University of Gothenburg

School of Business, Economics, and Law

Survival Bias and the Impact of HIV on Wealth in Sub-Saharan Africa

Johanna Dees

Maxim Brüls

Supervisor: Annika Lindskog

Master's thesis in Economics, 30 hec

Spring 2019

Graduate School, School of Business, Economics and Law, University of Gothenburg, Sweden

ABSTRACT

In this paper we estimate the effect of HIV-infection on household wealth. We use circumcision status as an instrumental variable, combined with DHS data, to obtain an exogenous variation in HIV on the individual level. We subsequently account for several biases in the data, including the early deaths of HIV-positive persons and the subsequently skewed dataset in favor of survivors. Specifically, we develop a rolling window, based on a three-stage environmental variables model to account for age- and wealth heterogeneity. This model then estimates the effect of HIV on wealth. Overall, the effects are large for poorer percentiles of the population and smaller for wealthier percentiles across a range of different subsamples.

KEYWORDS

HIV, Instrumental Variables Regressions, Rolling Window, Circumcision, Wealth, sub-Saharan Africa

Table of Contents

SECTION I: INTRODUCTION

The Human Immunodeficiency Virus, or HIV, has been well studied by researchers in various disciplines. Scientists have covered the medical, economic, and sociological aspects of HIV and HIV epidemics for well over thirty years. As of 2019, the virus has or had infected over thirty million people and brought little but poverty, sorrow, and death. It is odd, then, that the individual socioeconomic consequences of the epidemic confound researchers to this day.

Men and women in developed countries need no longer fear the death sentence that used to come with an HIV-infection. Treatment is cheap, effective, and readily available. The life expectancies of the afflicted are no longer cut short. The social ramifications may still be significant yet diminish with the ever-growing awareness that seropositive patients are neither dangerous nor a burden, and that the only occasions which oblige caution are sex and injury.

The picture in sub-Saharan Africa or in other, poorer parts of the world is much bleaker. In 2019, sub-Saharan Africa houses most of the world's HIV patients. Despite this, two out of five do not receive treatment. These unfortunate souls do not expect to live for more than a decade from the onset of the first symptoms. People who *do* receive treatment must often expend a significant part of their income. According to the United Nations Development Program (UNDP), anti-retroviral therapy (ART) costs several hundred dollars a year at the time of writing – a significant sum in developing regions. During the period under study, which ends in 2010, prices were as high as ten thousand dollars a year. Those who did not devote funds to treatment still lost in income and opportunities. Either way, HIV patients in developing regions were and are likely to be poorer than their HIV-free peers. Beyond the individual level, infection creates both emotional and financial issues for the family once the symptoms of AIDS emerge.

Despite the obvious necessity for a strong understanding of how HIV affects wealth, few studies have focused on the consequences on one's income, productivity, possessions, or any other individual measure of material welfare. We use person-level data for seventeen countries over an eight-year period to estimate the effect of HIV-infection on socioeconomic status – in our case, a wealth index constructed by USAID. To remove endogeneity, we use a circumcision dummy as an instrument. This has been used before and is the sole instrument in the literature that receives support from multiple sources.

Moreover, we devote a significant amount of time to discussing the biases and potential pitfalls that we and future researchers will encounter when using person-level data for HIV and other deadly diseases. The most important of these biases is caused by the relationship between a person's wealth and how long they survive with HIV, which in turn causes wealthy HIV-patients to be overrepresented in the data as one increases the age of the sample. An important contribution of this paper is to develop a way to deal with such biased data and consider the heterogeneity of the age regressor. We demonstrate that if one does not divide the sample in different age groups, one is certain to obtain biased results. In addition, we present the effect of HIV not as a coefficient, but a variable which behaves differently for different wealth groups.

Our summarized approach to this issue is as follows: first, we create age-based subsamples and observe how the effect of HIV on wealth becomes positive as the average age of the sample increases. This, presumably, confirms the survival bias. Subsequently, we create subsamples for different wealth groups. By doing this, we study the effect of HIV as a variable instead of a fixed

constant. Among the wealthiest groups the survival bias appears non-existent and the effect of HIV small.

Our models show that the wealth effect of HIV is largest for poor people and then gradually decreases with wealth. Simultaneously, the survival bias is, insofar as we can tell, strongest among the poorest groups. Determination of the size of the effects, however, is hard and requires additional efforts. The trend – that initial wealth status itself determines how much one's wealth is affected by HIV – seems quite certain and is confirmed by different models and different subsamples, as we will go on to demonstrate.

The thesis is divided into two parts: the first part provides a theoretical foundation and discusses the literature, theoretical framework, data, biases, and the econometric model. The second part contains the empirical work and sets up models to deal with the survivor bias.

Map of countries under study

PART ONE

This part contains the descriptive text. It describes the literature and the nature of HIV as well as the nature of the data and potential pitfalls. It concludes with the creation of a three-stage least-squares IV model, which is the workhorse model for our empirical part.

In this first half of the thesis, we discuss how HIV impacts wealth and how the data is biased toward survivors – people who live longer because they can afford treatment. It is filled with lengthy lists of variables and biases, included for completeness. The objective is to provide as solid and complete a foundation as possible for the empirical work, which is discussed in part two and contains most of the 'creative' content.

SECTION II: LITERATURE REVIEW

Most relevant literature may be divided into two large groups. The first group uses macroeconomic data such as GDP per country and HIV prevalence ratios to estimate the national and international effects of an HIV epidemic. The second group of literature uses individual level data and focuses on the determinants of HIV.

In contrast, the literature which studies the economic effect of being HIV-positive on the individual level is small and does not use individual-level data. Mostly, the opposite causality is researched: previous HIV-related studies that focus on the individual level try to uncover the factors which increase the probability, causally or not, to contract HIV. To our knowledge, both our econometric approach and the direction of the causality is nowhere to be found in the literature.

We would offer an additional, general critique on the literature – namely that results are often not more sophisticated than a single coefficient. This is true, contradictorily, of studies which estimate the coefficients from mathematical models which explicitly assume a non-linear effect. Instead, the literature almost universally assumes linear effects and makes few attempts to relax the ceteris paribus assumption to show how their results change over important variables. Clearly, the odds of contracting HIV do not scale linearly with wealth or, far that matter, any other influence. In our study we attempt to remedy this by treating the effect of HIV as variable over both age and wealth and refrain from presenting a single, linear, effect. The papers in section II (b) where chosen specifically because they, to some extent, avoided such linearity assumptions.

Section II (a): The Macroeconomic Effects of HIV

The literature abounds with studies which use macroeconomic effects on a country- or international level to evaluate the impact of HIV. Because HIV-prevalence rates are highest in sub-Saharan Africa, much literature has focused on this region.

Cuddington (1993) employs an extended version of the Solow model to approximate the macroeconomic effects of the AIDS in Tanzania. By incorporating the demographic effects of AIDS, he concluded that barring a change in policy the epidemic could bring about a decline in GDP of 15 to 25%. Furthermore, he concluded that per capita income levels could fall by roughly 10% by 2010 as a direct consequence of the virus.

Bonnel (2000) used data from forty-seven countries between 1990 and 1997 to run a cross-country regression. His findings suggest that HIV had slowed down the rate of Africa's per capita income growth by a yearly 0.7 percentage points and noted the reverse causality between economic growth and HIV. He further illustrated that HIV led to less economic growth and that economic growth may reduce HIV prevalence rates via education, infrastructure, employment opportunities, and female empowerment. Conversely, economic growth can also increase prevalence rates through labor migration and social change. In contrast with previous research, Bonnel estimated the contemporary, rather than future, effects of the disease. These had by 2000 already transformed from being a health issue to an economic and demographic problem for affected countries. He concluded that the disease caused a vicious cycle in the afflicted countries: the disease increased poverty as it prevented ill people from contributing to economic growth, which in turn led to increased infection rates due to poverty.

Arndt and Lewis (2000) used a Computable General Equilibrium model (CGE) with data from South Africa between 1997 and 2000. They estimated the effect of the disease on variables such as labor supply, death rates and HIV prevalence. The authors used demographic data to simulate and compare two scenarios: a hypothetical no-AIDS-scenario (on the assumption the economy continues to perform as in previous years) and an AIDS-scenario (on the assumption that the HIV prevalence rate and related factors affect economic performance). The authors' results suggested that the South African GDP level would have been 17% lower in the AIDS scenario by 2010 compared to the no-AIDS scenario. Furthermore, in the AIDS-scenario the GDP per capita dropped eight percent by 2010. South Africa's ongoing economic struggles lends support to their work.

Young (2005) also used South African data from the Demographic and Health Survey (DHS) but came to an alternative conclusion with respect to the impact of HIV on future living standards. His article, titled *The gift of the dying*, took into consideration two competing effects of the virus. On the one hand, HIV would have an impact on the accumulation of human capital of orphaned children. On the other hand, high infection rates in communities would lead to a decline in fertility rates. The latter can happen directly, through fewer unprotected sexual encounters, and indirectly, as the reduction in labor increases the value of women's time. Hence, women might decide to work instead of bearing children. Young (2005) concluded that the decline in fertility has a bigger impact, and that future generations in South Africa will have a higher consumption rate per capita. We note that we do not reject these papers out of hand, although they may seem estranged from other literature about HIV and economics. After all, there are countries, even in sub-Saharan Africa, in which HIV epidemics have not held back further economic development.

However, some studies like Kalemli-Ozcan (2003) suggested that, contrary to Young's findings, fertility does not decline due to HIV, and other more recent research found only a small overall effect on fertility (Durevall and Lindskog, 2016).

Mveyang et al. (2015) examined the effect of HIV on economic growth for several countries in sub-Saharan Africa. They used male circumcision and distance to the first outbreak of HIV as instruments to address endogeneity. Their results suggest that the disease did not significantly affect the economic growth in the region, but that inequality had increased. This result is consistent with our own findings that the effect of HIV grows larger for poorer people.

Section II (b): The Determinants of HIV on the Individual Level

The determinants of HIV are important control variables in our model. Many factors which influence the odds of contracting HIV are also strong determinants of wealth. Therefore, the determinants of HIV are probable causes of omitted variable bias if they are not included in the model.

Iorio et al. (2016) examined the relationship between HIV prevalence rates and education. They used DHS data for 39 sub-Saharan African countries and used an algorithm to define five different stages for the HIV epidemic at different points in time. As not all countries

reached the peak of the disease simultaneously, they find themselves in different stages in various years. We will use these stages as a control variable in our estimations, given in table II. Iorio's results suggest a U-shaped education gradient for HIV. Hence, education has a positive impact on the probability to become HIV-positive in the early stages, then the effect of education becomes zero and turns positive again for the last stages of the epidemic.

Fortson (2008) constructed an HIV gradient with DHS data to evaluate the determinants of HIV in several sub-Saharan countries. Her results suggest that more educated individuals are more likely to be HIV infected. Individuals with six or more years of education were 50 percent more likely to be HIV-positive. The reason for this difference may lie with the fact that more educated individuals marry later and thus have more sex partners before marriage. As the more educated population is also more likely to be wealthy, there is a positive HIV-wealth gradient (Fox, 2010).

In conclusion, wealth and HIV are highly correlated variables which partly determine each other, which is why we must take seriously any concerns about endogeneity and omitted variables. Generally, the literature points to a negative macroeconomic effect on wealth and productivity, albeit with notable exceptions. On the individual level, studies are mostly concerned with the causes, rather than the consequences, of HIV. This discrepancy is addressed in this thesis.

SECTION III: THEORETICAL DISCUSSION

Section III (a) is a discussion of the different ways in which HIV can affect wealth and lists six possible channels. Section III (b) contains the more interesting part and describes how the survival bias affects the data.

Section III (a): The Effect of HIV on Wealth

The channels through which HIV affects the wealth of an individual or a household depend on many factors and fluctuate over time. It is imperative to possess a rudimentary comprehension of the different stages of the disease as the physical impact on the patient varies significantly. Indeed, throughout most of their lives with HIV, patients experience relatively few physical symptoms. HIV goes through three stages before the patient succumbs and dies.

According to AVERT, after an initial feverish state (stage 1) which lasts several weeks, the virus goes dormant for many years (stage 2). During this long period, which has been known to last up to fifteen years without treatment, the person exhibits few symptoms and is not physically inhibited from maintaining their income. At a later stage, the HIV virus develops into AIDS (stage 3), which dramatically affects the physical well-being of the patient. Without treatment, the person usually dies within one or two years following the onset of the initial AIDS symptoms. If a sick person has access to anti-retroviral therapy (ART) their lifespan will increase dramatically compared to those without treatment (Cohen et al. 2011). Most HIV-positive persons in our dataset will not find themselves in serious physical discomfort, and we thus assume that the physical effects negatively affect wealth among HIV-positive patients only to a minor extent. The wealth effect of HIV is more likely passed on through several channels.

