.049,1.076)	p<0.0001	0.967(0.948,0.988)	0.002
	Cohabiting	1.082(1.059,1.106)	p<0.0001	0.989(0.963,1.015)	0.398
	Widowed	1.075(0.99,1.167)	0.086	0.973(0.896,1.058)	0.523
	Divorced	1.008(0.965,1.053)	0.72	0.922(0.881,0.964)	p<0.0001
	Separated	0.983(0.936,1.033)	0.498	0.911(0.867,0.958)	p<0.0001

Table 2. Crude and adjusted time ratios (TR) from parametric frailty models for men.

1.588, 95% CI = 1.275 - 1.977). Women who belonged to other religions were 49% more likely to engage in early sexual debut than Catholic women(AOR = 1.489, 95% CI = 0.679 - 3.266). Married women were 42% less likely to engage in early sexual debut than single women (AOR = 0.582. 95%CI = 0.418 - 0.809). The odds of early sexual debut were 48% lower for women who had separated from their husbands than for women who were single (AOR = 0.522, 95% CI = 0.330 - 0.827). Women who had consensual sex were 17% less likely to have engaged in early sexual debut than for women who were raped (AOR = 0.833, 95% CI = 0.700 - 0.992).

Discussion

Our analysis has shown that for men, early sexual debut was significantly associated with secondary or higher education, Lomwe ethnicity, Islam, being married, cohabiting or widowed and being born between 1965 and 1995. For women, early sexual debut was significantly associated with whether sex was consensual or forced, primary or higher education, Lomwe or other ethnicity, being a Moslem, being born during the 1990-1995 period and being married, cohabiting or separated. For both genders, province and the place of residence were not significantly related to early sexual debut.

Our findings show that secondary education was a - 1.980). Islamic women were 59% more likely to engage in early sexual debut than Catholic women (AOR = 1.591, 95% CI = 1.074 - 2.357). Similarly, the odds of early sexual debut were 59% higher for women of other Christian denominations than for Catholic women (AOR =

[enlarge]

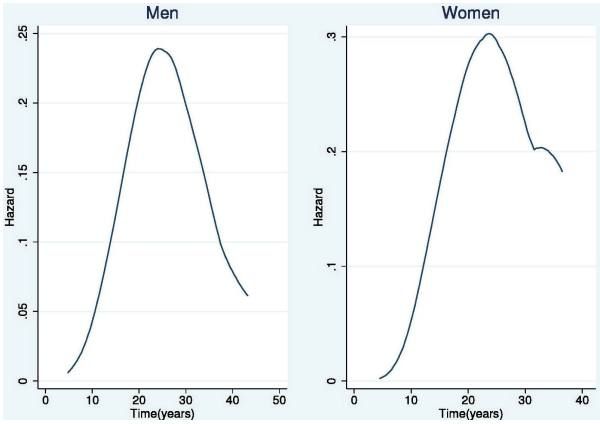

Figure 3. Hazard function plot.

		Crude Odds Ratios			
Variable	**Category**	**OR(95%CI)**	**p-value**	**OR(95%CI)**	**p-value**
Province	Northern	Reference	-	Reference	-
	Central	1.035(1.021,1.049)	p<0.0001	1.026(1.009,1.044)	0.003
	Southern	0.979(0.967,0.992)	0.001	0.990(0.974,1.006)	0.225
Residence	Urban	Reference	-	Reference	-
	Rural	0.975(0.962,0.988)	p<0.0001	1.011(0.999,1.024)	0.077
Education	None	Reference	-	Reference	-
	Primary	1.031(1.019,1.042)	p<0.0001	1.027(1.016,1.039)	p<0.0001
	Secondary or higher	1.147(1.132,1.162)	p<0.0001	1.122(1.106,1.139)	p<0.0001
Ethnicity	Chewa	Reference	-	Reference	-
	Tumbuka	0.984(0.969,0.999)	0.039	0.982(0.963,1.000)	0.054
	Lomwe	0.942(0.930,0.954)	p<0.0001	0.960(0.946,0.974)	p<0.0001
	Yao	0.932(0.918,0.946)	p<0.0001	0.970(0.951,0.989)	0.002
	Sena	0.973(0.955,0.992)	0.006	1.011(0.991,1.032)	0.279
	Ngoni	0.974(0.961,0.987)	p<0.0001	0.974(0.962,0.987)	p<0.0001
	Other	0.950(0.937,0.963)	p<0.0001	0.965(0.950,0.980)	p<0.0001
Religion	Roman Catholic	Reference	-	Reference	-
	Presbyterian	1.017(1.004,1.031)	0.011	1.002(0.989,1.015)	0.779
	Seventh Day Adventist	0.996(0.979,1.012)	0.605	1.000(0.984,1.016)	0.991
	Other Christian	0.966(0.956,0.976)	p<0.0001	0.981(0.971,0.991)	p<0.0001

protective factor against early sexual debut among both men and women. These results agree with McGrath *et al* (2009) who also reported that school had the effect of delaying sexual debut among South African men and women [39]. Another study by Pettifor at al (2008) also reported that staying in school protected South African girls from the risk of HIV infection [40].

In Malawi, most of the primary and secondary schools are owned by the government and most of them are co-educational. Furthermore, there are more primary schools than secondary schools. Secondary schools have four classes called Forms 1 through 4. Primary school have 8 classes called Standards 1 through 8. The recommended age for enrolling in Standard 1 is five years. The period before age 15 is usually spent in primary school. Schools provide an environment of great social interaction between boys and girls. This interaction between boys and girls sometimes culminates in hetero-sexual contact which is early sexual debut. In fact, Ngaiyaye (2000) reported that early sexual debut was the cause of high dropout rate on grounds of teenage pregnancy in Malawian schools [41]. To address this problem, the Malawi Government tried to integrated Life Skills and Sexuality and Reproductive Health in the curriculum of senior primary classes and all secondary school forms in 2002. Recently after assessing the impact of that pilot exercise,

	religions				
	Muslim	0.942(0.927,0.956)	p<0.0001	0.972(0.954,0.991)	0.003
	Other religions	0.981(0.941,1.023)	0.366	1.002(0.962,1.042)	0.938
Birth cohort	1960-1964	Reference	-	Reference	-
	1965-1969	1.006(0.986,1.027)	0.558	1.007(0.988,1.027)	0.463
	1970-1974	1.005(0.987,1.024)	0.572	1.001(0.984,1.019)	0.872
	1975-1979	1.018(1.000,1.036)	0.049	1.003(0.987,1.020)	0.696
	1980-1984	1.012(0.995,1.029)	0.178	0.991(0.975,1.007)	0.263
	1985-1989	1.017(1.000,1.034)	0.045	0.980(0.964,0.996)	0.013
	1990-1995	1.038(1.021,1.055)	p<0.0001	0.931(0.914,0.948)	p<0.0001
Marital status	Never married	Reference	-	Reference	-
	Married	0.886(0.875,0.896)	p<0.0001	0.870(0.857,0.883)	p<0.0001
	Cohabiting	0.877(0.862,0.892)	p<0.0001	0.867(0.851,0.883)	p<0.0001
	Widowed	0.878(0.859,0.897)	p<0.0001	0.855(0.835,0.876)	p<0.0001
	Divorced	0.887(0.870,0.905)	p<0.0001	0.873(0.855,0.892)	p<0.0001
	Separated	0.890(0.872,0.909)	p<0.0001	0.872(0.853,0.891)	p<0.0001
Sex consensual	No	Reference	-	Reference	-
	Yes	0.963(0.955,0.971)	p<0.0001	0.997(0.988,1.006)	0.483

Table 3. Crude and adjusted time ratios (TR) from parametric frailty models for women.

