Deuchert, Eva

Conference Paper

The Virgin HIV Puzzle: Can misreporting account for the high proportion of HIV cases in self-reported virgins?

Proceedings of the German Development Economics Conference, Zürich 2008, No. 10

Provided in cooperation with:
Verein für Socialpolitik

THE "VIRGIN" HIV PUZZLE:
Can misreporting account for the high proportion of HIV cases in self-reported virgins?

Eva Deuchert
St.Galler Research Centre for Ageing, Welfare and Labour Market Analysis
University of St Gallen, Switzerland

(Work in progress)

Summary

The Demographic and Health Surveys from Lesotho, Zimbabwe, and Malawi reveal that a significant proportion of HIV infections in adolescent women occurred in women who claim to be virgin. Two possible conclusions arise from this observation: adolescent women misreport sexual status or non-sexual risk is more relevant than previously asserted. This paper uses a nonparametric model to estimate the proportion of HIV infections associated with sexual activity under different assumptions on data accuracy. It shows that there is an inverse relation between data accuracy and importance of sexual HIV transmission. If all adolescent women in the considered sub-sample correctly report sexual activity, 70% of HIV infections cannot be attributed to sexual HIV transmission. The model predicts that more than 95% of HIV infections are due to sexual HIV infections, if a substantial proportion of self-reported virgins (between 40 and 90%) misreport sexual status.

JEL code: C14, I10, I12

Keywords: adolescent, HIV, misreporting, nonparametric modelling, sexual transmission

1 Introduction

It is widely believed that sexual HIV transmission is the dominant transmission mode in Sub Saharan Africa accounting for more than 95% of all HIV infections (see for example Schmid et al. 2004). This belief seems inconsistent with the data from the Demographic and Health Surveys (Table 1): In three of the most affected Sub Saharan countries, a substantial proportion of HIV positive adolescent women claim to be virgins. Two alternative explanations for this observation were proposed: (1) Adolescent women misreport sexual status (social desirability bias, Gavin et al. 2006). (2) A significant number of adolescents may have acquired HIV by non-sexual HIV transmission (Brewer et al. 2007).

Table 1: HIV Prevalence among adolescent women (age 15-19)

	HIV Prevalence		% of HIV infections who claim to be virgins
	Virgin	Sex. active	
Lesotho	5.3%	11.2%	36%
Malawi	1.8%	5.1%	20%
Zimbabwe	3.3%	10.4%	41%

Source: Demographic and Health Surveys (DHS), survey weights are not used

Since the Demographic and Health Surveys (DHS) do not include medical examinations, misreporting remains unobserved. It is therefore a priori unclear whether HIV infections in self-reported virgins are due to misreporting or due to nonsexual HIV transmission. This issue however, is of primary importance for public health research and policy: a high proportion of women misreporting sexual behaviour raises the question whether self-reported sexual data should be used to monitor the HIV/AIDS epidemic; a high proportion of non-sexual HIV transmission on the other hand may bring up the issue whether the prevention of non-sexual HIV risks is sufficiently covered by the current prevention paradigm. The present paper analyzes the trade-off between data accuracy and the importance of sexual HIV transmission.

The related economic literature on HIV/AIDS control is fairly new, since HIV/AIDS had not attracted a large amount of economic research during the 1990s. With the availability of large nationally representative surveys (such as the Demographic and Health surveys used in this paper), the control of the HIV/AIDS epidemic became subject to many studies in health and development economics. This research applies the economic toolbox to the analysis of risk-taking sexual behaviour and HIV infections (see for example de Walque 2007, Cogneau and Grimm 2007, Luke 2006, Oster 2007).

This new data however, has also revealed several anomalies that are unexplained by our current knowledge on sexual behaviour and HIV transmission: (1) The large regional variation in HIV prevalence in Sub Saharan Africa cannot be accounted to differences in self-reported sexual behaviours (Buvé et al. 2001a). (2) Studying HIV infection at the level of the cohabiting couples, de Walque (2007b) shows that in a high proportion of sero-discordant couples, the female partner is infected. This is at odds because most infected women did not report extramarital sexual relationships. (3) It was widely believed that the HIV/AIDS epidemic is associated to poverty, because it limits access to condoms, forces women into prostitution, and lowers individual incentives to adopt safer behaviours (see for example in Philipson and Posner 1995, Oster 2007). Several studies however, report a positive association between HIV infection and wealth (see for example Lachaud 2007, Mishra et al. 2007). (4) A high HIV prevalence in adolescent women who report that they have not initiated sexual activity is found by Gavin et al. (2006) and Buvé et al. (2001b).