(1) Decrease in human capital: Fortson (2008) argues that inhabitants of a region with a high prevalence of HIV are less inclined to invest in education. In her estimation, this effect was only shown for the younger cohorts and was especially strong for AIDS orphans. Because our study is limited to sub-Saharan Africa, this effect is unlikely to be strong as large numbers of people no longer go to school by the time that they could plausibly contract HIV, even though one could argue that individuals in sub-Saharan Africa are more likely to leave school early if they learn about the risk of infection or the reduced life expectancy. Furthermore, it is plausible that orphans in general will, on average, have received less education than non-orphans. Unfortunately, we do not have enough data to decide whether someone is an AIDS orphan or not. Moreover, papers that employ human capital theories have posited that the reduced life expectancy leads to unwillingness to invest in education in general, and that the effect is not secluded to orphans. The issue of human capital is discussed frequently in the literature, but we note that we believe this to only be a minor channel.
(2) Savings in anticipation of death: Freire (2002) used South Africa as an example to demonstrate that upon diagnosis many people save money to afford funeral costs and leave behind something for the remainder of the household. In this case, the individuals would spend less on material goods and save more. The wealth index, due to its construction, would decrease, though whether this constitutes a decrease in *actual* wealth is a topic of discussion.
(3) Treatment costs: For those who receive it, treatment is a constant, lifelong expenditure. According to Garmaise (2012), the cost of ART treatment has decreased significantly. In 2000,

annual ART treatment cost some 10 000 US dollars per person, while in 2012, the yearly cost could be as low 100 US dollars under the most favorable circumstances – e.g. that corruption or logistical inefficiencies did not increase costs. Median treatment costs for WHO-recommended drugs in low- and middle-income countries were as high as to 600 US dollars in 2007 and about 300 by 2012, and further declining thereafter Lange (et al., 2014). As we use data prior to 2010, we assume that the yearly cost of the treatment was high, especially to sub-Saharan African standards and that the poorest could not afford it. South Africa started offering ART treatment for free already in 2004 with, as we today know, limited success. The precise spread of ART treatments in sub-Saharan Africa is not well documented but general figures are available. These reveal a strong increase in treatment. According to Lange (et al., 2014), only 300 000 people in low- and middle-income countries received treatment in 2002. By 2012, merely a decade later, this had gone up to 9.7 million. In sub-Saharan Africa specifically, the number of treated patients grew from 50 000 to 7.5 million. Such a large increase in the poorest regions was made possible due to the availability of generic alternatives and lax patent enforcement, as well as large-scale aid initiatives. For the period under study in this thesis, 2003 – 2010, this means that there has been a dramatic increase in coverage and the money spent on treatment. We believe that this is the most important channel through which HIV affects wealth. Lastly, these numbers must be taken with some healthy skepticism. Sub-Saharan healthcare systems are not as robust as those in developed countries. Stock-outs are not uncommon and inept administering can lead to resistance to some treatments. It is important to keep in mind, therefore, that the phrase 'being on ART' does not consider the wildly varying quality of the treatment (Lange, et al., 2014).

(4) Indirect costs: These can be incurred through potential social fallout. Some communities harbor strong biases against HIV-positive people. Occasionally, ignorance plays a meaningful role. Our DHS data shows, for example, that a substantial part of Kenya's population was, in 2003, quite badly educated about how to prevent HIV. This may translate into an avoidance of seropositive people. This could also result in a lower wealth index – think a shopkeeper losing customers.

(5) Lifestyle changes: After a positive HIV test, it is well possible that people change their lifestyle to become healthier. Any sickness could be crucial for an HIV positive person, so we can assume that some individuals will refrain from excessive drinking and smoking or any other potentially dangerous and expensive lifestyles, especially as alcohol usage and risky sexual behavior are highly correlated (Kalichman et al., 2007). Because alcohol and cigarettes are luxuries in sub-Saharan Africa, teetotalism can lead to a higher income. However, individuals could also behave recklessly and indulge in an increasingly extravagant lifestyle after a positive HIV test. Stein et al. (2005) examined the behavior of seropositive patients and concluded that a high percentage of them was having unprotected sex, which lends some credibility to the claim that seropositivity does not stimulate a more responsible lifestyle *per se*.

(6) Increased investment: Infected individuals want to assure the wellbeing and health of their family. As the clanship system is very strong in most African tribes (Ankrah, 1993), the community will often assure a safe future for the household, especially the children, and sometimes even welcome the family into another family or create new households. This can influence the household wealth of an affected individual if the family is already benefitting from the clan system before the death of the individual. This support system could potentially lead to an unpredictable effect on wealth.

In general, it appears that the negative effects outweigh the positive ones. There is a strong case to be made that some channels may positively affect wealth – such as the adoption of a healthier lifestyle – but the net influence of HIV on wealth looks negative. The dominant channel depends largely on circumstance. In the first year of our dataset, 2003, for example, treatments costs were arguably negligible as treatment was virtually unavailable. Conversely, social costs are large in some societies and smaller in others. On the whole, treatment costs are most likely to have the largest impact.

Section III (b): Why HIV-positive Persons Grow "Wealthier" with Age

Technically, every person that has HIV in any of the countries under study has the same chance of being included in the data if one adjusts the weights. This is a problematic statement. People who can afford treatment and medical care will live much longer than those who cannot. Therefore, older HIV positive individuals are much more likely to be wealthy. This ensures that the data oversamples persons which can afford treatment as the age of the subject increases. Among the youngest people in our dataset, about up to the age of twenty-five, this is not a big problem as HIV is a slow-working virus and many people, even untreated, will live for at least several more years after the initial infection. However, as one studies older age groups, more and more HIV-positive persons who are wealthy enough to afford treatment will be included. Furthermore, due to their higher wealth, we assume that they also invest more on their health in general and therefore will survive longer, for example, they will eat more healthily.

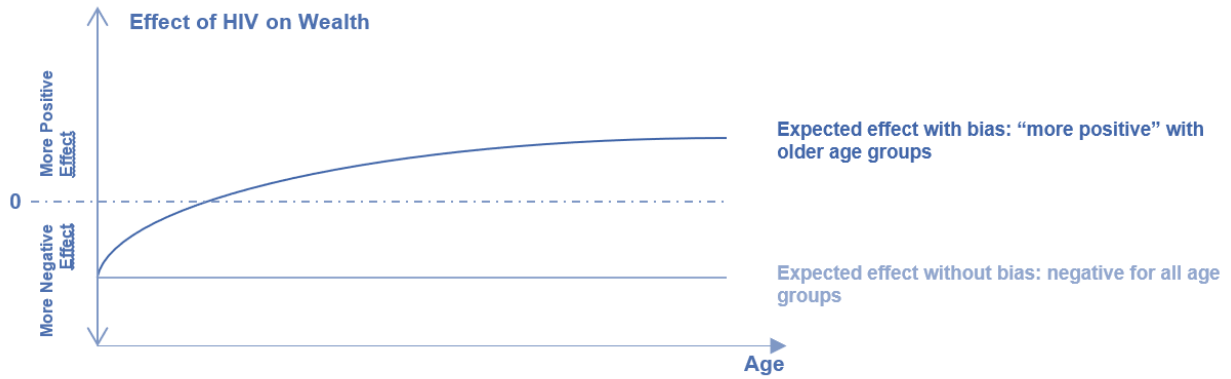
This wealth effect is magnified by the time period of our dataset, namely 2003 – 2010. Around this time ART treatments became widely available, as we discussed in section III (a), albeit often at a high cost. Before the widespread introduction of these treatments, wealth did not change one's lifespan significantly as doctors could do nothing other than to offer some relief and, occasionally, treat the symptoms of AIDS sufficiently well to marginally increase lifespan. However, there is a wide variation in how long individuals survive without treatment. This is mostly due to the environment, nutrition and wealth of the person. Even so, with the breakdown of the immune system, death was never far off once AIDS set in. In recent years, ART treatments have become cheaper at a fast pace, although they remain expensive for sub-Saharan countries. At the extremes, costs could be as high as 10 000 US dollars in 2000 and as low as 100 US dollars by 2012. Some snapshots of prices reveal how much prices could vary depending on time, location, and specific combination of drugs provided. In a report published in 2006, the WHO showed that different combinations could change treatment costs by an order of magnitude. By 2019, many Sub-Saharan African countries are offering free access to the ART treatment. The double combination of increasing availability yet high prices indicates that during the period we study (2003 - 2010) the wealth effect may be larger than in either the preceding or succeeding periods.

The relationship between wealth and health, disregarding HIV, has been well studied. There has been a growing literature regarding this relationship between health and wealth in Sub-Saharan Africa since the 1990s. Among others, Cockerham et al. (2000) noted that wealthy people live longer on average. Long and Deane (2015) also used DHS data from the same decade as we do, albeit for Tanzania. They show that wealthy people are more likely to be HIV positive, an effect

which is compounded by the fact that wealthy people live longer in general, regardless of HIV status. This is confirmed by numerous studies. This means that the older a person is when first contracting HIV, the wealthier, on average, he is too. Furthermore, we should note that the effects of being wealthy on lifespan even hold when ART treatment is unavailable as simply living in a more hygienic environment with access to doctors makes a difference for HIV-patients. After all, HIV drastically affects the immune system. Moreover, there are few "impoverished seniors" in our data to skew the results, as life expectancy in sub-Saharan Africa in 2005 was only 51 years (World Bank, 2019). Given an infant mortality rate of around 70 per 1000 births (World Bank, 2019), child mortality only slightly reduces life expectancy.[1] Lastly, Ataguba (et al., 2015) shows that there is a huge disparity in health conditions between the wealthy and the poor in South Africa due to infrastructure, employment, housing, et cetera.

We do, of course, use an instrument to make our HIV-variable exogenous. However, this variable is "HIV status" and indicates whether a person is positive or negative. How wealth affects a person's ability to deal with HIV is essentially *a different* variable, which does not exist in the data. Hence, while the odds of contracting HIV have been made exogenous from wealth, the effect of wealth on how HIV affects individuals remains. This issue is discussed further in section V.

This age bias appears to be so perverse that in many of our tests, HIV status appeared to have a positive effect on wealth as one increases the average age of the sample. This, of course, went squarely against our expectations as we assumed that people which have had HIV for a long time would be significantly poorer due to the cost of treatment. Instead, in an incorrect model, the reverse result comes out: because those who have managed to survive with HIV for a long period of time are significantly wealthier, incorrectly specified models tend to show that HIV has a positive effect on wealth.



[1] There are various ways to calculate life expectancy, but the most general way is to take the average age of death in the year of discussion. If we do this, then we can recalculate that if one excludes infant mortality, the life expectancy would increase from 51 to about 55: $\frac{51}{0.93} = 54.84$. Regardless of which specific measure of life expectancy of birth one chooses, the impact of infant mortality will not be more than a few years.

SECTION IV: DATA AND ECONOMETRIC MODEL

The first part describes the data and potential biases encountered. The second part introduces the variables and sets up the econometric model. The shape of our econometric model has been chiefly determined by the nature of our endogenous regressor, its instrument, and the bias created by the fact that wealthy people are over-represent in our HIV- positive subsample.

Section IV (a): Dataset

We use data from the Demographic and Health Surveys (DHS), which are funded by USAID, to conduct our study. This is individual data that contains several hundred person-specific variables for up to tens of thousands of people per country, although it can vary widely. Our data is from between 2003 and 2010 and includes only males. The limitation to these years and the exclusion of women is a consequence of our instrumental variable and the availability of HIV testing, which we discuss in section IV (b).

Table I: Countries included in the Dataset

All included sub-Saharan African countries.

Burkina Faso	2003
Cameroon	2004
Democratic Republic of the Congo	2007
Ethiopia	2005
Guinea	2005
Kenya	2003, 2008
Lesotho	2009
Liberia	2007
Malawi	2004, 2010
Mali	2006
Niger	2006
Rwanda	2005, 2010
Senegal	2005
Sierra Leone	2008
Eswatini (Swaziland)	2004
Zambia	2007
Zimbabwe	2006

Notes: All countries were selected for years *before* mass circumcision was introduced as a measure against the spread of HIV. As such, circumcision status is exogenous to HIV status.