		Crude Odds Ratios			
Variable	**Category**	**OR(95%CI)**	**p-value**	**OR(95%CI)**	**p-value**
Province	Northern	Reference	-	Reference	-
	Central	1.353(1.055,1.736)	0.017	1.063(0.744,1.519)	0.735
	Southern	1.520(1.195,1.934)	0.001	0.964(0.671,1.385)	0.842
Residence	Urban	Reference	-	Reference	-
	Rural	1.060(0.861,1.304)	0.585	1.035(0.820,1.306)	0.772
Education	None	Reference	-	Reference	-
	Primary	1.701(1.256,2.304)	0.001	1.151(0.822,1.61)	0.413
	Secondary or higher	0.915(0.658,1.273)	0.598	0.577(0.394,0.846)	0.005
Ethnicity	Chewa	Reference	-	Reference	-
	Tumbuka	0.834(0.619,1.124)	0.233	1.084(0.738,1.591)	0.682
	Lomwe	1.391(1.132,1.710)	0.002	1.671(1.274,2.190)	p<0.0001
	Yao	1.471(1.136,1.904)	0.003	1.190(0.800,1.771)	0.390
	Sena	1.329(1.010,1.748)	0.042	1.357(0.966,1.905)	0.078
	Ngoni	1.185(0.935,1.503)	0.160	1.275(0.994,1.635)	0.056
	Other	0.859(0.673,1.097)	0.223	1.023(0.740,1.414)	0.891
Religion	Roman Catholic	Reference	-	Reference	-
	Presbyterian	1.107(0.882,1.389)	0.380	1.080(0.851,1.370)	0.527
	Seventh Day Adventist	0.904(0.670,1.221)	0.511	0.840(0.610,1.158)	0.287
	Other Christian religions	1.084(0.902,1.303)	0.389	1.071(0.880,1.302)	0.494
	Muslim	1.647(1.278,2.123)	p<0.0001	1.560(1.065,2.284)	0.022
	Other religions	1.331(0.937,1.891)	0.110	1.431(0.989,2.072)	0.057
Birth cohort	1955-1959	Reference	-	Reference	-
	1955-1959	1.105(0.648,1.885)	0.713	1.130(0.658,1.938)	0.658
	1965-1969	1.823(1.120,2.966)	0.016	2.020(1.237,3.301)	0.005

	1970-1974	1.580(0.966,2.585)	0.068	1.653(1.007,2.716)	0.047
	1975-1979	1.294(0.809,2.069)	0.282	1.462(0.906,2.359)	0.119
	1980-1984	1.712(1.084,2.702)	0.021	1.808(1.134,2.883)	0.013
	1985-1989	2.200(1.398,3.462)	0.001	1.762(1.089,2.852)	0.021
	1990-1995	9.400(6.085,14.522)	p<0.0001	4.757(2.912,7.771)	p<0.0001
Marital status	Never married	Reference	-	Reference	-
	Married	0.223(0.191,0.260)	p<0.0001	0.418(0.324,0.538)	p<0.0001
	Cohabiting	0.216(0.160,0.290)	p<0.0001	0.416(0.294,0.589)	p<0.0001
	Widowed	0.061(0.008,0.445)	0.006	0.104(0.013,0.818)	0.032
	Divorced	0.515(0.331,0.801)	0.003	0.948(0.586,1.532)	0.826
	Separated	0.473(0.280,0.800)	0.005	0.755(0.426,1.338)	0.335

Table 4. Crude and adjusted odds ratios from logistic regression model for men.

		Crude Odds Ratios			
Variable	**Category**	**OR(95%CI)**	**p-value**	**OR(95%CI)**	**p-value**
Province	Northern	Reference	-	Reference	-
	Central	0.720(0.537,0.964)	0.027	0.709(0.486,1.036)	0.075
	Southern	1.514(1.147,1.997)	0.003	1.134(0.790,1.629)	0.495
Residence	Urban	Reference	-	Reference	-
	Rural	1.024(0.799,1.312)	0.852	0.778(0.596,1.017)	0.066
Education	None	Reference	-	Reference	-
	Primary	0.591(0.484,0.723)	p<0.0001	0.540(0.435,0.670)	p<0.0001
	Secondary or higher	0.263(0.189,0.365)	p<0.0001	0.200(0.135,0.297)	p<0.0001
Ethnicity	Chewa	Reference	-	Reference	-
	Tumbuka	1.094(0.752,1.591)	0.640	1.144(0.703,1.861)	0.588
	Lomwe	2.383(1.884,3.013)	p<0.0001	1.700(1.250,2.313)	0.001
	Yao	2.400(1.793,3.212)	p<0.0001	1.397(0.907,2.152)	0.129
	Sena	1.73(1.2170,2.460)	0.002	0.987(0.654,1.491)	0.952
	Ngoni	1.417(1.075,1.869)	0.013	1.297(0.967,1.741)	0.083
	Other	2.003(1.554,2.583)	p<0.0001	1.445(1.054,1.980)	0.022
Religion	Roman Catholic	Reference	-	Reference	-
	Presbyterian	0.890(0.658,1.204)	0.450	1.000(0.734,1.361)	0.998
	Seventh Day Adventist	1.377(0.966,1.964)	0.077	1.238(0.863,1.776)	0.246
	Other Christian religions	1.767(1.421,2.197)	p<0.0001	1.588(1.275,1.977)	p<0.0001
	Muslim	2.307(1.726,3.084)	p<0.0001	1.591(1.074,2.357)	0.021
	Other religions	1.573(0.721,3.428)	0.254	1.489(0.679,3.266)	0.320
Birth cohort	1960-1964	Reference	-	Reference	-
	1965-1969	1.000(0.688,1.454)	0.999	1.008(0.690,1.472)	0.969
	1970-1974	0.870(0.614,1.233)	0.433	0.907(0.624,1.318)	0.608
	1975-1979	0.765(0.541,1.081)	0.129	0.889(0.629,1.257)	0.506
	1980-1984	0.780(0.566,1.076)	0.130	1.024(0.726,1.444)	0.891
	1985-1989	0.729(0.528,1.005)	0.054	0.987(0.703,1.387)	0.940
	1990-1995	1.257(0.924,1.710)	0.145	1.438(1.008,2.050)	0.045
Marital status	Never married	Reference	-	Reference	-
	Married	0.631(0.480,0.830)	0.001	0.582(0.418,0.809)	0.001
	Cohabiting	0.750(0.520,1.081)	0.122	0.659(0.436,0.997)	0.048

Kalanda (2010) reported a significant change in the rate of drop outs from primary school on grounds of pregnancy after pupils were taught life skills and sexual reproductive health [42]. This shows that education is an effective intervention for controlling the sexual behaviour of Malawian school children. In fact, sex and HIV education interventions are believed to be effective in delaying sexual debut not only in Malawi but in developing countries in general [43].

The fact that primary education was a protective factor against early sexual debut among girls can be explained by girls' survival tactics to avoid getting pregnant so as to stay in school. Until a few years ago, a girl who became pregnant whilst in school was permanently rusticated - never to be allowed back in school even long after delivery. In fact, twelve years ago, Moore *et al*(2007) reported that in Malawi and Ghana for girls who remain in school: "Girls in school are more likely to want to avoid sexual intercourse to prevent pregnancy as pregnancy indefinitely suspends and potentially ends their schooling" [33]. Therefore, being in school may help both boys and girls to delay sexual debut since girls will try to avoid sexual contact at all costs. Therefore, boys will have nobody in school to have sex with.

Ethnicity seems to be an important factor in influencing age at early sexual debut in that men and women of Lomwe ethnicity were more likely to have early sexual debut than Chewa men and women. Moreover, women of other ethnicities were also

	Widowed	0.913(0.585,1.423)	0.687	0.823(0.482,1.406)	0.475
	Divorced	0.809(0.528,1.240)	0.331	0.681(0.421,1.102)	0.117
	Separated	0.569(0.375,0.864)	0.008	0.522(0.330,0.827)	0.006
Sex consensual	No	Reference	-	Reference	-
	Yes	0.844(0.714,0.999)	0.048	0.833(0.700,0.992)	0.040

Table 5. Crude and adjusted odds ratios from logistic regression model for women.

more likely to have early sexual debut than Chewa women. This may be explained by the differences in cultural traditions about sexuality. In Malawi, cultural traditions are a powerful force which governs people's sexuality. Some cultures regulate sexual behaviour. In fact, a study conducted in Norway by Traen and Samuelsen (2007) reported that “Human sexuality is shaped by cultural context and historical period” [44].