Several researchers have argued, that these anomalies are explained by misreporting (see for instance in de Walque 2007b, Gavin 2006). However, no study has systematically analyzed whether misreporting can account for these anomalies and which assumptions are needed if one wants to explain these findings with misreporting.

The paper contributes to this discussion by modelling the importance of sexual HIV transmission under different assumptions regarding data accuracy. It uses a non-parametric model based on the propensity score matching method that does not rely on assumptions regarding the sexual risk accumulation process. The finding suggests that an inverse relation between data accuracy and importance of sexual HIV transmission exists: If all adolescent women in the considered sub-sample correctly report sexual activity, 70% of HIV infections in the sub-sample cannot be attributed to sexual HIV transmission. The hypotheses that more than 95% of HIV infections are due to sexual transmission seems only valid, if we are willing to assume that between 40 and 90% of self-reported virgins misreport sexual activity.

This paper is organized as following: The next section outlines the model that is used to estimate the proportion of sexual HIV transmission. Section 3 introduced the data used in this study and discusses issues related to econometric methodology and the empirical implementation. Section 4 presents the results and the final section concludes.

2 A simple model of the importance of sexual HIV transmission

Sexual HIV transmission is typically modelled in the context of a Bernoulli-risk accumulation model (see for instance in Oster 2005), which each sexual episode or sexual partnership is treated as an independent trial and the probability of a "success" (HIV infection) on any one trial is assumed to equal the per-contact (or per-partner) probability of HIV transmission. This approach requires having knowledge about the number of partnerships, the number of unprotected sexual episodes per partner, the partner matching process, the HIV prevalence in partners, and the sexual HIV transmission efficiency. These variables are often not available and need to be based on assumptions, are measured only with a great uncertainty, or are also constraint by misreporting Estimates from these models are therefore attached with a great uncertainty (a discussion on this issue can be found in Deuchert and Brody 2007).

This paper uses therefore a non-parametric model that does not rely on parametrically modelling the sexual-risk accumulation process. This model compares the average proportion of HIV infections that can be traced to sexual activity to the total (observed) HIV prevalence in the sub-group.

Two important issues need to be considered: (1) Sexually active women may be also exposed to non-sexual risks (they may use injecting drugs or may be exposed to potentially unsafe health care; some women may have been HIV positive at their sexual debut). (2) DHS data relies on self-reported behaviour and some women may misreport sexual status.

Assuming that (correctly classified) sexually active respondents were also exposed to non-sexual risk factors, the proportion of HIV infections due to sexual transmission (θ) is equal to:

$$\theta = \frac{\Pr(A)\left[E(HIV \mid A=1) - E(HIV \mid A=0)\right]}{\mathrm{E(HIV)}} \tag{1}$$

E(HIV|A=1) is the expected HIV prevalence in sexually active respondents, and *E(HIV|A=0)* is the counterfactual which is equal to the expected HIV prevalence in the same group given they would not have initiated sexual activity. The difference between these two terms [*E(HIV|A=1) - E(HIV|A=0)*] is equal to the expected increase in HIV prevalence associated with sexual activity. The two remaining terms denote the probability to be sexually active [*Pr(A)*] and expected HIV prevalence in the sub-sample [*E(HIV)*].

While the expected HIV prevalence can be easily estimated by the sub-sample average, the remaining terms of equation (1) cannot be estimated as misreporting of sexual activity is not observable. Estimating the proportion of sexual HIV transmissions is only possible if we are willing to make assumptions about the proportion of women who misreport sexual activity. Assuming that between 0 and 100% of adolescent women who claim to be virgins lie about their sexual status, the probability to be sexually active can then be estimated by the portion of women who report to have initiated sexual activity plus the proportion of women who are assumed to lie about their sexual status.[1]

Expected HIV prevalence in the group of women who misreport sexual status is a priori not known and needs to be based on assumptions. This assumption has impact on the estimate for the proportion of sexual HIV transmissions: A high assumed HIV prevalence in women who misreport sexual activity increases the expected HIV prevalence of sexually active women; a low assumed HIV prevalence in women who misreport sexual status decreases the expected HIV prevalence of sexually active women. The assumption furthermore impacts the estimate for the counterfactual since this estimate will be based on data for ("real") virgins (see section 3.2). When the assumed prevalence in women who misreport sexual activity is high, HIV-positive women will be quickly select out the group of "real" virgins, reducing HIV prevalence in virgins and thus, the estimate for the counterfactual.