Using DHS data has several advantages. Firstly, the sample is quasi-perfectly representative of the population of the country, at least to the extent that it is practically possible to create such a sample. Another advantage is that it oversamples certain categories that could be of special interest.[2] This overrepresentation can then be adjusted for by using weights, which are provided. Thirdly, DHS data is commonly used in studies on HIV-related topics, meaning that the validity of the data has been well established in much previous literature. Lastly, DHS data is specifically designed to be compatible across different countries and years.

One issue with the study of HIV is that often only a 'small' part of the population has it, which results in a small sample. For some countries, we have only a few thousand persons in the dataset, of which perhaps only ten percent have HIV. This small subsample of only a few hundred HIV-positive persons severely limits how many factors one can control for. This is especially true in Africa, where societies are rarely homogenous and very diverse linguistically, culturally, geographically, et cetera. In order to control for these elements, we require many control variables, and thus as large a sample as possible. Moreover, by including different countries we can more easily control for a variety of the macroeconomic effects of HIV. This is especially important since many of the wealth-related effects of an HIV epidemic are often macroeconomic effects rather than microeconomic ones. We achieve this large set by combining twenty datasets into a single, large dataset containing over 110 000 men for a total of over a hundred million datapoints. The countries and years included are given in table II.

Section IV (b): Discussion of Biases and Limitations

Generally, DHS data is free of sample bias and assembled according to a very high standard. Any observation may be considered fully random: except where specified otherwise, any person in the dataset has the same chance of being included as any other person in the country. But while the data adheres to high standards, the nature of the topic is such that a myriad of biases are nonetheless present. The most important ones are discussed in this section.

(1) Endogeneity: The most obvious hazard is the simultaneous causality between wealth and HIV. The odds of a person contracting HIV are determined by many factors, including their social environment, age, income, educational status, culture, et cetera. As shown by Fortson (2008), many of these are influenced, directly or indirectly, by wealth. Even the usage of condoms is partially determined by wealth in the poorest countries, as wealthier individuals tend to use condoms more than poorer individuals – although the effect declines when controlling for knowledge about HIV (Ukwuani, et al., 2008). The issue of endogeneity is worsened because we do not know the direction of the bias. In the countries which suffer most from the HIV epidemic there is no observable correlation between wealth and HIV. Mishra (et al., 2007) showed that wealth increases the odds of getting HIV.[3] To the contrary, one can

[2] For example, if a country has a small HIV-positive population, the DHS can overrepresent HIV positive persons to ensure that there is a meaningfully large sample present in the dataset.

[3] More densely populated areas are wealthier and fertile ground for viruses such as HIV, while more educated individuals get married later and have more partners before their marriage.

argue that wealth reduces the chances of contracting diseases. For example, wealthier people might have more knowledge about HIV and can afford condoms more easily, et cetera.

Hence, the OLS estimates will be severely biased, and we cannot make any conclusive statements about the bias. Denote β as the true effect of HIV on wealth. We know that endogeneity bias in OLS is given by:

$$\hat{\beta} = \beta + \frac{cov(x,u)}{var(x)}$$

The direction of the bias is exclusively determined by the covariance which, by assumption, is nonzero. Since u is unobserved and there is no conclusive theory, we cannot say whether the bias is positive or negative, let alone make any statements about its size.

The ways in which wealth can determine at least partially one's odds of contracting HIV are too many to describe here, but the endogeneity issue is one of the main obstacles facing HIV-related research in general. The gravity of the situation is compounded by the fact that there are only very few instrumental variables available to us. In our case, we use male circumcision status as an instrument, leading to severe limitations on the data that we use. This issue is discussed in-depth later.

(2) Lack of data for women: Since we use male circumcision as our instrumental variable, the dataset is limited to exclusively men – less than half of the original observations. This obviously incurs a strong limitation, given that the effect on wealth of having HIV are potentially different for men and women as they often play vastly different roles in the household. Türmen (2003), evaluated the difference in genders with people that had the virus, as females are more likely to get infected. When using an instrument like male circumcision, as previously done by Mveyange et al. (2015), this creates a bias because women are not taken into consideration. However, we do note that according to Atkinson (1971) the "male head of the household is the representative individual," which from a wealth perspective is important. Even though this might not be true in the much of the Western world anymore, in most African countries the head of the household is still traditionally male. The DHS data also supports this point.

(3) Self-Selection: Self-selection bias is also present in our dataset. Most of our sample has undergone HIV testing, but a small subset of about 12% has not. While some of these individuals refused to take the test, some others were coincidentally not at home. This means that there are some ten thousand men in the data which, despite being offered a test, did not take one. This incurs an obvious self-selection bias. Usually, such an issue could be remedied by using some Heckman-type correction method. Unfortunately, this method does not work for discrete variables such as our main regressor, HIV status.

Moreover, the Heckman method requires us to use yet another instrumental variable, which risks increasing the inaccuracy of our estimates. Self-selection is not much discussed in other literature which employs DHS data. In the end, we judged it best to simply leave the self-selection issue be.

However, we examined the data to reveal more about the people who refused to take the test. In our dataset, we have a refusal rate of 11 percent. The refusal group is slightly more inclined to live in a rural area than in an urban area. Furthermore, the refusal group is on average 8 percent wealthier than the average individual and is more educated. We can assume that more educated, richer individuals are less trustworthy in institutions or are more concerned about their data. Nevertheless, across several variables we don't see enough deviation to significantly bias our estimations. In addition, our estimations, as we will later show, uncover some very pervasive trends among a variety of samples that would be hard to change significantly due to self-selection.

(4) Over-sampling: DHS data is not perfectly random: some groups are over- or under-sampled to ensure large enough samples of small-scale phenomena. This, again, biases the data one way or another – even though we do not know how. All the regressions in this study take this into account by applying the appropriate weights, which are included variables in the data. As such, over-sampling is no issue and does not cause a bias in our model.

(5) Household Wealth Index: The last 'bias' in our data is not so much a bias as it is an issue which one needs to be aware of. In sub-Saharan Africa the notion of individual wealth is not as well-developed as in wealthy nations. 90% of the population in developing countries is assumed to be unbanked at the end date of our dataset (Hinson, 2011), and the houses are typically owned by the household and not a single individual. Hence, our concept of measuring the wealth of an individual by measuring the wealth of a household is the 'best' way of determining how well off a person is from a material standpoint. While it may appear to be 'wrong' to look at the impact of HIV on household wealth, there is no alternative because most possessions are typically shared between all members, and there often is not enough income to save money on the individual level. In other words, the nature of many societies and economies of sub-Saharan Africa precludes, for many, the ability to create their own wealth. From that point of view, it may not even be desirable to find an individual measure of wealth for our dataset, even if it were possible.

This does mean that the effect of HIV on wealth becomes diluted because there are many members in the household which also contribute. A household consisting of a single individual will suffer much more heavily if that person gets HIV, while the wealth of a household of twenty persons may not be affected at all. We do control for this in our model. The fact that we use household wealth instead of some measure of individual wealth does not threaten the validity of our model.

These are the main biases and limitations that we have uncovered in our data and we act upon them to the best of our ability. The exclusion of women and the self-selection issue are the two main limitations that we have not corrected or altered our model for, mainly because doing so would introduce more severe problems, or because a correction is not possible. Over-sampling issues are swiftly rectified. For endogeneity we use an instrumental variable, and the bias introduced by the fact that wealthy people are overrepresent in the HIV subsample is considered in our model, as discussed in section III (b) and throughout section V.

Section IV (c): Variables

This subsection briefly discusses the form of our dependent variable (household wealth index), endogenous regressor (HIV status), instrument (circumcision), and discusses our approach toward control variables.

(1) Independent Variable: The DHS wealth index is created by surveying the material possessions of the household. Items included are the nature of construction (such as wall and floor materials), electronic equipment, plumbing and sewage, animals, et cetera. It contains hundreds of elements, although this can vary widely from country to country. In addition, the size of the distribution of the wealth index varies from country to country. Consequently, we cannot simply append different countries together. Another issue is that not all countries are equally wealthy. Appending a poor country with a low HIV rate to a wealthier country with a high rate can dramatically bias the estimates. We solve these issues by replacing the nominal distribution of the wealth index with wealth percentiles: every person's household wealth is given relative to the wealth of the rest of the population of the country. This means that there is an equal amount of people in each percentile, divided proportionally by country. If we then estimate the effect of HIV on wealth, our coefficient will tell us how many percentiles up or down the wealth ladder a household will go. An additional albeit minor advantage is that we know the exact distribution of our dependent variable when we segment our sample, as we will do in our empirical part.

    After standardizing the index through percentiles, we can simply append the data for all countries. This means that lower percentiles can include persons that are wealthier than some included in the top percentiles, simply because they live in a wealthier country. However, we are interested how wealth is impacted relative to a person's immediate environment. If, for example, we used a monetized index and our model would show that HIV has a negative annual effect of fifty dollars, this would be a meaningless result as fifty dollars is nothing in some countries and a monthly wage in others. By contrast, working with percentiles makes countries much more comparable. Lastly, we note that this uniform distribution of the dependent variable does not violate OLS assumptions and has no impact on point estimates or confidence intervals.

(2) Endogenous Regressor: Our HIV variable is simply a dummy which indicates whether a person is HIV negative or not. The interpretation is straightforward, but its discrete nature means that we cannot use a simple two-stage least squares regression. This is discussed in detail in part section IV (d). Moreover, all observations with inconclusive tests or tests for the HIV-2 virus (a rare variant found in Western Africa) were deleted from the set.

(3) Instrument: Because our HIV-variable suffers from serious endogeneity issues, we obviously need an instrument. One issue is that we are limited to variables included in the dataset, as there is no other data that identifies individuals compatibly with the DHS data.

    Before 2010 men we mainly circumcised as a matter of culture and tradition that dates back decades or even centuries. It is very unlikely that the decision to circumcise was determined by wealth. Thus, any correlation between wealth and circumcision, which can most definitely

be found, can be taken to be non-causal. The long history and tradition behind circumcision across a variety of vastly different cultures strengthens the exogeneity of a circumcision variable.

Male circumcision has been used as an instrument to account for the endogeneity of HIV by Mveyange (2015), Kalemli-Ozcan (2006) and Werker (et al., 2006). As an instrument, circumcision must be correlated with HIV, but cannot be causally correlated with wealth. Circumcision has been proven to be quite effective in reducing the risk of contracting HIV. Over 40 studies have shown a negative effect of male circumcision on HIV prevalence rate. According to Szabo et al. (2000), male individuals who are circumcised are 2 to 8 times less likely to get HIV infected. Medical research suggests that the lower infection rates are due to the removal of HIV receptor in the foreskin. As this is the main point of entry for the virus, circumcision basically makes it harder for the virus to enter a body. Almost 50 percent of the HIV patients are male (women's prevalence rates are slightly higher in our data) and most of these men get infected through sexual intercourse. Circumcision, then, makes for a good tool to shrink infection rates. According to AVERT, circumcision only prevents the transmission of HIV to the circumcised men but does not protect the female or male partner of these men.

In many cultures circumcision at a young age is a common tradition, especially in Islamic countries. Murdock et al. (1959) researched the circumcision practices in Africa among the different ethnic groups and provided a good overview. This data is prior to the outbreak of the HIV epidemic. According to Murdock circumcision rates were around 50% in sub-Saharan Africa.

Bailey (2007) noted the different HIV prevalence rates across sub-Saharan African countries. It is puzzling to see that the prevalence rate has been relatively stable in some countries, while it exploded and is still increasing in other countries. The latter are typically called the "HIV Belt." There seems to be an inverse correlation between HIV prevalence rate and traditional circumcision rate (Moses et al. 1990).

In 2008, the World Health Organization recommended mass circumcision as a policy to contain HIV epidemics. Since then, many sub-Saharan countries experienced large increases in circumcision rates. The problem here is that HIV becomes causal to being circumcised, which invalidates its use as an instrument. This is the reason that we do not use data more recent than 2008, unless we could determine that there was no increase in the rate of circumcision. This was only the case for Lesotho (2009), Malawi (2010) and Rwanda (2010), where circumcision rates did not increase from the previous years. For all other countries, circumcision rates either increased drastically due to the new policy, or there was no data available to make an informed judgement, and we consequently excluded these countries.