Birth cohort was also an important factor which is associated with early sexual debut among men when other factors were adjusted for. This finding agrees with the reported decrease of the age at sexual debut among men between 2005 and 2010 [4]. The fact that birth cohort was not a significant risk factor for women is hard to explain because evidence exists of an increase of the age at first sex among Malawian women between 2000 and 2010 [4]. Nevertheless, the findings for both men and women agree with Zuma *et al* (2011) who reported birth cohort as one of the risk factors of early sexual debut in South Africa [32].

Religion affects sexuality since religious teaching may include imparting instructions concerning sexual behaviour. The current study has found that being a Muslim predisposes men and women to early sexual debut. This finding is hard to explain. However, this may be due to peer pressure within the social network of the religious community [39]. Fatusi and Blum (2008) had similar results in Nigeria and McGrath *et al* (2009) also had similar findings in South Africa [39]. Religious leaders are influential in religious communities and so may be important partners in effecting behavioral change in the community.

The predominantly small time ratios among men and women combined with the results from regression analysis imply that with some few exceptions for women, Malawian men and women initiate sex faster on average. Exceptions are women who have at least primary school education. Indeed, according to MDHS2010, the reported median ages at first marriage were 17.1, 17.5, 20.4 and 24.5 years respectively for women with no education, with primary, secondary and higher education respectively [4]. The main reasons for this are unknown.The main limitation of the current research is that the data was obtained from survey participants through recall by ORC Macro International and the Malawi National Statistical Office. In such cases, the problem of recall bias is inevitable. Nevertheless, considering the nature of the topic and the sample sizes, recall bias cannot invalidate or dilute the results of the current research [34] [45].

Conclusions

Early sexual debut is associated with ethnicity, education, religion, birth cohort, marital status and consensual sex. It is quite remarkable that for both men and women education was a protective factor against early sexual debut. Education should be used as an intervention for delaying sexual debut. Policies aimed at achieving this goal should encourage boys and girls to stay in school at the same time taking into account the differences in ethnicities and religions of people.

Declarations

Competing interests

There are no competing interests.

Acknowledgments

I thank ICF Macro International for allowing me to use the MDHS2010 data.

Funding

The study was sponsored by NUFU and the Research Council of Norway contract/grant number: 177401/V50.

Authors' contributions

HM conceived the study, obtained the data, analyzed the data, drafted the manuscript and revised it.

References

1. Pettifor A, O'Brien K, MacPhail C, Miller W, Rees H. Early coital debut and associated HIV risk factors among young women and men in South Africa. Int Perspect Sex Reprod Health. 2009;35:82-90 pubmed publisher
2. Stöckl H, Kalra N, Jacobi J, Watts C. Is early sexual debut a risk factor for HIV infection among women in sub-Saharan Africa? A

systematic review. Am J Reprod Immunol. 2013;69:27-40 pubmed publisher

3. Carael M, Holmes K. Dynamics of HIV epidemics in sub-Saharan Africa: introduction. AIDS. 2001;15:S1-4 pubmed
4. National Statistical Office (NSO), ORC Macro. Malawi Demographic and Health Survey 2010. Zomba: NSO and ORC Macro; 2010.
5. Misiri H, Edriss A, Aalen O, Dahl F. Estimation of HIV incidence in Malawi from cross-sectional population-based sero-prevalence data. J Int AIDS Soc. 2012;15:14 pubmed
6. Lufesi N, Andrew M, Aursnes I. Deficient supplies of drugs for life threatening diseases in an African community. BMC Health Serv Res. 2007;7:86 pubmed
7. Chimbiri A. The condom is an 'intruder' in marriage: evidence from rural Malawi. Soc Sci Med. 2007;64:1102-15 pubmed
8. Potdar R, Mmari K. Factors influencing sexual initiation, multiple partners and condom use among male slum youth in Pune, India. Glob Public Health. 2011;6:843-58 pubmed publisher
9. Levandowski B, Pearson E, Lunguzi J, Katengeza H. Reproductive health characteristics of young Malawian women seeking post-abortion care. Afr J Reprod Health. 2012;16:253-61 pubmed
10. Mkandawire P, Tenkorang E, Luginaah I. Orphan status and time to first sex among adolescents in northern Malawi. AIDS Behav. 2013;17:939-50 pubmed publisher
11. Malawi Human Rights Commission. Cultural Practices and their Impact on the Enjoyment of Human Rights, Particularly the Rights of Women and Children in Malawi. Lilongwe.
12. Munthali A, Zulu E. The timing and role of initiation rites in preparing young people for adolescence and responsible sexual and reproductive behaviour in Malawi. Afr J Reprod Health. 2007;11:150-67 pubmed
13. Morris B. Animals and ancestors:An ethnography. Oxford, UK: Berg; 2000.
14. van Breugel JWM. Chewa Traditional Religion. Blantyre,Malawi. Blantyre,Malawi: CLAIM; 2001.
15. Madise N, Zulu E, Ciera J. Is poverty a driver for risky sexual behaviour? Evidence from national surveys of adolescents in four African countries. Afr J Reprod Health. 2007;11:83-98 pubmed
16. Banda C. Teen prostitution: Making ends meet in Malawi. 2005. Available from: www.greatreporter.com/content/teen-prostitution-making-ends-meet.htm
17. Kalanda BF. Empowering young sexual workers for safer sex in Dowaand Lilongwe districts of Malawi. Malawi Medcol Journal. 2010;22(1):10-1.
18. Geubbels E, Bowie C. Epidemiology of HIV/AIDS in adults in Malawi Malawi Medcol Journal. 2006;18(3):99-121.
19. Baumgartner J, Waszak Geary C, Tucker H, Wedderburn M. The influence of early sexual debut and sexual violence on adolescent pregnancy: a matched case-control study in Jamaica. Int Perspect Sex Reprod Health. 2009;35:21-8 pubmed publisher
20. Ma Q, Ono-Kihara M, Cong L, Xu G, Pan X, Zamani S, *et al*. Early initiation of sexual activity: a risk factor for sexually transmitted diseases, HIV infection, and unwanted pregnancy among university students in China. BMC Public Health. 2009;9:111 pubmed publisher
21. Kaestle C, Halpern C, Miller W, Ford C. Young age at first sexual intercourse and sexually transmitted infections in adolescents and young adults. Am J Epidemiol. 2005;161:774-80 pubmed
22. Soares A. Early patterns of sexual activity: age cohort differences in Australia. International Journal of STD & AIDS. 2007;14(11):745-752.
23. Zhao FH, Tiggelaar SM, Hu SY, Xu LN, Hong Y, Niyazi M, *et al*. A multi-center survey of age of sexual debut and sexual behavior in Chinese women: suggestions for optimal age of human papillomavirus vaccination in China. Cancer Epidemiology. 2012;36(4):384-90.
24. Stueve A. Early sexual initiation and subsequent sex-related risks among urban minority youth: the reach for health study. Fam Plann Perspect. 2001;33:268-75 pubmed
25. Rissel C, Richters J, Grulich A, de Visser R, Smith A. Sex in Australia: first experiences of vaginal intercourse and oral sex among a representative sample of adults. Aust N Z J Public Health. 2003;27:131-7 pubmed
26. Guo W, Wu Z, Qiu Y, Chen G, Zheng X. The timing of sexual debut among Chinese youth. Int Perspect Sex Reprod Health. 2012;38:196-204 pubmed publisher
27. Woynarowska B, Tabak I. [Risk factors of early sexual initiation]. Med Wieku Rozwoj. 2008;12:541-7 pubmed
28. Caminis A, Henrich C, Ruchkin V, Schwab-Stone M, Martin A. Psychosocial predictors of sexual initiation and high-risk sexual behaviors in early adolescence. Child Adolesc Psychiatry Ment Health. 2007;1:14 pubmed
29. Cremin I, Mushati P, Hallett T, Mupambireyi Z, Nyamukapa C, Garnett G, *et al*. Measuring trends in age at first sex and age at marriage in Manicaland, Zimbabwe. Sex Transm Infect. 2009;85:i34-40 pubmed publisher
30. Slaymaker E, Bwanika J, Kasamba I, Lutalo T, Maher D, Todd J. Trends in age at first sex in Uganda: evidence from Demographic and Health Survey data and longitudinal cohorts in Masaka and Rakai. Sex Transm Infect. 2009;85:i12-9 pubmed publisher
31. Zaba B, Pisani E, Slaymaker E, Boerma J. Age at first sex: understanding recent trends in African demographic surveys. Sex Transm Infect. 2004;80:ii28-35 pubmed
32. Zuma K, Mzolo T, Makonko E. Determinants of age at sexual debut and associated risks among South African youths. African Journal of AIDS Research. 2011;10:189-94.
33. Moore A, Awusabo-Asare K, Madise N, John-Langba J, Kumi-Kyereme A. Coerced first sex among adolescent girls in sub-Saharan Africa: prevalence and context. Afr J Reprod Health. 2007;11:62-82 pubmed
34. Glynn J, Kayuni N, Floyd S, Banda E, Francis-Chizororo M, Tanton C, *et al*. Age at menarche, schooling, and sexual debut in northern Malawi. PLoS ONE. 2010;5:e15334 pubmed publisher
35. Macro International. Malawi Demographic and Health Survey 2010:Preliminary Report. Maryland, USA: MEASURE DHS; 2011; Available from. Available from: pdf.usaid.gov/pdf_docs/PNADU780.pdf
36. Lindsey JK, Mersch G. Fitting and comparing probability distributions with loglinear models. Computational Statistics and Data Analysis. 1992;13:73-384.
37. Heidari S, Abdool Karim Q, Auerbach J, Buitendijk S, Cahn P, Curno M, *et al*. Gender-sensitive reporting in medical research. J Int