In this paper, I will assume that HIV prevalence in women who will be reclassified as sexually active [$E(HIV|A^{**})$] lies between HIV prevalence in ("real") virgins [*E(HIV|V)*] and HIV prevalence in women who report sexual activity [$E(HIV|A^{*})$]:

[1] This implicitly assumes that potential bias goes only into one direction (women who report sexual activity have correctly reported sexual status and misreporting occurs only in self-reported virgins).

$$E(HIV \mid V) < E(HIV \mid A^{**}) < E(HIV \mid A^{*}) \qquad (2)$$

The first part of inequality (2) is justified because women who had sexual intercourse but misreport sexual status have been exposed to sexual risk. Therefore, HIV prevalence should be higher than in the group of virgins.[2] The second part seems also feasible since it is unlikely that women with sexual risks above average (such as women with sugar daddy relationships or prostitutes for example) can conceal that they are sexually active (even though they may misreport risk behaviours). Furthermore, it seems unlikely that many women who misreport their sexual status know that they are HIV positive and try to hide the infection: Only few women who claim to be virgins are aware of their serostatus, and HIV prevalence in virgins who know their status is not significantly higher than women who do not know HIV status.[3] Thus, unless we are willing to assume that women systematically misreport HIV tests and sexual status,[4] people do not seem to misreport sexual status because they know that they are HIV positive. For these reasons, it seems unrealistic to assume that women who misreport their sexual status have extraordinary high HIV prevalence rates.

3 Empirical Strategy

3.1 Data

The data is from the Demographic and Health Surveys (DHS) in Lesotho, Zimbabwe, and Malawi. DHS conduct interviews on maternal and child health, family planning, nutrition and related issues, and also collect biological and clinical data such as testing for HIV (a description of DHS surveys can be found on ORC Macro's webpage http://www.measuredhs.com/ or in Mishra et al. 2006). The countries are chosen because they have sizeable HIV epidemics with HIV prevalence rates (measured as average HIV prevalence in adults age 15 to 49) exceeding 10% and because the Demographic and Health Surveys reveal a substantial HIV prevalence in adolescent women (age 16-19) who claim to be virgins. The data is pooled across countries to maximise the size of the cross-sectional sample. The analysis is restricted to women, age 16-19 for which HIV test, sexual activity status and all covariates of interests are available (N=3085). 39% of the respondents in the sub-sample report to be sexually active. HIV prevalence is 9.3% in the group of sexually actives and 3.6% in the group of self-reported virgins. The descriptive statistics for this sub-sample are provided in the appendix (Table A1).[5]

3.2 Econometric Methodology

Three estimates are necessary to estimate the proportion of HIV cases that are due to sexual HIV transmission (see equation 1): (i) expected HIV prevalence in adolescent women, (ii) the proportion of adolescent women who (correctly) report sexual activity, and (iii) the increase in HIV prevalence associated with sexual activity.

---

[2] This implicitly assumes that no HIV/AIDS related deaths occur in individuals who have been infected with HIV by sexual transmission. This assumption seems feasible since the latent period of HIV is approximately 10 years (Pantaleo et al. 1993, Morgan and Whitworth 2001), so that respondents should be still alive even they acquired HIV early in their sexual career.

[3] In Lesotho, Zimbabwe, and Malawi only 12% of adolescent women have been tested for HIV, 35% of these women report that they are virgins. HIV prevalence is slightly higher in self-reported virgins if they have been tested (3.5% vs. 4.4%) but the difference does not appear to be statistically significant (Pearson chi2(1)= 0.23).

[4] Women have been asked whether they have ever been tested for HIV. It was pointed out during the interview that they do not have to disclose their HIV status.

[5] Sample weights are not used because pooled non-random sub-samples from three different countries are used.

Expected HIV prevalence (i) is estimated by the sub-sample average of positive HIV test results. The proportion of adolescent women who (correctly) report sexual activity (ii) is estimated by the sub-sample average of women who are classified as sexually active, once a randomly selected group of self-reported virgins, with average HIV prevalence defined by inequality (2), is reclassified as sexually active.