We must discuss the potential threats to the exclusion restriction of our instrument. Religion could validate the relevance of our instrument if circumcision is done as a religious tradition which is not affected by wealth. This could be the case for the Muslim population, were religion is not determined by wealth. The DHS data shows that areas with a high percentage of circumcision rates are also typically home to many Muslim households. Therefore, we control for that with a religion variable, which indicates if a person is Christian, Muslim, et cetera.

The same situation occurs if an ethnicity performs circumcision as part of their culture, as this might also affect the economic outcome of the household. Michalopoulos and Papaioannou (2013) found evidence that the ethnic regional centralization and the economic performance of a region are highly correlated. Ethnicity did not work for our estimation, as it was often missing, so we used the variable language as a control instead. In Africa, each ethnic group has their own tribal language, which means that there are often several languages in a region.

The male circumcision variable is simply a dummy, and any observation whose values was not either 'circumcised' or 'not circumcised' was deleted.[4] This was only a tiny minority of observations. Altogether, the case in favor of circumcision as an instrument is much stronger than for other candidates.

Following this discussion of our main variables, we now briefly discuss the control variables and their selection. Both wealth and the odds of contracting HIV are subject to an immense amount of diverse influences. The risk of significant omitted variable bias is therefore quite large, and we must control for as many factors as possible. This merits a discussion on controls that were included, as well as several controls that were unavailable or deliberately excluded. The main reason for including a control is to adjust for omitted variable bias. As such, most controls were included based on strong correlations with both the dependent variable and instrument, and the regressor.

Table II lists our control variables. Note that most of these variables are uniform across different DHS datasets. Dozens of additional control variables were considered and tested, mainly regarding cultural attitudes, but were often discarded for a variety of reasons. The chief criterion for any control is theoretical relevance, and each control must remove omitted variable bias from either the main regressor or the instrument. As such, we do not include controls that are indeterminant of HIV, wealth, or circumcision. Doing so is neither theoretically interesting nor practically wise. There can also not be any technical issues. Control variables must be compatible across datasets and cannot result in the dropping of too many observations, as this would create a sample bias. Furthermore, there are several excluded control variables. These are discussed below.

(1) Any direct identifier of ethnicity: The DHS dataset does, in fact, contain a variable called 'ethnicity,' but the number of binary variables required was utterly out of proportion. We estimated that for our largest dataset over a thousand different ethnicities were included. Normally, the solution to this problem is to combine similar ethnicities into larger groups. This requires expert knowledge about local tribes and cultures – knowledge which we do not possess. To mediate the situations, we spend a significant effort creating a variable which controls for 67 languages. This combined with country, region, and religion dummies should compensate for the absence of a variable for ethnicity.
(2) Data related to anti-retroviral treatment (ART) or patient history: We mentioned earlier that availability and cost of ART therapy changed significantly over the years covered in our dataset. Moreover, the ability to pay for therapy plays an important role in the interpretation of our model. Sadly, no variables that provide direct information about ART are included in the

[4] E.g. "don't know."

DHS data. We compensate by including an HIV-age interaction term, as we will discuss in section IV (d). Furthermore, the gradual decrease in price of ART should be captured, at least in part, by our time variable.

We observe a high R^2 ratio with many different sets of control variables. Obviously, a high R^2 is not our immediate goal, but higher ratios indicate fewer omitted variables and therefore less implicit bias, provided the model is correctly specified. Moreover, if one assumes that there are many joint determinants of HIV and wealth, then it is expected to see at least reasonably high levels of R^2 with the inclusion of the correct control variables.

Table II: Included Control Variables

This table lists *all* control variables used at any time throughout the paper.

Variable	Level	Type	Notes / values
Age	Individual	Continuous	Also included as squared term
Country	Macro	Categorical	
Year	Time	Continuous	
Household size	Household	Continuous	
Relationship to head of household	Individual	Categorical	Head of household Son of head
Urban-rural	Individual	Categorical	Urban Rural
Religion	Individual	Categorical	Catholic Protestant Muslim Traditional/animist No religion Other
Highest Educational Level	Individual	Categorical	No education Primary Secondary Higher
Language	Individual	Categorical	67 languages
Region	Macro	Categorical	~ 100 regions
Total children ever born	Individual	Continuous	
Stage of HIV epidemic	Macro	Categorical	Stage 0 Stage 1 Stage 2 Stage 3 Stage 4+
De facto place of residence	Individual	Categorical	Capital, Large City Small City Town Countryside
Ethnicity	Individual	Categorical	*Missing countries:* Lesotho, Zimbabwe, Rwanda, Liberia, Swaziland
Number of deceased children	Household	Continuous	
Dataset ID	Macro	Categorical	Identifies the original dataset
Marital Status	Individual	Categorical	

Notes: Not all controls are included in all regressions. Ethnicity is excluded unless specifically mentioned otherwise.

Section IV (d): Econometric Model

We have already excluded the use of a simple OLS model due to endogeneity concerns. The next logical model is a two-stage-least-squares regression. Even though our endogenous variable is binary, this should produce consistent estimates even if there is some misspecification in the first stage (Angrist et al., 2001). This is especially useful, as it is hard to fully correctly specify any regression related to HIV and wealth. Both HIV and wealth are determined by so many different influences that it is very hard to exclude all omitted variable bias in the first stage, even though our instrument is well-defended in the literature. The downside of this model is that it is biased (Adams et al., 2009).

The next logical option is a probit or logit first stage, followed by an OLS second stage. This is technically a valid model, but the first stage must be specified almost exactly right, or one risks severe inconsistency. Angrist (2009) mentions that unless this is the case, it is still better to use a simple 2SLS.

The solution is offered by Adams et al. (2009). They propose a three-stage model which combines the advantages of both the normal 2SLS and a probit IV regression. In the first stage, HIV is regressed on circumcision using a probit model. The second and third stage are a normal 2SLS model, but instead of using circumcision as the instrument, we use the fitted values from the probit stage. A vector of control variables C is included throughout. Denote x as our binary endogenous variable, $\ddot{x}$ as the fitted values from the first stage, and $\tilde{x}$ as the fitted values from the second stage. Further, the instrument circumcision is denoted as z and the main dependent variable household wealth as y. Φ is the cumulative probability function.

$$Pr(x = 1|z, C) = \Phi(\theta_0 + \theta_1 z + C\delta)$$

$$x = \alpha_0 + \gamma\ddot{x} + C\eta + \epsilon$$

$$y = \alpha_1 + \beta\tilde{x} + C\kappa + u$$

In this model, the coefficient $\hat{\beta}$ will tell us how many percentiles a household moves up or down the wealth index due to someone having HIV. Note that $\ddot{x}$ and $\tilde{x}$ are continues variables denoting the probability that someone will be HIV-positive.

This is the core model that we will use in this study. However, in its current shape, the only bias accounted for by this model is the endogeneity, which is removed by the instrument. We will now also modify the model in order to attempt to isolate the bias caused by the early deaths of poor HIV-positive persons.

We noted earlier that the data is biased since wealthy persons are overrepresented in the subsample of HIV-positive persons. In an ideal world with perfect data, this problem would easily be accounted for by including a variable which measures how long the person has had HIV. If we do this, then we would see that persons who have had HIV for a very long time are not, in fact, poor, and the regression could 'take into account' the fact that people in older age categories have had HIV for a much longer time than young people. The bias caused by surviving for longer due to being able to afford treatment would then be isolated in this interaction term, along with some other effects. This data is not available to us, and there is no way to reconstruct this variable

using averages. We can, however, attempt to isolate this bias by using an interaction term which relates HIV to age instead. We rewrite the third stage as:

$$y = \alpha_1 + \beta_1 \tilde{x} + \beta_2 \tilde{x} \cdot age + C\kappa + u$$

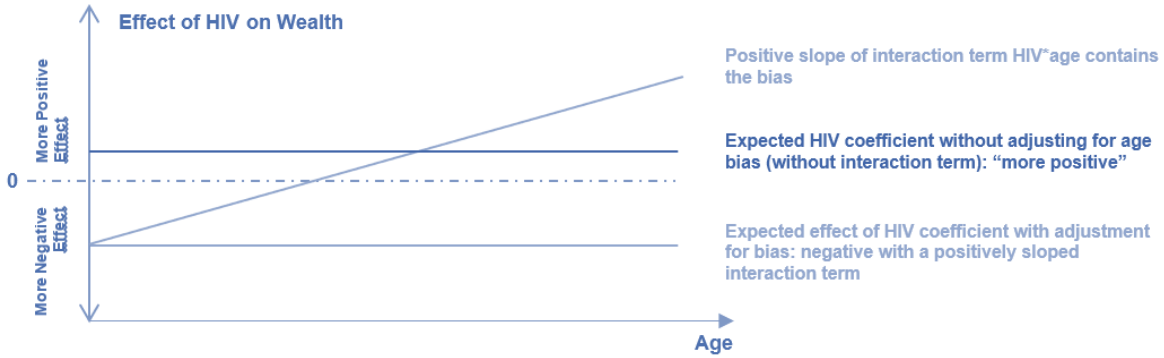
Now, the total effect on wealth is given by $\hat{\beta}_1 + \hat{\beta}_2 \cdot age$, instead of just $\hat{\beta}_1$. The idea here is that age works as a proxy for how long a person has been infected. Clearly, this is not accurate. We do know for certain, however, that the amount of years that someone has been infected must be smaller than their age.

The bias caused by over-representation of wealthy persons is, in essence, a case of omitted variable bias: the variable which shows how long a person has had HIV is omitted, and part of the effect is transmitted to the HIV dummy and the age variable. Therefore, introducing the interaction term reduces the bias, although it does not completely remove it.

We can now say something about the size of the bias by calculating at which age, on average, the net effect of having HIV becomes positive: if this age is early, then there is an indication that poor people are very underrepresented in the HIV-positive subsample. If the slope is steeper, then the association with HIV becomes rapidly more positive with age, indicating that wealthy persons are rapidly becoming overrepresented. Some early exploration of the data did indeed show that the correlation between HIV and wealth becomes increasingly positive as one looks at older age groups.

Table III: Regression Models

IV Stage	Model I: Excluding Interaction Term	Model II: Including Interaction Term
1	$Pr(x = 1\|z, C) = \Phi(\theta_0 + \theta_1 z + C\delta)$	$Pr(x = 1\|z, C) = \Phi(\theta_0 + \theta_1 z + C\delta)$
2	$x = \alpha_0 + \gamma\ddot{x} + C\eta + \epsilon$	$x = \alpha_0 + \gamma\ddot{x} + C\eta + \epsilon$
3	$y = \alpha_1 + \beta\tilde{x} + C\kappa + u$	$y = \alpha_1 + \beta_1\tilde{x} + \beta_2\tilde{x} \cdot age + C\kappa + u$



Notes: The vector of control variables can vary, which changes the sample size and may introduce some bias.

The previous part covered the theoretical foundation required for an in-depth study. This part focuses on the empirical work and contains most of the 'creative' content.

We have discussed how survivor bias affects the sample: as one increases the average age of the sample, the HIV-positive population becomes wealthier. This then leads to erroneous statistical results. In this second half of the thesis, we will study this effect with real data.

In the first instance, we run our two IV-models, including and excluding the HIV-age interaction term. This shows, as expected, a positive effect of HIV over age – a strong indication that the sample indeed becomes biased. To further investigate this, we create a new type of model (rolling window), in which we use our IV model on different subsamples over different ages to make any age heterogeneity irrelevant. We then use this technique to show that the age-related survivor bias does not appear among the wealthier groups in society. This strengthens the theory that wealthy persons live longer because they can afford treatment, and in turn bias the sample.

If we then look at how different wealth groups are affected by HIV, the wealthy appear unaffected and the poor heavily affected.

PART TWO

- Simple IV regressions
- Age heterogeneity
- Rolling window over age
- Rolling window over age and wealth
- Integrations over different wealth groups
- Discussion on robustness

SECTION V: EMPIRICAL WORK – IV and ROLLING WINDOWS

In this part we run our models many times with different subsamples in order to uncover trends in the data that are not self-evidently detectable from a single regression output. Notably, our hypothesis that the data is biased toward a positive age-effect due to the survival of wealthier people needs to be studied carefully. All models include the same set of control variables and differ only in terms of the subsample. We also note that for every coefficient or set of coefficients estimated in this paper, our three-stage instrumental variables approach was used.

The first part studies the output of the instrumental variables regressions. Thereafter, we discuss the age heterogeneity in the sample and propose a rolling window model to address both the heterogeneity and the survival bias. This is then followed by a more advanced rolling window model which also considers the wealth distribution. Finally, the chapter is concluded by developing a summary statistic for the impact of HIV on wealth for groups of differing economic status.