AIDS Soc. 2012;15:11 pubmed publisher

38. Klein JP, Moeschberger ML. Survival Analysis:Techniques for Censored and Truncated Data. New York: Springer-Verlag; 1997.
39. McGrath N, Nyirenda M, Hosegood V, Newell M. Age at first sex in rural South Africa. Sex Transm Infect. 2009;85:i49-55 pubmed publisher
40. Pettifor A, Levandowski B, MacPhail C, Padian N, Cohen M, Rees H. Keep them in school: the importance of education as a protective factor against HIV infection among young South African women. Int J Epidemiol. 2008;37:1266-73 pubmed
41. Ngaiyaye R. Sexuality and Reproductive Health Education for primary schools in Malawi: Baseline Report. Lilongwe,Malawi: Ministry of Education Sports and Culture; 2000.
42. Kalanda BF. Life skills and reproductive health education changes behaviour in students and teachers:Evidence from Malawi. Educational Research and Reviews. 2010;5(4):169-74.
43. UNAIDS. Global Report of the AIDS Epidemic 2010; Available from: ww.unaids.org.
44. Traeen B, Samuelsen SO. Sweet 16 and never been kissed?Experiences from a longitudinal Norwegian study. Electronic Journal of Human Sexuality. 2007;10:6.
45. Fatusi A, Blum R. Predictors of early sexual initiation among a nationally representative sample of Nigerian adolescents. BMC Public Health. 2008;8:136 pubmed publisher

ISSN : 2334-1009

Paper III

Global Advanced Research Journal of Medicine and Medical Sciences Vol. 1(6) pp. 145-153, July, 2012
Available online http://garj.org/garjmms/index.htm
Copyright © 2012 Global Advanced Research Journals

Full Length Research Paper

Cancer incidence in Malawi: Time trends in Blantyre 1996-2005 and predictions up to 2015

Humphrey Misiri[1,2], Charles Dzamalala[3], Abdi Edriss[6], D Max Parkin[4], and Freddie Bray[1,5]

1 Department of Biostatistics, Institute of Basic Medical Sciences, University of Oslo, Oslo, Norway.
2 Department of Community Health, College of Medicine, Blantyre, Malawi.
3 Department of Pathology, College of Medicine, Blantyre, Malawi.
4 CTSU, University of Oxford, UK
5. Norwegian Cancer Registry, Oslo, Norway
6 Bunda Collage, University of Malawi, Lilongwe, Malawi

Accepted 28 June, 2012

As in many sub-Saharan countries, communicable diseases have been given greater public health priority in Malawi, although the magnitude of the cancer burden is increasing as a result of demographic changes, as well as the impact of the HIV pandemic. To be able to describe the patterns of cancer between 1996 and 2005 and to predict the incidence and total burden of cancer for 2015,we analysed data from the Malawi National Cancer Registry for the period from 1996 to 2005. We obtained age standardized incidence rates for the most common cancers in Malawi. Linear trend models were used to predict incidence rates and the burden of cancer for Blantyre for 2015. The most common cancers, in terms of age-standardized (world) incidence rates were Kaposi sarcoma (50.5 per 100,000 for males, 26.4 for females), cervical cancer (49.3 per 100,000), oesophageal cancer (22.3 for males, 14.6 for females), non-Hodgkin lymphoma (6.6 for males, 5.3 for females), eye cancer (4.4 for males, 5.4 for females), and breast cancer (11.9). Predictions based on the rather rapidly increasing trends would yield an upper limit of 2512 cases of cancer in Malawi by 2015; an absolute percentage increase of 193.4 % and 242.7 % among males and females, respectively. Based on our analysis we conclude that incidence rates of cancer in Blantyre have been increasing between 1996 and 2005. Apart from the AIDS pandemic in Malawi, population growth and ageing will also contribute to the projected threefold increase in the number of cancer cases.

Keywords: Cancer, incidence, predictions, time trends

INTRODUCTION

An estimated 12.7 million new cases of cancer worldwide occurred in 2008, and 7.6 million cancer deaths (Ferlay2008). The cancer burden in high-income areas has surpassed that of low- and medium-income areas, in which 80% of the global population currently reside. The profile of the most common cancers differ markedly by level of income within a given country or region; cancers of the prostate, lung and colorectum – frequent in westernized countries – do not, for instance, figure among the five most common cancers in Sub-Saharan Africa, where infection-related Kaposi sarcoma, and cancers of the cervix and liver often dominate, alongside female breast cancers (Parkin, Sitas et al. 2008).

As in many sub-Saharan countries, communicable diseases have been afforded a greater public health priority in Malawi, given the magnitude of the disease burden relative to cancer (World Health Organisation 2004). Malawi has been particularly affected by the HIV

*Corresponding author E-mail: hmisiri@gmail.com, humphrey.misiri@medisin.uio.no; Tel/Cell: +265-888-342-864: Fax: +265-1874-700

pandemic, with the prevalence of infection in adults aged between 15 and 49 years in Blantyre city estimated to be 22.3% in 2004 (National Statistical Office of Malawi 2005), the highest prevalence in the country; the national prevalence of HIV in Malawi was estimated at approximately 12% circa 2007 (National Aids Commission of Malawi 2008).

However, it is clear that non-communicable diseases, including cancer, will become an increasing burden on health services in Malawi, as in other low and middle income countries (Samb 2010). Knowledge of the cancer profile in Malawi can be derived from the Malawi Cancer Registry, which, although recording data from various parts of the country, has aimed at complete population coverage of the Blantyre District (urban and rural divisions). The first report covered the first five years of registration (1994-8) (Banda, Parkin et al. 2001), while more recent results (2000-2001) have been published in Parkin et al (2003) (Parkin 2003). The profile of childhood cancer was reported by Mukiibi et al (1995) (Mukiibi1995) and Banda & Liomba (1999) (Banda 1999). There was a relatively high incidence of Kaposi sarcoma, as well as cancers of the oesophagus and cervix uteri, with more modest rates of prostate, breast and bladder cancer, as well as non-Hodgkin lymphomas.

The current report provides an update of the incidence to 2005, with the aim of describing and interpreting the incidence trends in the urban population of Blantyre within the last decade (1996-2005). We focus on the major types of cancer observed in the Blantyre population: Kaposi sarcoma, non-Hodgkin lymphoma, eye and oesophageal cancer, as well as cancers of the breast and cervix in women. Based on these trends, we predict the future cancer burden up to 2015; such information is valuable in planning cancer services and deciding on the allocation of finite resources to prevention, treatment and palliative care, as well as providing a baseline against which the success of interventions can be measured.