The increase in HIV risk associated with sexual activity (iii) is estimated using the propensity score matching method (Rosenbaum and Rubin 1983). The general problem, which this estimator tries to solve, is the counterfactual problem: While it is possible to observe the HIV status of women who report to be sexually active, it is not possible to observe the HIV status of the same woman in case she would not have initiated sexual activity.

Propensity score matching relies on the potential outcome framework, where the treatment effect equals the difference in the potential outcome for the same individual or group. Translating this principle into the problem of estimating the risk increase imposed by sexual activity, the propensity score matching estimator seeks to estimate the average treatment effect on the treated (*ATT*) by estimating the difference between HIV status expected for individuals who have initiated sexual involvement [$E(HIV^{A=1}|A=1)$] and expected HIV status for the same group of women in case they would not have been sexually active [$E(HIV^{A=1}|A=0)$]:

$$ATT = E\left(HIV^{A=1}\,|A=1\right) - E\left(HIV^{A=1}\,|\,A=0\right) \tag{3}$$

If the decision to start sexual involvement would be a random event, one could estimate the ATT by differencing HIV prevalence rates of sexual active women and virgins. However, people decide to initiate sexual activity and the same variables that influence the decision on sexual debut may also be related to potential outcome (HIV prevalence). Thus, the naïve estimator would be biased. If however, all confounding variables that simultaneously affect the decision on sexual debut and non-sexual risk can be observed, potential outcomes are independent by the "treatment" selection. This conditional independent assumption (CIA) allows estimating the treatment effect defined in equation (3) since HIV prevalence rates of virgins is identical in expectation to the unobserved counterfactual outcome of sexually active women with the same characteristics *X*:

$$E\left(HIV^{A=0}\,|X=x, A=0\right) = E\left(HIV^{A=1}\,|X=x, A=0\right) \tag{4}$$

Matching means now that a sub-sample of sexually active persons and virgins is selected, whose covariates are similar (enough) so that equation (4) holds. HIV prevalence is then compared between groups that differ only by their activity status but are otherwise equal (ceteris paribus assumption).

The advantage of propensity score matching is, that one does not have to match directly on observable characteristics (which would become more and more difficult the larger the set of characteristics), but to match on the propensity score i.e. on the probability of sexual activity (Rosenbaum and Rubin 1983). Thus, instead of conditioning on an n-dimensional variable, women are matched on the basis of a single scalar variable.

The paper uses thernel matching procedure. Kernel matching takes a kernel-weighted average of the comparison group observations near each treated observation to construct the counterfactual. The weight for each observation of the comparison group is in proportion to the distance between the propensity scores of the control and the treated.

3.3 Identification strategy

The conditional independence assumption is not directly testable. To make CIA creditable, it is therefore essential to understand the way how HIV is transmitted and the decision process of initiating sexual activity. This is necessary to identify all variables which jointly influence outcome (HIV prevalence) and treatment (sexual activity).

HIV can be transmitted through direct exposure to contaminated body fluids, such as blood, semen, cervical and vaginal secretions, and breast milk. The transmission routes are (1) sexual transmission (unprotected vaginal, oral, or anal intercourse), (2) blood-to-blood transmission (blood transfusions, exposure to HIV-contaminated needles, syringes, and other sharp objects), and (3) mother-to child transmission (during pregnancy, labour, delivery, and breast feeding).

The highly active antiretroviral therapy (HAART) was developed in 1997 so that effective HIV treatment had not been available to women belonging to this sub-sample at infant age. Because life expectancy for untreated HIV-positive infants is only 3.75 years (Marston et al. 2005), long term survivors of mother-to-child transmission should be rare in this age-cohort. Therefore, it is relatively unlikely that HIV cases in this sub-sample are caused by mother-to-child transmission, and infections in ("real") virgins must be through exposure to contaminated blood.