Section V (a): Simple Instrumental Variables Models

In this section we run model I, which is our three-stage IV model which excludes an interaction term, and model II, which includes one. In both cases, circumcision was considered a strong predictor and easily passed weak identification tests. The theoretical aspects of these models are discussed in section IV (d).

We begin by running the regression model without the age-HIV interaction term. The results are presented in table IV. We observe that the coefficient on HIV is insignificant: clearly, a straightforward instrumental variables regression does not uncover any effect of HIV on wealth at all. Most of the variation is explained by factors such as education, which is one of the strongest controls, and whether one lives in an urban or rural environment. We note that most of the regressions in this thesis give fairly similar results for the controls: things such as education and environment generally receive large, significant coefficients, while age (the control, not the interaction term) does not. There are some exceptions, but none of these show us anything useful. Further, note that since we have a substantial amount of control variables (around one hundred), we decided to summarize them and simply present how many within each category were found to be significant. Some were dropped, as detailed in tables IV and V. The controls are not generally of interest to us except in their ability to remove omitted variable bias. and we will not discuss them much. The same set of controls is used to ensure that results are comparable across regressions and prevent any unintended bias from creeping in. Moreover, it also shows that the model is consistent under many different samples. Circumcision appeared to be a very good instrument and yielded a Cragg-Donald F-statistic of 2458 – well above the cut-off value of 10.

The fact that the coefficient on HIV is insignificant is already a strong indicator that the effect is unlikely to be large. However, the dataset is sizeable, the area under study huge. A simple regression will not do. Let us run model II, which includes the age-HIV interaction term. The results are given in table V. The Cragg-Donald F-statistic is 927. This time the results show an entirely different picture: the main effect of HIV is quite negative, and the interaction term is positive. Both

are highly significant which lends some support to our hypothesis that the subsample of HIV-positive patients becomes biased toward the wealthy as age progresses. Specifically, the effect estimated by model two, which is shown in the graph in table V, is given as:

$$\Delta\, Wealth = -16.87 \; + 0.42 \; \cdot age$$

One way of looking at this is that the effect of HIV is neutral, overall, among people aged 41 as $-16.87 \; + \; 0.42 \; \cdot 41 \; = \; 0$. In the same vein, people younger than that should experience negative effects of HIV and people older than that should experience positive effects. Intuitively, one might think that having HIV starts out as a bad flu and after a few decades transforms into a pleasant, wealth-enhancing experience. Naturally, this is absurd. The only way that these estimates can be right is if these statements are true for the different age groups on *average*, and do not otherwise describe the process a seropositive patient goes through. A better way to look at this output is to understand that the HIV subsample of young people is, on the whole, poorer than their HIV-negative peers, while the opposite is true for older persons. However, this confirms only part of our theory: that older age groups are biased toward positive coefficients. This, we suspect, rests on the fact that wealthy people are more likely to survive longer with HIV and thereby become overrepresented. Thus, confirming our theory also requires that we test that HIV does not significantly impact the wealthy.

Table IV: MODEL I OUTPUT

Independent Variable: Household Wealth Index, 100 percentiles. **Sample size:** 55 730 males **Age:** +15yo. **Centered R^2** = 0.55. For the **categorical variables**, we report how many categories out of the total number were significant at the five percent level (e.g. three out of five religions had a significant coefficient). Brief explanations for those variables are given in the notes under the table. ***In this model HIV has no effect on wealth.***

Cragg-Donald F-statistic of the instrumental variables stage: 2458

Variable	Coefficient	p-value	lower 95%	upper 95%	Note
HIV	0.06	0.969	-3.09	+3.22	Dummy
Urban-Rural	-25.00	0.000	-26.28	-23.72	Dummy
Country	-	9/11[8]			Categorical
Language	-	23/52[1]			Categorical
Province	-	69/99[2]			Categorical
Epidemic Stage	-	1/4[3]			Categorical
Dataset ID	-	1/13[4]			Categorical
Education Level	-	3/3[5]			Categorical
Environment	-	3/3[6]			Categorical
Religion	-	3/5[7]			Categorical
Year	2.05	0.000	+1.23	+2.89	Continuous
Provincial HIV rate	2.07	0.964	-86.74	+90.889	Continuous
Dead Children	0.02	0.842	-0.22	+0.27	Continuous
Household Size	0.37	0.000	0.31	+0.42	Continuous
Children Born	-0.40	0.000	-0.51	-0.29	Continuous
Age	-0.06	0.317	-0.18	+0.06	Continuous
Age^2	0.0024[9]	0.005	+0.000719	+0.004042	Continuous
Constant	-4065.34	0.000	-5733.10	-2397.59	Constant

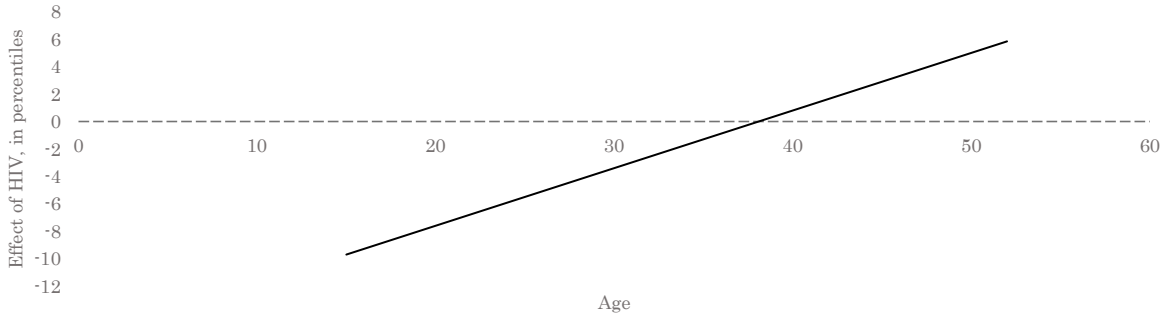
[1] These 52 include languages dropped due to multicollinearity. [2] These 99 include provinces dropped due to multicollinearity. [3] Three multicollinear, presumably with country. [4] All but one omitted. [5] All education levels are highly significant (p = 0.000) and are among the largest positive coefficients in the regression. [6] All were negative between one and nine percentiles of the wealth index. [7] Positive and negative values with five (absolute) percentiles of the wealth index. [8] All coefficients were very positive with more than ten percentiles. [9] The average age in our dataset is 30 years old, thus a ballpark-accurate interpretation is 30^2 * 0.0024 ~ 2 percentiles.

Table V: MODEL II OUTPUT

Independent Variable: Household Wealth Index, 100 percentiles. **Sample size:** 55 730 males **Age:** +15yo. **Centered R^2** = 0.55. For the **categorical variables**, we report how many categories out of the total number were significant at the five percent level (e.g. three out of five religions had a significant coefficient). Brief explanations for those variables are given in the notes under the table.

Cragg-Donald F-statistic of the instrumental variables stage:927

Variable	Coefficient	p-value	lower 95%	upper 95%	Note
HIV	-16.87	0.00	-27.70	-6.04	Dummy
HIV*age [9]	0.42	0.00	+0.16	+0.68	Continuous
Urban-Rural	-25.12	0.000	-26.41	-23.83	Dummy
Country	-	8/11[1]			Categorical
Language	-	28/52 [5]			Categorical
Province	-	70/99[4]			Categorical
Epidemic Stage	-	1/4[7]			Categorical
Dataset ID	-	1/13[8]			Categorical
Education Level	-	3/3[3]			Categorical
Environment	-	3/3[6]			Categorical
Religion	-	3/5[2]			Categorical
Year	2.10				Continuous
Provincial HIV rate	0.54	0.991	-89.14	+90.23	Continuous
Dead Children	0.07	0.007	-3.20	-0.49	Continuous
Household Size	0.36	0.000	+0.31	+0.42	Continuous
Children Born	-0.39	0.000	-0.50	-0.28	Continuous
Age	0.01	0.876	-0.12	+0.13	Continuous
Age^2	0.001[10]	0.258	-0.0008	+0.0029	Continuous
Constant	-4156.017	0.000	-5829.24	-2482.79	Constant



[1] All coefficients were very positive by more than ten percentiles. [2] Mix of both negative and positive coefficients smaller than five percentiles. [3] All were positive and significant. Education is arguably the most important control for wealth in both models. [4] These 99 include provinces dropped due to multicollinearity. [5] These 52 include languages dropped due to multicollinearity. [6] All were negative and like in model I between one and nine percentiles of the wealth index. [7] Three multicollinear, presumably with country. [8] All but one was collinear, presumably with country (most countries have only one year). [9] Note that this variable takes a value of zero for HIV-negative persons, and the value of *age* for HIV-positive persons. [10] The average age in our dataset is 30 years old, thus a ballpark-accurate interpretation is $30^2 * 0.001 \sim 1$ percentile.

One way to superficially test whether the wealthy are less affected than the poor is by running the same models on a subsample of the wealthiest ten percentiles. The output can be found in the appendix. Doing so reveals insignificant coefficients on both the HIV coefficient and the age-HIV interaction term. This procedure, however, is problematic because the sample window is too narrow:[5] if only the ten wealthiest percent are included, then there are bound to be a number of persons who 'jump the border' between the 89th and the 90th percentile, which in turn biases the estimates. In other words, there may be several HIV-positive patients in the top ten percentiles which were *outside* of this group before the became ill, and conversely there may be several HIV-positive patients outside the top ten wealth percentiles which were *inside* this group before they were infected. In this case, the reliability of the estimates depends on the true effect of HIV on wealth: if the true effect is very small, the results are reliable. If the true effect is large, then the results are no longer interpretable. In a later model, we will use this technique with a very wide window so that almost one-third of the width of the wealth index is included, and the bias is minimized. For now, even though the regressions on the top ten percent wealthy confirm our theory, we must discard it on the grounds that it is statistically unreliable.

Section V (b): Age Heterogeneity

At the end of section V (a) we attempted to uncover the relationship between age, wealth, and HIV by running the models on the wealthiest persons and then compare the coefficients with the original models. Because the sample window – the top ten wealthiest percentiles – was too narrow, we knew that the results were not reliable.

This means that we must find another way to study the effects over wealth *and* age. At this point during our work we realized that we had overlooked a fairly obvious feature common to many sub-Saharan countries: the skewed age pyramid and the resulting heterogeneity. Most sub-Saharan countries have massive young populations and a comparatively tiny number of seniors. We do have a control for age in our regressions, and age squared, and the interaction term between HIV and age, but the functional form of these coefficients[6] does not address this age heterogeneity well enough. Figure 1 demonstrates the extent of the problem.

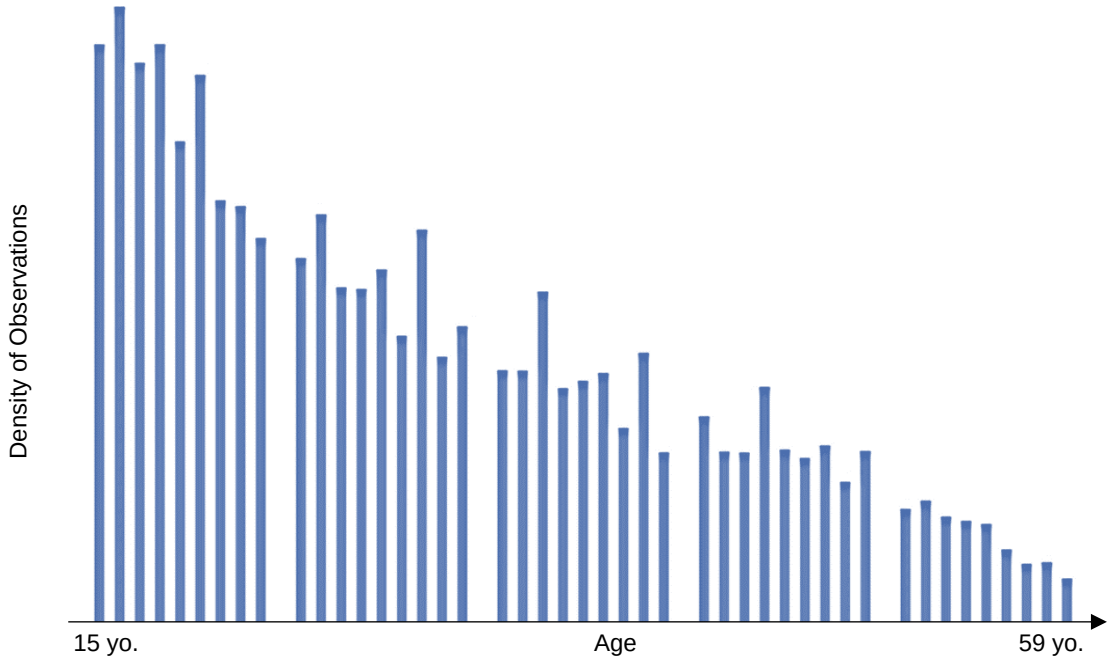
The youngest group is so massively overrepresented that even the age-related coefficients will fit the younger observations much better than the older ones, by virtue of the simple fact that there is more "loss" to be minimized with young people and the linear and quadratic shapes simply do not fit the data well enough. Any loss function, least-squares or otherwise, will be biased to fit younger age groups better. By consequence, we cannot simply state that all age-related effects are captured by controls and suspend our worries, as its linear nature prohibits it from capturing all the variation. But even if the age groups were evenly distributed, we would still have to defend the result that the effect of HIV is strictly linear over time – which we cannot. Much other HIV-related literature which proposes elaborate theoretical models, then estimates a coefficient, which in the case of a dummy such as HIV is just a constant, and then draws its conclusion. The effect of HIV is likely variable over many factors such as age and wealth, and we must treat it as such.