MATERIALS AND METHODS

The Malawi National Cancer registry was established in 1989, and became population-based for the Blantyre District (Urban and Rural), in 1993. Registration is carried out by a programme of regular visits by cancer registrars to all hospitals (government and private) in the District, with data recorded on cases of cancer from hospital records departments and clinical services, where cases might have been diagnosed or treated. There is no comprehensive system of death registrations in Malawi, and thus death certificates are not used as a source of information. Cancer diagnoses are coded according to the International Classification of Disease for Oncology (ICD-O). The CANREG system provided by IARC is used for the recording of cases; this includes checks at data entry for potential duplicates as well as for impossible or unlikely codes or combinations of codes, and provides automatic code conversions to ICD-10.

Incidence data were extracted from the registry database by cancer site (coded using ICD-10), sex and age group (ages 0-14, 15-24, 25-34,35-44, 45-54, 55 and over) and year of registration (1996-2005). Censuses were carried out in 1998 and 2008. The population of Blantyre District (urban and rural) was 809,397 in 1998 and slightly more than 1 million in 2008 (National Statistical Office of Malawi). The data extracted and analysed in this paper are restricted to this population, with residential status defined as persons who have lived in Blantyre for at least six months (Banda, Parkin et al. 2001). About 10% of registrations are incomplete due to missing ages.

Data on the underlying population-at-risk came from two sources. Population estimates for Blantyre for 1996 to 1997 and 1999 to 2005 were estimated via linear interpolation (Gerald 1984) using census data for the years 1987 (National Statistical Office of Malawi) ,1998 (National Statistical Office of Malawi) and 2008, whereas more recent population data for 2000 and 2015 were available by sex and age from the tables of national population projections for 2009-2030 produced by the National Statistical Office of Malawi (NSO) (National Statistical Office of Malawi).

Trend analysis was restricted to the most frequent cancer types observed in Blantyre over the 10-year period. Annual age-specific and age-standardised incidence (world) rates (ASR) for 1996-2005 were calculated, where applicable, by sex for Kaposi sarcoma (ICD-10 C46), cancer of the eye (ICD-10 C69), non-Hodgkin lymphomas (ICD-10 C82-C85, C96), oesophageal cancer (ICD-10 C15), female breast (C50) and cervix cancer (ICD-10 C53). Of the 372 eye cancers, 75% were squamous cell carcinomas of the conjunctiva and 20% retinoblastomas). We also ascertained a residual group comprising “all other cancers combined” (see footnote, Table 1) to enable the prediction of the total cancer burden in 2010 and 2015. The all-ages ASR has been corrected for cases of unknown age by multiplying by T/K, where T is the total number of cancer- and sex-specific cases and K is the corresponding number for which age is known.

The study used time-linear models (Hakulinen and Dyba 1994; Dyba 1997; Dyba and Hakulinen 2000; Dyba and Hakulinen 2008) to predict cancer incidence for 2015 by extrapolating linear trends seen in the recent past, with the base of prediction set at 1996-2005. In the models, the number of cases of cancer in each age-sex group were assumed to have a Poisson distribution (Dyba and Hakulinen 2000). Incorporated in the prediction methods are specific models suitable for increasing and decreasing trends. As *a priori* analyses indicated increasing trends for the major cancer types in Malawi, we chose the following candidate models that avoided the prospect of an exponential explosion in incidence (Model1) and positive estimate of age standardised rate (Model2):

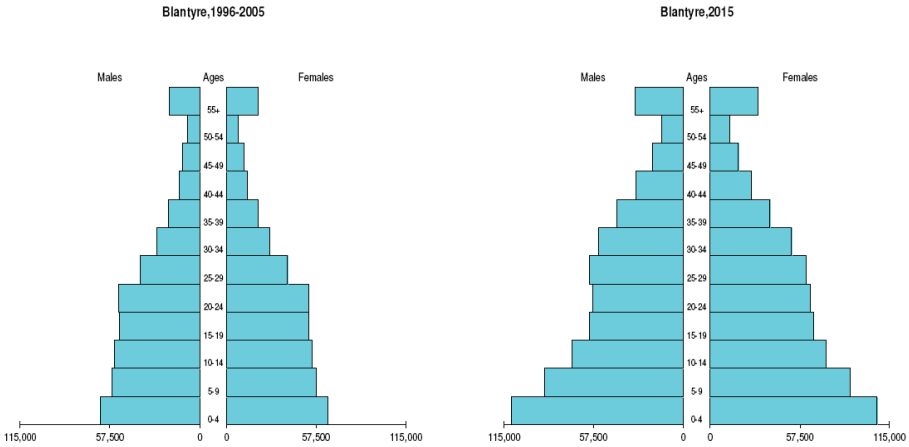

Figure 1. The Population Pyramids of Blantyre District
Left Panel: Mean annual population for 1996-2005

Right Panel: Projected population for 2015

$$E\left[\frac{c_{it}}{n_{it}}\right] = \alpha_i + \beta_i t \text{ ..... (1)}$$ and

$$E\left[\frac{c_{it}}{n_{it}}\right] = \alpha_i + \beta t \text{ ..... (2)}$$

where

c_{it} is the number of cases of cancer in age group i and year t

n_{it} is the number of people (pyr) in age group i and year t

$E\left[\frac{c_{it}}{n_{it}}\right]$ *is the expected value of incidence in age group i and year t*

α_i *is the intercept parameter - baseline incidence for age group i.*

β_i *is the slope parameter for age group i*

β *is a common parameter for all age groups*

For most combinations of age, sex and cancer site, Model 1 was chosen, because the larger number of parameters gave a better fit. In some cases, however, Model 1 could not be localized due to data limitations, and then Model 2 was fitted instead. Model 2 was fitted for Kaposi sarcoma in women.

The fitted models were used to estimate predicted incidence rates for 2015. These incidence rates were then multiplied by the population forecasts for Blantyre to yield the corresponding number of predicted cancer cases.

To quantify the changes in the annual number of new cases of cancer in 2015 from the mean annual number observed 1996-2005, the approach by Møller et al (2002) (Moller, Fekjaer et al. 2002) was adopted, with the net differences partitioned into changing risk of cancer and changing age structure and population size.

RESULTS

Population projections

The average and projected age-specific population pyramids for 1996-2005 and 2015 are displayed in Figure 1 in the left and right panel respectively.

The average population in 1996-2005 and population projection for 2000 estimated at 858,049 and 848442 respectively are projected to increase to 1,283,333. The population forecast for 2015 indicates that there will be 51% more inhabitants in Malawi than during the period 1996-2005 and 2000. Furthermore, the mean projected number of people aged 50 years or over in 2015 is forecasted to be 51% greater than the estimated population 1996-2005.