It is of interest to estimate the HIV risk *increase* imposed by sexual activity. Therefore, it is necessary to adjust for all confounders that potentially influence blood-to-blood HIV transmission and the decision to initiate sexual activity. Blood-to-blood transmission occurs by direct contact with contaminated blood mainly through the usage of sharp objects. The two most common blood-to-blood transmission modes are needle sharing (i.e. injecting drug users) and unsafe medical procedures (i.e. contaminated blood transfusions, re-use of contaminated injection and surgery equipment).[6]

Relatively little is known on injecting drug use in Sub Saharan Africa but it is suggested that injecting drug use in an emerging HIV risk as a result of the expanding international drug traffic (Beckerleg et al. 2005; Kloos and Mariam 2007). Needle sharing is very common, and high risk behaviours such as drawing blood back in a syringe and passing it to other drug users (which is used as a method to avoid pains from withdrawal) is frequently observed (McCrudy 2005; McCurdy et al. 2007). Risk factors that lead to injecting drug use and needle sharing have not been studied in the context of Sub Saharan Africa but European data suggest that among psychological factors and family background, low socio-economic status (Reinherz et al. 2000; Poulton et a. 2002) and low education (Johnson et al. 1995, Stronski et al. 2000) are associated with illicit substance abuse.

Health care safety is a major concern (WHO 2002a). Several studies document the lack of basic supplies and infrastructure (see for example the Health Service Provision Assessments: GSS 2003; MoHR 2003), the re-use of injection equipment (Hutin et al. 2003), or the usage of untested blood transfusions (WHO 2002b) in Sub Saharan Africa. Thus, access to potentially unsafe health care may increase the risk of HIV infections for virgin and sexually active adolescents.

The demand for (potentially unsafe) health care is typically influenced by the attributes of the disease or health state (such as the severity of the illness), the costs of health care, financial and geographic accessibility, barriers that arise from low bargaining power, and individual socio-economic status (see for example in Hjortsberg 2003; Sahn et al. 2003).

[6] Unsafe medical procedures are effective in transmitting HIV/AIDS as the outbreak of HIV in children attending the Al-Fateh Hospital in Benghazi, Libya has demonstrated (see for example in de Oliveira et al. 2006).

Sexual debut is early in Sub Saharan Africa, with median age at first sex less than 20 in almost all Sub Saharan African countries (Zaba et al. 2004). The early sexual debut is often explained by socio-cultural differences (see for example in Jordan 2004), economic constraints, which may force women to initiate sexual partnerships (for example "sugar daddy relationships", Luke 2006), and the low bargaining-power of women, which leaves them little choice over their own sexual behaviour (Clark et al. 2006).

Based on the analysis above, factors that jointly influence non-sexual risks and sexual debut can be categorized into the following groups: (1) socio-economic variables, (2) socio-cultural variables, (3) socio-geographic variables, and (4) bargaining power.

The Demographic and Health Surveys provide a large set of covariates that can be used to approximate and describe these categories such as age, education, household wealth, occupation, religion, regional indicators, and a set of self-reported barriers to health care, which approximate geographic accessibility, individual bargaining power, as well as moral values. Descriptive statistics for these variables can be found in the Appendix (Table A1).

Unfortunately, DHS does not provide information on current income and ethnic groups (for Zimbabwe and Lesotho). Income may increase bargaining power to negotiate sexual activity and may also be an important factor that impacts exposure to non-sexual risks. However, wealth, education, and occupation may approximate income. Also cultural differences may impact sexual debut and exposure to non-sexual risks (for example female genital mutilation). This analysis adjusts for regional differences, which may be closely related to differences in ethnic composition. Thus, even some doubts on the validity on CIA could be raised, it seems reasonable to assume that these missing factors (conditional on all the other variables) play only a minor role.

3.4 Estimation of the propensity score

Table 2 present the results of the Probit estimation for the probability that the respondent reported sexual activity (propensity score if no misreporting is assumed).[7] A higher age is associated with a higher likelihood to report sexual activity. Women living in richer households are less likely to report sexual activity. However, sexual activity does not seem to be related to poverty since there is no statistically significant difference between the other wealth quintiles. Women who are working seem to be more likely to report sexual activity but statistical significant effects can be only found for women working in sales or agricultural jobs. Surprisingly, variables that could be associated with barraging power (such as education or the need to get permission for getting health care) are not significantly associated with sexual activity. Roman-atholic Christians and Muslims are less likely to report sexual intercourse then Christians belonging to other churches, Traditionalists or women who are not religious, probably because these religions have stricter sexual morals. Finally, women who fear that there is no female health provider are significantly less likely to report sexual activity.