[5] Note that the sample *size*, at around 5000 persons, is sufficient for a good estimation.
[6] The age coefficient and interaction terms are straight lines, and the age squared a parabola.

In other words, we need another model which treats the effect of HIV as a *variable* and not a single statistic.

In summary, we have in our theoretical section stipulated that as we increase the age of the sample group, there is a growing bias toward wealthy HIV patients that survive longer. With some reservations, the existence of an age-related bias is confirmed by our simple instrumental variable models. The sizeable amount of age heterogeneity in the sample may, however, also account for this bias.

Figure 1: Distribution of age in the sample.



Contrary to expectations the graph does not decrease continuously but exhibits certain spikes. This is presumably because some persons do not know their age accurately, as it is consistently a multiple of five, such as thirty or forty, where the spike occurs.

Section V (c): A Rolling Window Model over Age

A simple solution to this problem would be to run the regressions on different ages and then glue the results together to obtain a trend of the effect of HIV over age. In our case, the sample does not allow this. The data behaves to erratic, and among older ages we almost run the risk that the variables outnumber the sample size. Moreover, some ages are overrepresented. For example, there appear to be significantly more persons aged thirty than there are aged twenty-nine, twenty-eight, or thirty-one. This is because people who do not know their age accurately are grouped in round numbers. Thus, it is hard to justify running regressions for age groups of one year. Thankfully, there are other methods. Below we run an estimation reminiscent of the rolling window techniques sometimes used in finance to model volatility – a stubbornly erratic variable which

does tend to wander around a bit. This involves picking a window for a subsample, for example all people aged fifteen to thirty, and running the regression model. We record the coefficient for the effect of HIV on wealth for the age interval 15 – 30. Subsequently, the same is done but for a window shifted forward one year, the age interval 16 – 31, and the coefficient is recorded again. Note that the samples overlap. The window size is an arbitrary choice between maintaining a large enough sample yet still having enough results to reliably uncover a trend. This continues until the entire sample has been covered, and all the coefficients can be glued together to show how they evolve over time. We will run two rolling window models, both of which are completely insensitive to the fact that the dataset contains more young people than old people. Therefore, the loss function no longer biases the effect of HIV toward young people. As a result, the effect of HIV is no longer given by a single number, but by a curve determined by age. The results are presented in table VI. A figure illustrating the construction is given in appendix B.

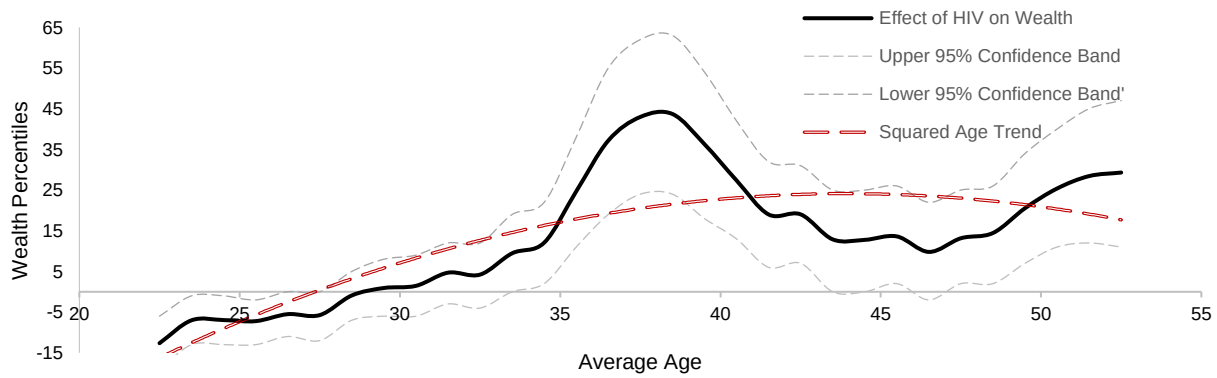
This is equivalent to "releasing" the interaction term from its linear straitjacket. Therefore, we now only use model I. The results of our estimations, 31 regressions in total, are recorded in table VI. When we ran model II for the first time, in table V, the interaction term was positively sloped and gave a positive effect around age 40. Our new estimations show a similar picture. The effect is initially negative, then around age 30 it becomes positive and goes up drastically, only to decrease and increase again over time. Essentially, the effect of HIV over age has now become badly behaved variable. There is not much intuitive explanation for it, and it might be easy to dismiss it as either bad data, which we were very careful to avoid, or as a problem with the model. But we know from the literature that models I and II are consistent. Furthermore, of all the biases we have determined to exist and discussed previously, none adequately explains the shape of this graph. Where does this big bulge for middle-aged people come from?

This phenomenon is typical of the least-squares solutions in most regressions, or that of any other loss function. What presumably happens here is the result of an inherent issue of the regressions concept, namely that they will always provide an average or insignificant coefficient if two large subsamples behave in divergent ways.

So far, our rolling window model has shown *more-or-less* the same thing as model II: an initially negative effect of HIV which goes up over age. The problem now is that model II is not realistic, unless one takes a linear effect to be realistic, and that the rolling window estimation does not give us a result that we can satisfyingly explain without resorting to peculiar intellectual gymnastics. If this behavior is due to divergent movements among subsamples, then which subsamples should we look at? When we tested our model on the wealthiest ten percent, the output of which is stored in the appendix, we noticed that the effects of HIV appeared different for this group than for the whole sample, which gives us a good place start.

TABLE VI: Rolling Window Estimations over different age groups

Sample Age Interval	Average Age	Effect of HIV on Wealth	p-value	Upper 95% Confidence Band	Lower 95% Confidence Band	# of Observations
15 – 30	22.5	-12.64	0	-19	-6	30040
16 – 31	23.5	-7.01	0.022	-13	-1	28922
17 – 32	24.5	-6.96	0.018	-13	-1	27471
18 – 33	25.5	**-7.22**	0.012	-13	-2	26037
19 – 34	26.5	-5.5	0.061	-11	0	24460
20 – 35	27.5	-5.73	0.053	-12	0	24385
21 – 36	28.5	-0.9	0.776	-7	5	24287
22 – 37	29.5	0.93	0.79	-6	8	22223
23 – 38	30.5	1.46	0.685	-6	9	21471
24 – 39	31.5	4.7	0.229	-3	12	21103
25 – 40	32.5	4.22	0.312	-4	12	20813
26 – 41	33.5	9.46	0.052	0	19	19844
27 – 42	34.5	12.07	0.018	2	22	19410
28 – 43	35.5	24.69	0	11	38	18687
29 – 44	36.5	37.17	0	19	55	17834
30 – 45	37.5	42.98	0	24	62	17638
31 – 46	38.5	43.7	0	24	63	16601
32 – 47	39.5	36.28	0	18	54	16070
33 – 48	40.5	27.37	0	13	42	15509
34 – 49	41.5	19.01	0.005	6	32	14940
35 – 50	42.5	19.03	0.002	7	31	14200
36 – 51	43.5	12.88	0.047	0	25	12599
37 – 52	44.5	12.76	0.047	0	25	12220
38 – 53	45.5	13.61	0.027	2	26	11304
39 – 54	46.5	9.8	0.107	-2	22	10595
40 – 55	47.5	13.16	0.024	2	25	10140
41 – 56	48.5	14.45	0.019	2	26	9333
42 – 57	49.5	20.54	0.002	7	34	8803
43 – 58	50.5	25.39	0.001	11	40	8122
44 – 59	51.5	28.44	0.001	12	45	7618
45 – 60	52.5	29.31	0.001	11	47	6412



These are thirty-one different regressions on different but overlapping subsamples of our dataset. By taking overlapping averages, we can create more accurate estimations than with a subsample of a single age, yet still uncover a trend over age.

Section V (d): A Rolling Window Model over Age *and* Wealth

From the previous results we concluded that controlling for age, no matter how sophisticated the model, is not enough. Our theory rests on the fact that wealth creates a bias in the survival rate of HIV patients, which in turns biases the apparent effect of HIV. If this is the case, then we may consider a model which controls for both age *and* wealth. But before doing this we must address the concerns laid out at the end of section V (a) in more detail.

We indicated that running the regression on only the top ten percentiles of the wealth index could jeopardize the estimates as there was a chance that the effect of HIV was too big to get a reliable estimate. In the model we run next, we expand the window to thirty percentiles of the wealth index to accommodate this. The odds of the effect being larger than this are probably remote: for nearly one third of people included in the data the effect *cannot* be bigger, as one cannot drop below the zeroth percentile. If we also consider that for wealthier persons the effect is likely to be small, it becomes very hard to conceptually justify a coefficient near thirty or larger. While there are still some out-of-sample issues, we believe therefore them to be minor. We further note that Fortson (2008), did a similar procedure in a study when she restricted certain areas based on HIV prevalence rates. Specifically, areas with low and high HIV infection rates were excluded at some point. In this case HIV was the dependent variable, and wealth was the regressor.

Another common concern regarding this sort of subsampling is that there may not be enough variance left in the dependent variable. Given the nature of our dependent, a wealth index with an even distribution, this is not an issue.

Finally, because we subsample based on the value of the dependent, we stress stringently that the model described hereunder is not a case of the so-called *sampling-on-the-dependent*.[7] We select a sample based on objective criteria; we do not arbitrarily pick observations to prove a desired point.

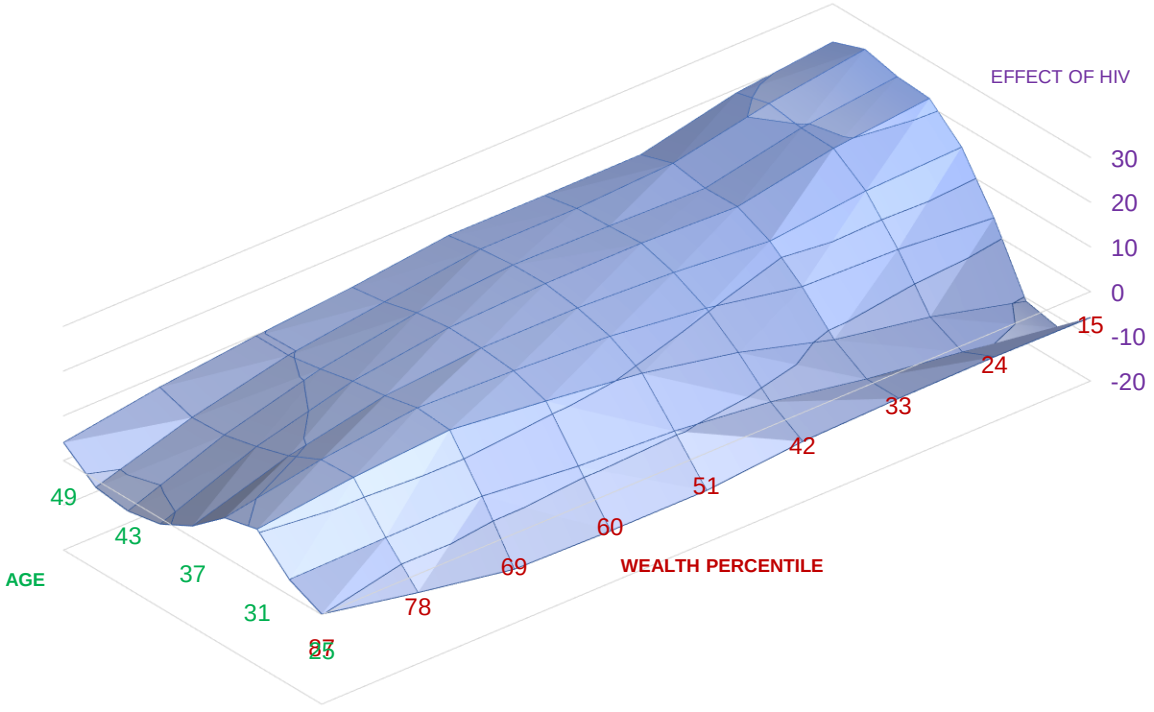
Now, then let us create a three-dimensional model with two rolling windows: one over age, as we have just done, and one over wealth. This means that each window will contain an age group, such as 15 – 30 years, and within that group a subset determined by wealth, such as the bottom 30 percentiles. Then we run model I on this subsample and record the coefficient. The results are given in table VII. Again, the window sizes are a compromise between sample size, and the number of coefficients that we need in order to obtain a trend. We were also limited by the number of regressions we could run, as each one takes a significant amount of computing power. Overall, we ran no less than eighty-one regressions over different windows to obtain our results. Each number recorded in the table is the result of our three-stage instrumental variables regression – model I.