Age standardised incidence estimates (1996-2005)

A mean annual number of 791 cases of cancer were diagnosed 1996-2005 in Blantyre (Table 1), with a slight majority of these occurring in men (51.8%). Almost half of the male cancers were Kaposi sarcoma (ASR 54.1, comprising 47.9% of all male cancer cases), with oesophageal cancer (ASR 21.5, 10.5%), non-Hodgkin lymphoma (ASR 7.0, 7.3%) and eye cancer (ASR 4.7, 4.4%) the next three most frequent cancers among

Table 1. Mean annual number of cases and age standardised rates (ASR) by sex group for selected sites for Blantyre (1996-2005)

Site	ICD10	Mean annual number of cases1996-2005	%	ASR(World)	Morphological verification (%)
Males					
Oesophagus	C15	43	11.0	21.5	18.5
Kaposi sarcoma	C46	196	48.0	54.1	10.6
Eye	C69	18	4.4	4.6	79.0
Non-Hodgkin lymphoma	C82-85,C96	30	7.3	7.0	66.6
Other cancers#		123	30.0	56.7	55.7
All cancers		410	100.0	143.0	32.0
Females					
Oesophagus	C15	24	6.3	13.4	12.4
Kaposi sarcoma	C46	98	26.0	27.8	12.6
Breast	C50	23	6.0	11.4	67.4
Cervix uteri	C53	103	27.0	47.6	51.6
Eye	C69	20	5.2	5.7	82.1
Non-Hodgkin lymphoma	C82-85,C96	21	5.5	5.4	68.8
Other cancers#		93	24.0	42.7	66.5
All cancers		382	100.0	152.8	46.2

#Other cancers comprise Oral cavity (ICD-10 C00-08), Nasopharynx (ICD-10 C11), Other pharynx(ICD-10 C09-10,C12-C13), Stomach(ICD-10 C16),Colon and rectum (ICD-10 C18-21),Lliver (ICD-10 C22),Bronchus and lung (ICD-10 C33-34),Skin melanoma (ICD-10 C43),Other skin cancers (ICD-10 C44),Ovary (ICD-10 C56),Male breast cancer(ICD-10 C50),Other female genital cancers (ICD-10 C51-52,C57), Penis (ICD-10 C60),Prostate (ICD-10 C61),Bladder (ICD-10 C67),Thyroid (ICD-10 C73) ,Hodgkin's disease (ICD-10 C81),Leukaemia (ICD-10 C91-95) and a residual of other neoplasms whose ICD10 code follows:ICD10-C14,ICD10-C90,ICD10-C22-C26,C30-34,C40-44,C17-C45, ICD10-C47-C49, ICD10-C62-C66,ICD10-C71-C72,C74-C77,ICD10-C80

Malawian men.

All other cancer forms in combination constituted the remaining 30.1% of the male cancer burden over the 10-year period. The male: female ratio of the five cancer types diagnosed in both men and women indicated the preponderance of cancers in men, including Kaposi sarcoma (M:F ratio of 2.0), oesophageal cancer (1.8), non-Hodgkin lymphoma (1.4) and other cancer forms (1.3); eye cancer was the exception, with a M:F ratio of 0.9.

In women, over half of all diagnosed cancer cases in Malawi were either cervical cancer (ASR 47.6, constituting 27.0% of female cancer cases) or Kaposi sarcoma (ASR 27.8 per 100,000, 25.7%). Eye cancer (ASR 5.7, 5.2%), non-Hodgkin lymphoma (ASR 5.4, 5.5%), oesophageal cancer (ASR 13.3, 6.3%), and breast cancer (ASR 11.3, 6.0%) were the next four most frequent types of female cancers observed, and the six neoplasms explained over two-thirds of the total female cancer burden.

The basis of diagnosis of 36% and 42% of cancers in men and women respectively was morphological verification (MV). Most of cancers of the eye, breast and NHL were diagnosed using cytological or histological examination of tissue. Just a small proportion of Kaposi sarcoma and oesophageal cancer in both genders were diagnosed by cytology or histology.

Incidence trends

Age-specific incidence rates for Kaposi sarcoma, breast, oesophageal, cervical and eye cancer, as well as other cancers combined are shown in Figure 2.

For Kaposi sarcoma and cancer of the eye, incidence rates are greatest in the age group 35-44 in both sexes. For oesophageal cancer (men and women), breast (women) and cervical cancer (women) and other cancers (men and women), the incidence rates continue to rise with increasing age.

The age-standardised rates of Kaposi sarcoma, non-Hodgkin lymphoma, oesophageal and eye cancers, as well as other cancers combined, and the fitted trend between 1996 and 2005, together with the predicted rates for 2015 are shown below for men and women in Figures 3 and 4 respectively. Also shown is the average rate for the period 1996-2005, projected forward to 2015 (assuming no further change).

The cancer-specific trends are generally increasing in both sexes between 1996 and 2005, although there is considerable year-on-year random variation evident in the trends. In men, incidence rates of Kaposi sarcoma, non-Hodgkin lymphoma, oesophageal and eye cancer are predicted to increase (Figure 3).

Similarly, in women incidence rates of non-Hodgkin

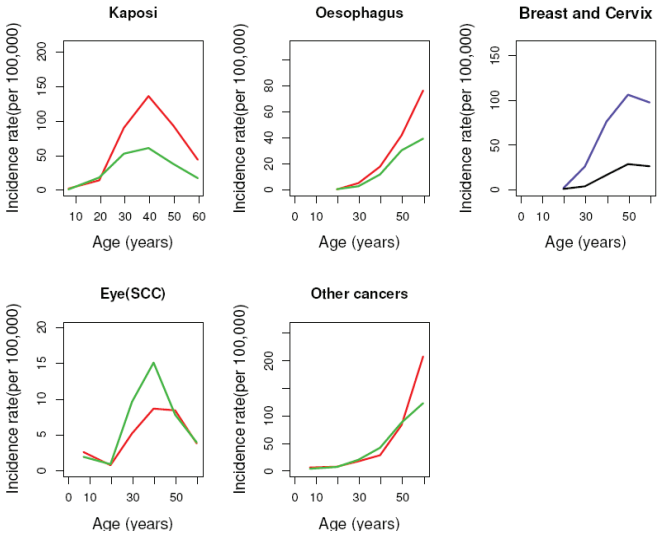

Figure 2. Age-specific incidence trends among males and females of Blantyre district, Malawi 1996-2005. Red line – Males; Green line- Females; Blue line - Cervical cancer; Black line - Breast cancer

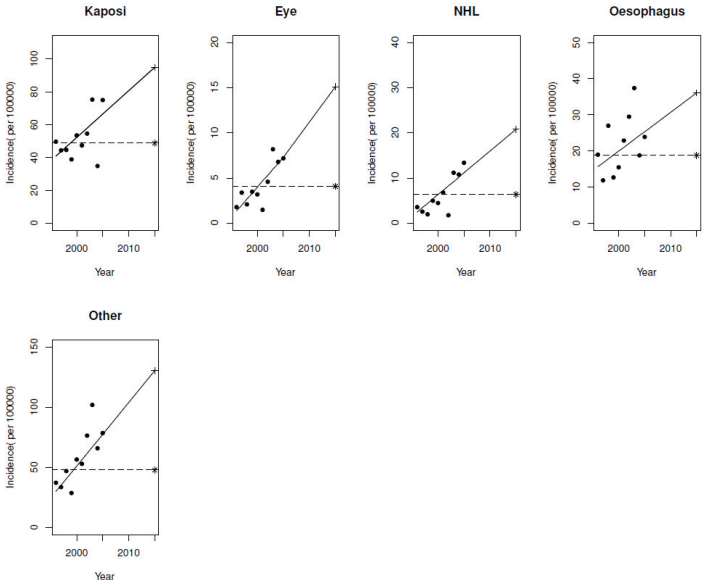

Figure 3. Age-standardised incidence trends for Kaposi sarcoma, and cancers of the oesophagus, eye and all other cancers combined among males in Blantyre, Malawi for 1996-2005 and projections for 2015.
+ 2015 Projected standardized incidence rate; * 1996-2005 Mean standardized incidence rate