Matching requires that the propensity scores of the treatment group (sexually active) and control group (virgins) overlaps. If propensity scores do not overlap (for example because included control variables are perfect predictors), matching cannot be performed and observations will be excluded from the analysis. Consequently, the estimated effect does not represent the average effect for the full sample but have to be redefined at the mean effect for individuals falling in the common support (region where propensity scores overlap). With a small common support, resulting estimators may be of little relevance because they apply only for small groups and are not representative for the full sample. With the given data however, the

[7] The estimates for the propensity scores for the models with misreporting are not discusses but can be obtained from the author on request.

common support is not a problem because the range of propensity score values is very similar for virgins and sexually active women (see Appendix, Figure A1).

Table 2: Results of the estimation for the propensity score

	Coef.		z
Age	0.48	***	23.43
Self reported barriers to access health care (big problem vs. small problem)			
Getting permission to go	-0.01		-0.11
Getting money needed	0.10		1.64
Distance to health facility	-0.03		-0.35
Having to take transport	0.03		0.35
Not wanting to go alone	0.06		0.95
No female health provider	-0.21	***	-2.60
Highest educational level:			
Primary	0.12		0.49
Secondary, higher	-0.32		-1.31
Wealth			
Poorer	0.01		0.12
Middle	-0.09		-1.04
Richer	-0.04		-0.39
Richest	-0.44	***	-3.89
Occupation			
Sales	0.26	**	2.02
Agricultural	0.38	***	4.40
Household & domestic	0.05		0.42
Services	0.06		0.24
Skilled & unskilled manual	0.06		0.35
Clerical	0.35		1.28
Residence: Rural	-0.06		-0.60
Religion			
Other Christian	0.12	*	1.70
Muslim	0.05		0.29
No Religion, Traditional	0.42	***	3.13
_cons	-8.38	***	-16.66
Regional variables			
R2			0.24
N			3085

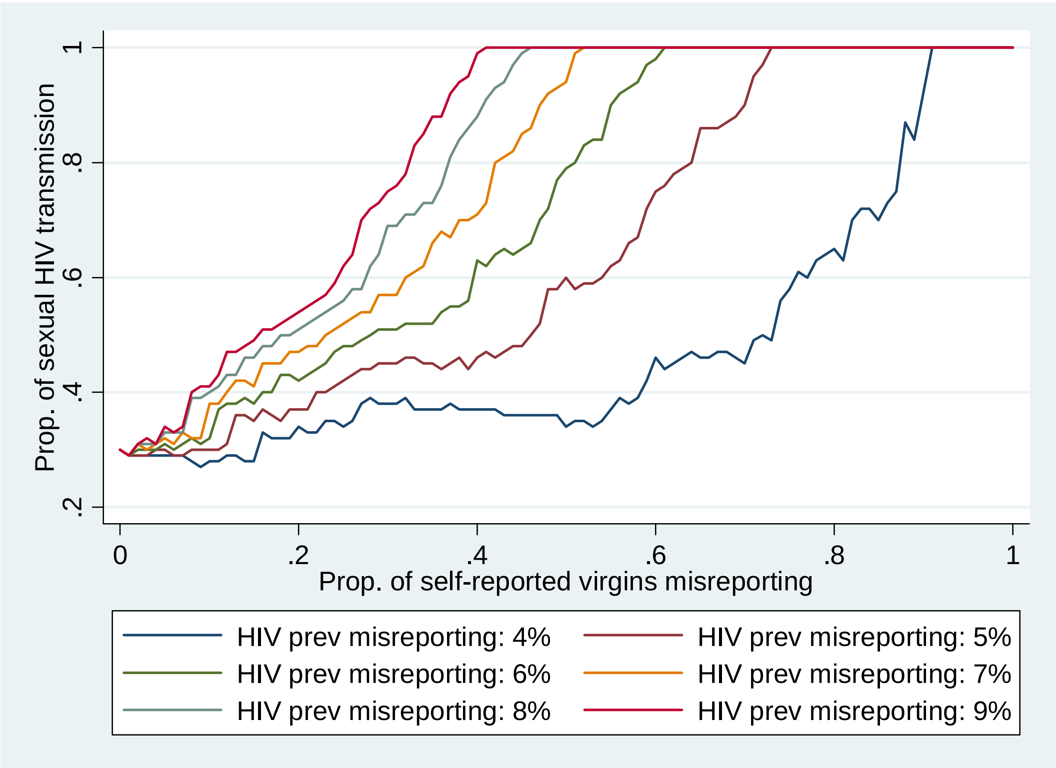
Note: Depended variable = 1 for women who report sexual activity (no misreporting assumed); significance levels: *** p<0.01, ** p<0.05, * p<0.1; omitted categories: Highest education level: none, Wealth: Poorest, Occupation: None, Residence: Urban, Religion: Roman-Catholic; coefficients for regional variables are not reported but available from the author on request

4 Results

This section provides the estimates for the proportion of HIV infections caused by sexual transmission under various scenarios. These scenarios differ by the assumptions on the proportion of virgins who misreport, and by the assumed HIV prevalence in women who misreport.