[7] The archetype example of *sampling-on-the-dependent* is proving that swimming is deadly by looking at people who drowned and excluding those that did not.

TABLE VII: REGRESSIONS BY WEALTH AND AGE CATEGORIES

* Each purple number in the table represent the coefficient to HIV status, i.e. the effect of HIV on wealth, for a given subsample based on age and wealth. For example, the first number, -5.8, is model I run on a subsample of the poorest 15 – 35-year-olds.

Age (right) Wealth (below)	15 - 35	18 – 38	21 - 41	24 - 44	27 - 47	30 - 50	33 - 53	36 - 56	39 – 59
0 – 30	-5.8	-12.39	-10.78	3.58	15.11	21.71	22.29	24.1	21.44
9 – 39	-5.56	-8.81	-5.28	5.54	14.33	19.56	19.79	20.99	19.1
18 – 48	-5.85	-5.82	-1.34	6.23	12.11	15.51	15.34	16.62	14.25
27 – 57	-6.46	-1.92	4.84	10.44	14.11	16.37	15.9	15.53	14.14
36 – 66	-8.36	0.6	9.23	13.17	15.21	16.35	16.22	15.47	14.19
45 – 75	-8.38	1.78	12.04	14.47	15.27	15.52	14.55	12.67	11.65
54 – 84	-7.93	4.09	14.6	15.22	13.63	12.57	10.99	9.77	9.8
63 – 92	-4.12	3.84	13.02	12.66	9.21	6.99	5.26	4.89	7.36
72 - 100	0.12	3.53	10.48	8.72	2.59	-1.29	-2.69	-1.68	4.03

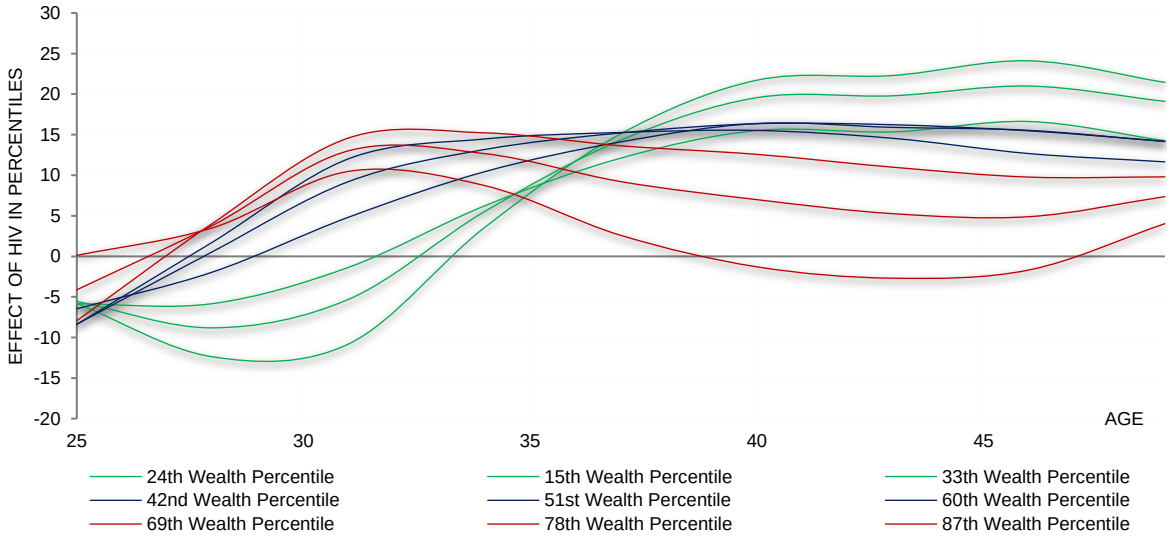


Several things become evidently clear: the graph contains higher values for the effect of HIV among the poorest percentiles, and values close to zero for the wealthiest percentiles. The effects also become more extreme for the poorest thirty percentiles: both the most negative and the most positive values are found in this subsample. This indicates that the differential survival rate is strongest there. In addition, young age groups are universally negatively affected by HIV, but for older age groups this no longer holds. Again, this points toward a differential survival rate. For the wealthier groups the coefficient of HIV becomes negative again. Generally, this model confirms what we see from the graph in table VI, except for the big bulge among thirty-something year-olds. But this is just a matter of perspective. Let us take the effect over different age groups and represent it as a set of curves. In the graph in table VIII, the red curves roughly represent the top third of the wealth index, the blue curves represent the middle third, and the green curves represent the bottom third. They are each very differently affected by HIV. This graphic a just a refined side image of the three-dimensional graph, tweaked to show how the different trends in certain subsamples can lead a general model to generate misleading coefficients. Most notably, the bulge is completely gone now. Instead, we notice that around the age of thirty-five all coefficients are very positive, but not nearly among the largest. However, there are no coefficients closer to zero to "average out" the effect, and thus a model on the entire sample of people in their thirties will throw out a large coefficient. Moreover, there is significantly less variation around that age, which likely contributed to the larger confidence intervals seen in table VI. Conversely, we can see that around the age of forty-five, the coefficient for poor persons is largest, while it is smallest for wealthier persons. Despite clear evidence that certain people in this group are strongly affected by HIV, it would only show up as a moderately sized coefficient. Note that the graph only starts at average age 25 due to the nature of the model, and thus does not in fact indicate that for the wealthiest percentiles, HIV has an immediate positive effect starting at age 15.

Also note that for wealthier percentiles the coefficients are consistently smaller than for the poorest ones. In the strictest sense, this is an endogeneity issue. Our instrument removes the effect of wealth on the odds of contracting HIV. What it does not remove is 'the effect that wealth has, on the effect of wealth of HIV.' It is easy to confuse these two, but they are in fact quite different. Indeed, a very wealthy person spending currency on treatment will not see their position change on the wealth index, while a very poor person will see a change. Thus, the size of the change is very much determined by initial wealth. This is one of the reasons for why it is so important to treat the effect of HIV as a heterogenous effect, as we can track how this effect changes over different wealth groups. By using rolling window estimations, we essentially make the heterogeneity irrelevant.

Overall, it appears that the poorer one is, the more negative the impact of HIV is among younger groups, and the more rapidly it grows positive as the group grows older – presumably, this positive effect is caused by the differential survival rate. Simultaneously, the effects are more flattened out among the wealthier percentiles, as we would expect, because most receive treatment and there is no differential survival rate.

TABLE VIII: TOTAL IMPACT OF HIV PER WEALTH GROUP



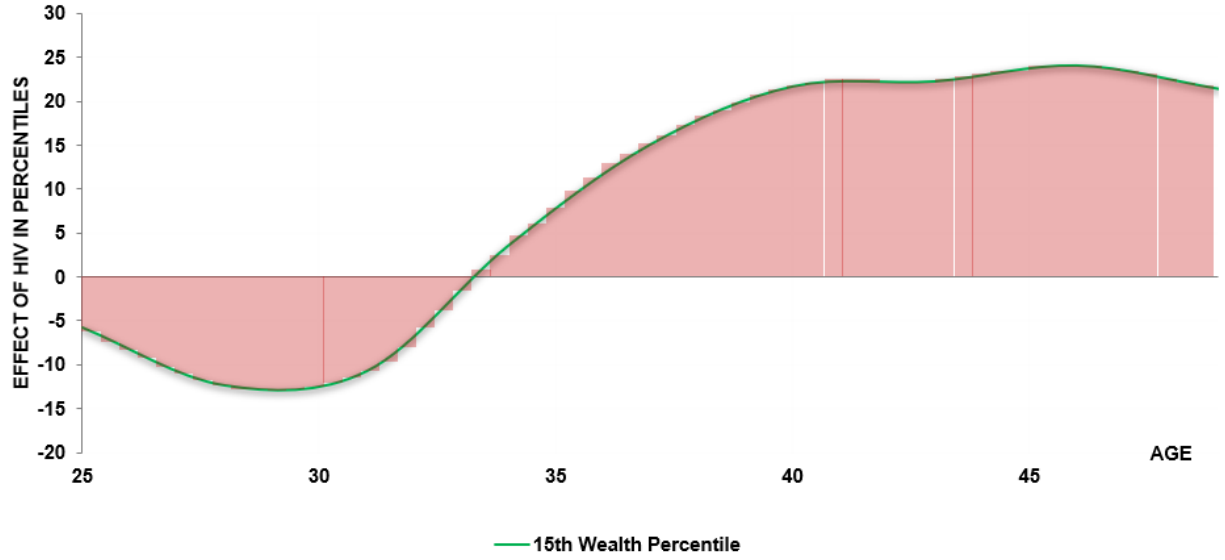
Each curve on the top graph represents a different wealth percentile of the sample. It is essentially a sideways view of our three-dimensional graphic. Every curve here represents the effect of HIV over age for a single wealth group. The nine lines thus represent the entire wealth distribution of the population. Green curves represent the poorest percentiles, while the red ones represent the wealthiest percentiles. The blue ones are average.

Wealth Average Percentile	Actual Sample Used
15	0-30
24	9-30
33	18-48
42	27-57
51	36-66
60	45-75
69	54-84
78	63-93
87	72-100

Right now, the effect of HIV is shown as a variable – a curve in our graphs. We can manipulate this to determine which groups suffer the strongest bias from differential survival. Logically, the most affected subsamples should show the largest change in coefficient over age. In other words, a subsample whose coefficient which is very negative at young ages and very positive at older ages is heavily affected by the bias, and vice versa. A good way to determine which group is most heavily affected is by taking the curves given in the table above and integrating them over the wealth effect of HIV. This calculates the surface area between the horizontal axis and the curve. Simultaneously, this gives us a good indication of which groups are most heavily affect by HIV, as larger areas indicate larger wealth effects.

By doing this, we can say with confidence which groups are more, or less, affected than others by ranking according to the gross surface area below their respective HIV curves. This surface is unitless (unless one believes an age-wealth-percentile to have interpretative value), but we believe it to be an ordinal ranking otherwise. In other words, groups with similar numbers will also be affected to a similar degree. Note that we do not speak of a positive or negative effect. After all, the coefficient of HIV may be positive over age, but it can hardly be called a positive effect as its positivity is caused by the survival bias. Our ranking merely indicates in which groups HIV has

had a large effect and in which ones it has had a small effect. We use the data from table VIII to approximate the areas in a rough fashion, as integrating and calculating a sixth-degree polynomial best-fit curve is prone to human error. Overall, this only affects the numbers marginally. An example is shown hereunder:

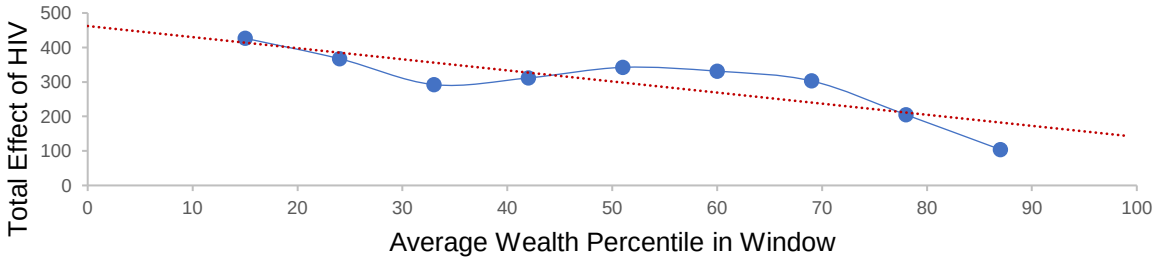


And so, several things become clear: the effect of HIV on wealth cannot easily be recorded by one or two coefficients. Instead, it acts as a variable over age and wealth, and possible over other factors. From our estimations, we cannot draw a numerical conclusion about the effects, or it would not be wise to do so. It is much better to look at the larger trends that appears once one accounts for several biases in the data, such as the skewed population pyramids common to sub-Saharan countries.