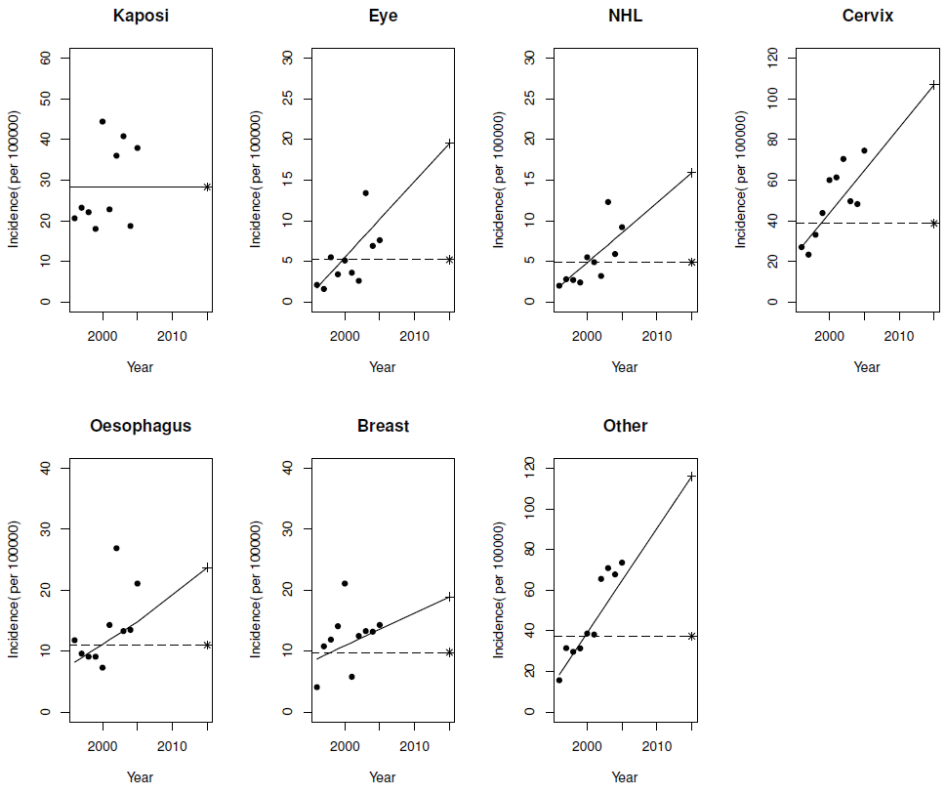

Figure 4. Age-adjusted incidence trends of cervical cancer, breast cancer, Kaposi sarcoma, oesophageal, eye and other cancers among females in Blantyre, Malawi for 1996-2005 and projection for 2015.
+ 2015 Projected standardized incidence rate; * 1996-2005 Mean standardized incidence rate
The bold line is the projected trend and the dotted line is mean of ASR's of all the years from 1996 to 2005

lymphoma, oesophageal and eye cancer are predicted to increase (Figure 4). However, the age standardised incidence of Kaposi sarcoma in women is predicted to be constant up to 2015. In both sexes, the incidence rates of all other cancers are also predicted to increase. There are large year-to-year variations in the trends of the incidence rate of Kaposi sarcoma in men and other cancers. These, however, are unlikely to cause similar changes in incidence.

Total cancer burden prediction for 2015

The mean annual number of male cancer cases in Blantyre for 1996-2005 was 410. If the rates observed in this decade were maintained, the numbers would increase to 618 by 2015, a 51% increase due to the projected increase (and ageing) of the population (Table 2). From the trends-based model, an increase to 1147 cases is predicted by 2015 (Table 2); a rise of 180%.

The numbers of non-Hodgkin lymphoma, eye and other cancer cases among men are projected to have the greatest increase of over 200% between 1996-2005 and 2015.

For Kaposi sarcoma, the number of cases is projected to increase by over 120%. The smallest increase in number of cases due to risk is predicted for oesophageal cancer. The total annual number of female cases of cancer in 2015 is predicted to increase by 841 cases which is an increase of over 200%. There are large projected increases in the number of female cases of oesophageal cancer, Kaposi sarcoma, non-Hodgkin lymphoma, eye cancer, cervical, breast cancer and other cancers.

Table 2. Percentage changes in the projected annual number of cases for 2015 relative to the mean annual number of cases for 1996-2005 in Blantyre, Malawi by sex and site.

		OBSERVED 1996-2005 Mean annual cases	PREDICTED CASES 2015					
			No change in rates			Prediction		
Site	**ICD10**	**N(%)**	**N(%)**	**%Change**	**N(%)**	**Lower**	**Upper**	**%Change**
Males								
Oesophagus	C15	43(10.5)	65(10.5)	51.2	89(7.8)	44	135	107.0
Kaposi sarcoma	C46	196(47.8)	299(48.4)	52.6	454(39.6)	78	830	131.6
Eye	C69	18(4.4)	27(4.4)	50.0	76(6.6)	52	100	322.2
Non-Hodgkin lymphoma	C82-85;C96	30(7.3)	46(7.4)	53.3	145(12.6)	98	191	383.3
Other#		123(30.0)	181(29.3)	47.2	383(33.4)	300	465	211.4
Total		410(100.0)	618(100)	50.7	1147(100)	572	1721	179.8
Females								
Oesophagus	C15	24(6.3)	39(6.3)	62.5	62(5.1)	36	89	158.3
Kaposi sarcoma	C46	98(25.7)	159(25.6)	62.2	221(18.1)	98	344	125.5
Breast	C50	23(6.0)	38(6.1)	65.2	59(4.8)	25	93	156.5
Cervix	C53	103(27.0)	171(27.5)	66.0	328(26.8)	245	411	218.4
Eye	C69	20(5.2)	33(5.3)	65.0	94(7.7)	52	137	370.0
Non-Hodgkin lymphoma	C82-85;C96	21(5.5)	33(5.3)	57.1	93(7.6)	47	139	342.9
Other#		93(24.3)	149(24)	60.2	366(29.9)	307	426	293.5
Total		382(100.0)	622(100)	62.8	1223(100)	810	1639	220.2

#Other cancers comprise Oral cavity (ICD-10 C00-08), Nasopharynx (ICD-10 C11), Other pharynx(ICD-10 C09-10,C12-C13), Stomach(ICD-10 C16),Colon and rectum (ICD-10 C18-21),Lliver (ICD-10 C22),Bronchus and lung (ICD-10 C33-34),Skin melanoma (ICD-10 C43),Other skin cancers (ICD-10 C44),Ovary (ICD-10 C56),Male breast cancer(ICD-10 C50),Other female genital cancers (ICD-10 C51-52,C57), Penis (ICD-10 C60),Prostate (ICD-10 C61),Bladder (ICD-10 C67),Thyroid (ICD-10 C73) ,Hodgkin's disease (ICD-10 C81),Leukaemia (ICD-10 C91-95) and a residual of other neoplasms whose ICD10 code follows:ICD10-C14,ICD10-C90,ICD10-C22-C26,C30-34,C40-44,C17-C45, ICD10-C47-C49, ICD10-C62-C66,ICD10-C71-C72,C74-C77,ICD10-C80

DISCUSSION

The growth and ageing of the world population will lead to an increase in the global burden from non-communicable diseases including cancer, and the increase is projected to be particularly marked in low and middle income countries (Mathers 2006). Our analysis confirms that similar changes can be anticipated in Malawi; the shifting demographics will inevitably increase the numbers of cancer cases diagnosed each year. However, our results suggest that, over and above this effect, the increasing incidence rates will result in an even larger contribution of cancer to the disease burden, with the number of cases of cancer in 2015 predicted to be up to three times those observed on average over the study period 1996-2005.

Projections of cancer incidence rates make the implicit assumption that the trends observed in the past that are used as a prediction base (1996-2005 in this paper) are accurately measured, and that they will continue into the future, either unchanged (as in our short term predictions), or with pre-defined modifications (Bray 2006). This assumption was validated by exploratory analyses.

The observed trends in 1995-2006 include rather dramatic increases in incidence for almost all cancers considered, and it seems likely that the temporal pattern observed is partly based on artefact, due, for example, to changes in the completeness of case ascertainment by the cancer registry in the last few years. During the decade under study there were changes to the registration methodology that would have contributed. Several new hospitals serving the catchment population were opened, and added to the case finding sources, and active search for cases was extended to all hospitals, having been confined prior to 1999 to the major teaching hospital (Queen Elizabeth Central Hospital), with reliance on passive surveillance at smaller centres. Given these uncertainties concerning the reasons for the increasing trends in the period 1996-2005, as well as the implausibly high rate of change for some cancer sites, it would be prudent to interpret the rates and number of cases based on the predicted change in incidence with extreme caution as they surely represent the upper limit of the cancer burden in 2015. Conversely, those based on stability of rates (those observed in 1996-2005) represent the lower limit of future burden circa 2015. Nevertheless population growth and ageing alone will yield an increase in cancer cases of 448 - 57% by 2015, assuming rates remain fixed. Blantyre is the commercial capital of Malawi and migration into Blantyre City is always high as people look for financial or job opportunities. Besides, the total fertility rate of Blantyre -4.3% by 2004 - is high enough to

lead to the projected increase in the total population in 2015.