Assuming that adolescent women did not misreport sexual status, the average risk increase associated with sexual activity equals 0.04 [95% bootstrap CI: 0.02-0.07].[8] The proportion of women who classify themselves as sexually active is equal to 0.39. Given the HIV prevalence in the considered sample of 5.8%, the model predicts that 29.68% [95% bootstrap CI: 13.59 - 47.01] of HIV infections can be traced to sexual HIV transmission.[9] Thus, if all adolescents have correctly reported their sexual status, 70% of HIV infections cannot be explained by sexual HIV transmission and must be due to other causes.

Figure 1: Estimated proportion of sexual HIV transmission under different scenarios



Note: The Epanechikov kernel with a bandwidth of 0.06 is used to construct weights for estimating the HIV risk increase associated with sexual activity.

The analysis above assumes that all women correctly report their sexual status. If this is not the case and a significant proportion of virgins misreport sexual status, the proportion of sexual HIV transmissions varies (see Figure 1). With increasing proportion of virgins who are assumed to misreport sexual status, the proportion of sexual HIV transmission increases. The speed of this process however, depends on the assumed HIV prevalence in women who misreport sexual status: When the assumed HIV prevalence in women who misreport sexual status is low (4%), the proportion of HIV infections caused by sexual transmission remains relatively constant even if 40 to 50% of self-reported virgins are reclassified as sexually active. The slope of the curve becomes steeper only if a significant share of women (> 50%) misreports sexual activity. The reason for this is that HIV prevalence in ("real") virgins (used

[8] The bootstrap estimator is based on 100 repetitions.

[9] The naïve estimator, which sets the counterfactual equal to HIV prevalence in virgins, would predict that 38.3% of HIV infections are due to sexual HIV transmission.

to estimate the counterfactual) declines only slightly once women who are assumed to misreport are sampled out, while HIV prevalence in sexually active women declines sharply. As a result, the HIV risk increase associated with sexual activity declines sharply, which (partly) offsets the increase in the proportion of sexually active individuals. If the assumed HIV prevalence in women who misreport sexual status is close to HIV prevalence in women who report sexual activity (9%), HIV prevalence in the group of sexually active remains constant but HIV prevalence in virgins declines sharply. Because this decreases the estimate for the counterfactual, HIV risk increase associated with sexual activity increases, causing a sharp raise in the estimate for the importance of sexual transmission.

Figure 1 demonstrates nevertheless that a substantial proportion of self-reported virgins need to be reclassified as sexually active, to assume that HIV is almost exclusively sexually transmitted. In the scenario, where HIV prevalence in women who misreport sexual activity is 4%, more than 90% of self-reported virgins need to be re-classified as sexually active to obtain an estimate for sexual HIV transmission of 95% and higher. If HIV prevalence rate in women who misreport sexual status is close to HIV prevalence in women who report sexual activity (9%), still more than 40% of self-reported virgins need to be reclassified as sexually active.

It is a priori unclear, to which degree adolescent women misreport sexual behaviour. A first insight can give the Multicentre Study of Factors Determining the Different Prevalences of HIV in Sub-Saharan Africa, which estimated the magnitude of misreporting by testing for various sexually transmitted infections among women aged 15-24 years who reported that they had never had sexual intercourse (Buvé et al. 2001b). The study concludes that between 6 and 18% of women had misreported their sexual activity.[10] Thus, assuming that 20% of virgins misreport sexual status, the model predicts that between 34 and 54% of HIV cases are due sexual HIV transmission, between 46 and 66% of HIV infections remains unexplained by sexual HIV transmission. However, it is clear that misreporting rates drawn from studies on other regional settings may not be representative for the considered sub-samples. Therefore the estimates can give only a first insight but should not be treated as an estimate for the importance of sexual HIV transmission.