Two trends are so obvious that they almost conclusively exist:

1. **Poor subsamples are affected more by HIV than wealthy groups.** This is self-evident, as the poor suffer more from a cut in their budget than the wealthy. Moreover, they may not receive adequate care. For the wealthiest subsample, the effect of HIV is tiny.
2. **There exists an effect over age which makes a group of HIV-positive people wealthier as the average age of the group increases.** This is explained by the differential survival rate between the wealthy and the poor. Notably, in the wealthiest subsamples this effect disappears.

TABLE IX: TOTAL IMPACT OF HIV



The blue line on this graph was constructed by calculating the integrals of the curves in table VIII. To estimate the total effect of HIV for a given wealth percentile, we integrated the function that shows the effect over age for that percentile on the wealth index. We then take this result, the area under the curve, to be the "total" impact of being HIV-positive on one's wealth for a given person in a given wealth percentile. We collected the nine different results for the nine different wealth groups in the table on the right-hand side. Overall, a poor household is affected "four times as heavily" as a wealthy household. These results are unit less and cannot be interpreted in a straightforward way. However, they can be used to rank different groups in terms of how heavily they are impacted by HIV. Every dot on this graph represents the area under one curve on in table VIII.

Wealth Average Percentile	Total Impact (Ordinal)
15	426.66
24	367.455
33	292.275
42	311.7
51	342.4125
60	331.3425
69	302.9625
78	204.9
87	104.0475

Another reason to focus on the larger trends is that our confidence intervals are somewhat wide, however they are still narrow enough that the trends are statistically certain – i.e. the confidence band around the estimated curve is not so wide that the curve could realistically trend differently. This can be seen in table VI, for example. Moreover, this is not intended to be a final study, and our results at least point toward several variables that must be explored thoroughly if one wants to accurately determine the effects of HIV and similar diseases.

Section V (e): Discussion on Robustness

The exact same controls were used in all models to ensure perfect comparability across the different models, partly because the sample size can change significantly depending on the in- or exclusion of certain controls. However, this was done also out of necessity: the results of almost 115 three-stage regressions are recorded one way or another in our models, each quite intensive to compute, and customizing the set of controls for each is a laborious exercise. We do note that our test runs with different sets yielded similar results. Moreover, all our control variables were carefully chosen based on a theoretical foundation – which means that the inclusion of a few insignificant ones does not jeopardize the results. Furthermore, controls behave logically. Education and environment, for example, are always among the strongest predictors of wealth.

Although the controls are fixed, the subsamples are not. As our rolling window models show, the coefficient *trends* as we change the subsample in ways that we can explain or, at least, intuitively understand. They behave consistently, and the models appear to uncover actual underlying patterns in the data. Obviously, they don't spout out random numbers. Overall, our models perform well under a variety of subsamples and different probability distributions of the data.

SECTION VI: CONCLUSION

In general, we find that in sub-Saharan Africa the effect of HIV on wealth is stronger for poorer groups than for wealthier groups, although a numerical conclusion cannot be drawn. Our various models show a similar trend: poor people are more affected, and the effect grows positive over age – most likely due to a bias incurred by the fact that people who survive longer tend to be wealthier, which in turn leads to them being overrepresented in the sample.

We paid specific attention to the numerous biases and limitations hidden in the data. The endogeneity caused by the causal relationship between wealth and the odds of contracting HIV was removed by using circumcision as an instrument. Furthermore, the age heterogeneity was, as best as possible, neutralized by using rolling window estimations. We applied the same technique to uncover the different effects of HIV on groups of various means.

Our general findings could be tested by alternative research in more developed countries. Our case rests on the idea that poorer people are affected because they cannot afford care, while wealthier people can. In developed countries with accessible healthcare, this means that our models should show insignificant effects over both age and wealth groups.

Moreover, this is to our knowledge the first study which examines the effect of HIV on wealth on the individual level, filling the gap between macro-level studies, and micro-level studies which examine the opposite causal relationship, namely that wealth is a strong determinant for HIV infection.

Lastly, there are many diseases in the world, though not perhaps as widespread as HIV, which exhibit similar patterns in the sense that they can have a strong impact on people's lives, but that this may not show up in the data. Lupus comes to mind. Too often, the causes of those diseases are investigated, but not the effects on the patients that live with them. We hope that our thesis can provide some inspiration for future endeavors.

BIBLIOGRAPHY

Adams, R., Almeida, H., & Ferreira, D. (2009). Understanding the relationship between founder CEOs and firm performance. *Journal of Empirical Finance*, 136-150.

Angrist, J. D., & Krueger, A. B. (2001). Instrumental Variables and the Search for Identification: from Supply and Demand to Natural Experiments. *Journal of Economic Perspectives*, 69-85.

Ankrah, E. M. (1993). The impact of HIV/AIDS on the family and other significant relationships: the African clan revisited. *AIDS care*, pp. 5-22.

Arndt, C., & Lewis, J. D. (2000). The macro implications of HIV/AIDS in South Africa: a preliminary assessment. *South African Journal of Economics,*, pp. 380-392.

Ataguba, J. E., Day, C., & McIntyre, D. (2015). Explaining the role of the social determinants of health on health inequality in South Africa. *Global health action*, p. 28865.

Atkinson, A. B. (1971). The distribution of wealth and the individual life-cycle. *Oxford Economic Papers*, pp. 239-254.

AVERT. (2019). *AVERT ORG*. Retrieved from https://www.avert.org/learn-share/hiv-fact-sheets/circumcision

Bailey, R. C. (2007). Male circumcision for HIV prevention in young men in Kisumu, Kenya: a randomised controlled trial. *The lancet*, pp. 643-656.

Bonnel, R. (2000). HIV/AIDS and economic growth: a global perspective. . *South African Journal of Economics,*, pp. 820-855.

Cockerham, W. C. (2000). The social gradient in life expectancy: the contrary case of Okinawa in Japan. *Social science & medicine*, pp. 115-122.

Cohen, M. S. (2011). Prevention of HIV-1 infection with early antiretroviral therapy. *New England journal of medicine*, pp. 493-505.

Cuddington, J. T. (1993). Modeling the Macroeconomic Effects of AIDS, with an Application to Tanzania. *The World Bank Economic Review,*, pp. 173-189.

Durevall, D., & Lindskog, A. (. (2016). Adult mortality, AIDS, and fertility in rural Malawi. *The Developing Economies,*, pp. 215-242.

Fortson, J. G. (2008). The gradient in sub-Saharan Africa: socioeconomic status and HIV/AIDS. *Demography*, pp. 303-322.

Fox, A. M. (2010). The social determinants of HIV serostatus in sub-Saharan Africa: an inverse relationship between poverty and HIV? *Public Health Reports*, pp. 16-24.

Freire, S. (2002). Impact of HIV/AIDS on saving behaviour in South Africa. *Manuscript, University Paris I.*

Garmaise, D. (2012). *Aidspan*. Retrieved from Aidspan: http://www.aidspan.org/gfo_article/more-people-ever-art-cost-treatment-continues-decline

Hinson, R. E. (2011). Banking the poor: The role of mobiles. . *Journal of Financial Services Marketing*, pp. 320-333.

Iorio, D. &.-L. (2016). *Education, HIV Status and Risky Sexual Behavior: How Much Does the Stage of the HIV Epidemic Matter?* Retrieved from SSRN: https://poseidon01.ssrn.com/delivery.php?ID=573003097124109023102021016106075102004031054052030066103127101126126084090027071007006034063101028059032066017099004125071097046016071077042065003082095079122070096068080057089027001090064091002087124068123082O

Kalemli-Ozcan, S. (2012). AIDS,“reversal” of the demographic transition and economic development: evidence from Africa. . *Journal of Population Economics*, pp. 871-897.

Kalichman, S. C. (2007). Alcohol use and sexual risks for HIV/AIDS in sub-Saharan Africa: systematic review of empirical findings. *Prevention science*, p. 141.

Lange, J. M., & Ananworanich, J. (2014). Review: the discovery and development of antiretroviral agents. *Antiviral Therapy*, 5 - 14.

Long, D., & Deane, K. (2015). Wealthy and Healthy? New evidence on the relationship between wealth and HIV vulnerability in Tanzania. *Review of African Political Economy*, 376 - 393.

Michalopoulos, S., & Papaioannou, E. (2013). Pre-colonial ethnic institutions and contemporary African development. *Econometrica*, pp. 113-152.

Mishra, V. B. (2007). A study of the association of HIV infection with wealth in sub-Saharan Africa.

Moses, S. B.-A. (1990). Geographical patterns of male circumcision practices in Africa: association with HIV seroprevalence. *International Journal of Epidemiology,*, pp. 693-697.

Murdock, G. (1959). *Africa: itspeoplesandtheirculturehistory.* McGraw-Hill New York.

Mveyange, A. S. (2015). Does HIV/AIDS matter for economic growth in sub-Saharan Africa? . *WIDER Working Paper.*

Stein, M. H. (2005). Alcohol use and sexual risk behavior among human immunodeficiency virus–positive persons. . *Alcoholism: Clinical and Experimental Research,*, pp. 837-843.

Szabo, R., & Short, R. V. (2000). How does male circumcision protect against HIV infection? *Bmj,*, pp. 1592-1594.

Türmen, T. (2003). Gender and HIV/aids. *International Journal of Gynecology & Obstetrics*, pp. 411-418.

Ukwuani, F. A. (2003). *Condom use for preventing HIV infection/AIDS in sub-Saharan Africa: a comparative multilevel analysis of Uganda and Tanzania. .* Journal of acquired immune deficiency syndromes (1999).

UNDP. (2017). *United Nations Development Program*. Retrieved from https://www.undp.org/content/undp/en/home/presscenter/pressreleases/2015/11/27/breakthrough-brings-cost-of-hiv-treatment-to-under-100-per-patient-per-year.html

Werker, E. A. (2006). Male circumcision and AIDS: the macroeconomic impact of a health crisis. *Division of Research, Harvard Business School.*

World Bank. (2019, March 13). *The World Bank Data*. Retrieved from https://data.worldbank.org/indicator/SP.DYN.LE00.MA.IN?locations=ZF

World Health Organization. (n.d.). *Progress on global access to HIV antiretroviral therapy.* Retrieved 05 07, 2019, from World Health Organization: https://www.who.int/hiv/fullreport_en_highres.pdf

Young, A. (2005). The gift of the dying: The tragedy of AIDS and the welfare of future African generations. . *The Quarterly Journal of Economics,* , pp. 423-466.

APPENDIX A

MODEL I OUTPUT, top ten wealth percentiles

Independent Variable: Household Wealth Index, 100 percentiles. **Sample size:** 5139 males of the top ten wealth percentiles **Age:** +15yo. **Centered R^2** = 0.1503.

Variable	Coefficient	z-stat	p-value	Standard deviation	lower 95%	upper 95%	Note
HIV	1.47	1.30	0.193	1.13	-0.74	3.67	Dummy
+ controls[(1)]	.	.	.	.	.	.	.

(1) Controls omitted for brevity. The same controls as in tables IV and V were used.

MODEL II OUTPUT, top ten wealth percentiles

Independent Variable: Household Wealth Index, 100 percentiles. **Sample size:** 5139 males of the top ten wealth percentiles **Age:** +15yo. **Centered R^2** = 0.1471.

Variable	Coefficient	z-stat	p-value	Standard deviation	lower 95%	upper 95%	Note
HIV	-3.57	-0.85	0.397	4.22	-11.83	4.69	Dummy
HIV*age	0.12	1.22	0.222	0.10	-0.08	0.33	Continuous
+ controls[(1)]	.	.	.	.	.	.	.

(1) Controls omitted for brevity. The same controls as in tables IV and V were used.

APPENDIX B