AIDS-related cancers (Kaposi sarcoma, non-Hodgkin lymphoma and eye cancers) comprised 60% of cancers in men in 1996-2005, and 36% of cancers in women; a further 26% of female cancers were cancers of the cervix, also considered to be “AIDS-defining”. While it is possible that the increases in incidence of these cancers relates to the epidemic of HIV-AIDS, it should be noted that, although the prevalence of HIV increased in the years up to 1998 (when it peaked in adults aged 15-49 at around 15% (UNAIDS 2011)), prevalence has been declining since then, to around 11% in 2009. In addition, the availability and provision of antiretroviral therapy – which suppresses manifestations of HIV and AIDS such as cancer - has increased, from almost zero circa 2004, to more than 50% of AIDS patients in 2009 (UNAIDS 2011). Similar changes in Uganda have resulted in quite marked declines in the incidence of Kaposi Sarcoma (Parkin 2010). The incidence of Kaposi sarcoma, the cancer with the strongest association with HIV, has increased at slower rate than any other of the malignancies considered, and the increasing incidence of the “other cancer” category (12-13% per year in 1996-2005), which does not include any HIV-related malignancies, also implies that HIV infection is unlikely to be the principal cause of the increase in incidence. The big increase in the projected number of cases of Kaposi sarcoma is almost entirely due to population increase and ageing, since incidence rates are projected to remain fixed.

Mlombe et al (2009) had previously noted the increasing incidence of oesophageal cancer (Mlombe 2009). The increase in the observed incidence trend of breast cancer trends during 1996-2005 is similar to that observed in Kampala, Uganda in 1991-2006 (where it was ~5% per year (23)). Risk factors for breast cancer are increasing in prevalence in the population of Malawi, for example, contraceptive use among currently married women increased from 13.0% in 1992 to 30.6% in 2000 and 32.5% by 2004 and declining fertility, although it is unlikely that these could account for such large increases in incidence.

Our analysis illustrates that cancer is already a significant health problem in Blantyre. Projections based on demographic projections, or on incorporating recent trends reveal that cancer is very likely to become a major public health issue in the coming decades. While some of the trend-based predicted increase in incidence most likely results from a changing accrual of cases over time, part of the increase in the cancer burden in Malawi probably in part the result of rising trends in several cancers common in western countries, linked to social and economic transitions and consequent changes in lifestyle. Increases in the prevalence of tobacco consumption and HIV-induced immune-suppression will also have an important effect on the risk of cancer by 2015. The forecasted population growth (due to natural means and internal migration) and ageing in Malawi will certainly mean there are more cancer patients in Blantyre in Malawi in the next decades, irrespective of the true temporal patterns. This demands that urgent consideration be given to the orientation of preventive and curative services for cancer (and other non-communicable diseases) in the years ahead (Samb 2010).

ACKNOWLEDGEMENTS

We are very grateful to Macro International for granting us permission to use the MDHS2004 data. Humphrey Misiri acknowledges support from the Norwegian Programme for Development, Research and Education (NUFU). The study was also partially supported by funds from the Research Council of Norway through contract/grant number: 177401/V50.

REFERENCES

Banda LT, Parkin DM, Dzamalala CP, Liomba NG (2001). Cancer incidence in Blantyre, Malawi 1994-1998. Trop Med Int Health. 6(4):296-304.

Banda LTL, Liomba NG (1999). Malawi National Cancer Registry 1991-1995. International Incidence of Childhood Cancer. D. M. Parkin, Kramarova, E., Draper, G.J., Masuyer, E., Michaelis, J., Neglia, J, Qureshi, S., Stiller, C.A. Lyon,France, IARC. 2: 31-34.

Bray F, Møller B (2006). "Predicting the future burden of cancer." Nat Rev Cancer **6**(1): 63-74.

Dyba T, Hakulinen T (2000). "Comparison of different approaches to incidence prediction based on simple interpolation techniques." Statistics in Med. **19**(13): 1741-1752.

Dyba T, Hakulinen T (2008). "Do cancer predictions work?" Euro. J. Cancer **44**(3): 448-453.

Dyba T, Hakulinen T, Paivarinta L (1997). "A simple non-linear model incidence prediction." Statistics in Med. **16**(20): 2297-2309.

Ferlay J, Shin HR, Bray F, Forman D, Mathers C, Parkin D (2008). "Estimates of worldwide burden of cancer in 2008, GLOBOCAN."

Gerald CF, Wheatley PO (1984). Applied Numerical Analysis. Taipei,Taiwan, Addison-Wesley Publishing Co.

Hakulinen T, Dyba T (1994). "Precision of Incidence Predictions Based on Poisson Distributed Observations." Statistics in Med. **13**(15): 1513-1523.

Mathers CD, Loncar D (2006). "Projections of Global Mortality and Burden of Disease from 2002 to 2030." PLoS Med **3**(11).

Mlombe Y, Dzamalala C,Chisi J, Othieno-Abinya N (2009). "Oesophageal cancer and Kaposi sarcoma in Malawi: a comparative analysis." Malawi Med. J. **21**(2): 66-68.

Møller B, Fekjaer H, Hakulinen T, Tryggvadóttir L, Storm HH, Talbäck M, Haldorsen T (2002). Prediction of cancer incidence in the Nordic countries up to the year 2020. Eur. J. Cancer Prev. 11(S1):S1-96.

Mukiibi JM, Banda L, Liomba NG, Sungani FC, Parkin DM (1995). Spectrum of childhood cancers in Malawi 1985-1993. East Afr. Med. J. 72(1):25-29.

National Aids Commission of Malawi (NAC) (2008). HIV and Syphilis Sero-Survey and National HIV Prevalence and AIDS Estimates Report for 2007. Lilongwe,Malawi.**:** 1-38.

National Statistical Office of Malawi (NSO) "1998 Census Results." Retrieved 29 April 2011, from http://www.nso.malawi.net/index.php?option=com_content&view=article&id=127%3A1998-population-and-housing-census&catid=8&Itemid=3.

National Statistical Office of Malawi (NSO). Malawi Population and Housing Census 1987, Summary of Final Results. . Zomba, Malawi Government.

National Statistical Office of Malawi (NSO). "Population projections for Malawi." Retrieved 29 April,2011, from http://www.nso.malawi.net/index.php?option=com_content&view=article&id=134%3Apopulation-projections-for-malawi&catid=8&Itemid=3.

National Statistical Office of Malawi (NSO), Macro ORC (2005). Malawi Demographic Survey 2004. , NSO and ORC Macro.

Parkin DM, Ferlay J, Hamdi-Chérif M, Sitas F, Thomas JO, Wabinga H, Whelan SJ (Ed) (2003). Cancer in Africa._. IARC Scientific Publication

No 153. Lyon,France, IARC.
Parkin DM, Nambooze S, Wabwire-Mangen F, Wabinga HR (2010). "Changing cancer incidence in Kampala, Uganda, 1991-2006." Int. J. Cancer. **126**: 1187-1195.
Parkin DM, Sitas F, Chirenje M, Stein L, Abratt R, Wabinga H (2008). Part I: Cancer in Indigenous Africans--burden, distribution, and trends. Lancet Oncol. 9(7):683-692..
Samb B, Desai N, Nishtar S, Mendis S, Bekedam H, Wright A, Hsu J, Martiniuk A,Celletti F, Patel K, Adshead F, McKee M, Evans T, Alwan A, Etienne C (2010). "Prevention and management of chronic disease: a litmus test for health-systems strengtheningin low-income and middle-income countries." Lancet **376**: 1785-1797.
UNAIDS (2011). "UNAIDS: Epidemiological Factsheet- Malawi" Retrieved 29 April, 2011, from http://www.unaids.org/en/regionscountries/countries/malawi/.
World Health Organisation (WHO) (2004). "Global Burden of Disease: Disease and Injury Country Estimates " Lancet Retrieved 29 April 2011, from http://www.who.int/healthinfo/global_burden_disease/estimates_country/en/index.html.