The sensitivity of the results is checked using alternative specifications to estimate the propensity score: (i) It is estimated whether the inability to control for ethnic groups results in a significant bias. This sensitivity check includes dummy variables for ethnic groups in Malawi, while for Lesotho and Zimbabwe only one dummy variable is created (one for Lesotho and one for Zimbabwe). (ii) To check the sensitivity for omitted socio-economic variables, I omit wealth, education, and occcupation. The results for these specifications are not presented but are very similar to the ones presented in Figure 1. Therefore, the overall conclusion, that a substantial proportion of self-reported virgins need to misreport sexual status when the hypothesis that HIV is almost exclusively sexually transmitted is accurate, remains largely unchanged by omitting and adding variables.

Nevertheless, important information are lacking and the analysis above relies on assumptions, which validity needs to be investigated by future research: Since no information on women who potentially misreport sexual activity is available, I had to randomly select women who claim to be virgin to reclassify them as sexually active. This proceeding may be oversimplifying and the analysis could have been greatly improved if some indicators for misreporting (for example inconsistent answers, questions on symptoms of sexually transmitted diseases, etc) were available. Furthermore, the analysis implicitly assumes that the risk increase in HIV

[10] Gavin et al. (2007) report also that approximately 8% of the HIV-positive females who reported no sexual history did report a history of STI symptoms. Unfortunately, Gavin et al. do not report how many adolescents who reported sexual history also report a history of STI symptoms; therefore it is not possible to estimate a misreporting rate for Zimbabwe.

prevalence associated with sexual activity is due to sexual HIV transmission. This does not need to be the case because sexually active women are more likely to be exposed to non-sexual risk factors (hormonal injections are a common contraception mode; pregnant women receive antenatal health care; sexually active women receive treatment for sexually transmitted infections).[11] It is not possible to adjust for variables that approximate exposure to health care, because available variables on access to health care (injections in the previous months, visits of health facilities) are constraint by reverse causality: HIV positive women are more likely to be in contact with the formal health care sector because they seek treatment for HIV/AIDS related diseases. Because of the inability to control for the demand for health care, the analysis above may overestimate the importance of sexual HIV transmission. The analysis could be improved if information about the symptoms or the reasons why women search for searching treatment would be available. Since this information is currently not collected, this issue is left for future research.

5 Conclusions

This study models the trade-off between data accuracy and importance of sexual HIV transmission using non-parametric propensity score methods. The results demonstrate that a large fraction of HIV infections remains unexplained by sexual HIV transmission if one assumes that women have correctly reported sexual activity. Misreporting can explain the high proportion of "virgin" HIV cases, but at the "cost" of data accuracy: The HIV epidemic in adolescent can be explained by sexual HIV transmission only if we are willing to assume that a substantial proportion of virgins (between 40 and 90%) misreport sexual status.

Two possible implications for public health policy arise from this research: (2) If misreporting is assumed to be moderate, non-sexual risk factors account for more HIV infections than previously assumed. This raises the question whether non-sexual risk factors are sufficiently covered by the current prevention paradigm and whether or not prevention interventions aimed to prevent non-sexual HIV need to be scaled-up. (2) The hypothesis that HIV is almost exclusively sexually transmitted implicitly requires that misreporting of sexual behaviour is severe. If this is true, self-reported sexual behaviour is inappropriate to monitor the epidemic and studies trying to link recent drop in HIV prevalence to changes in sexual behaviour (for example Asiimwe-Okiror et al. 1997, Gregson et al. 2006) should therefore be treated with caution.

Because misreporting is not available, this study cannot provide an estimate for the importance of sexual and non-sexual transmission modes. Better data can greatly improve our current knowledge on the importance of different HIV transmission modes. Surveys should collect detailed information on all relevant transmission modes, such as sexual HIV behaviours, health care, and illicit drug use. To validate the effects of social desirability on self-reported sexual behaviour, surveys should include testing for other sexually transmitted infections for all respondents and/or should include test questions to detect inconsistent answers. Additionally, carefully designed effectiveness trials could help to evaluate whether quality improvement of the health care system or needle-exchange programs are (cost-) effective in preventing HIV/AIDS in Sub Saharan Africa.

[11] For example 36% of the sexually active adolescents and only 13.7%% of self-reported virgins reported that they visited a health care facility in the past 12 months. Estimating the probability of visiting a health care facility (Logit model) conditional on self-reported sexual activity and all control variables of Table 2 results in a 2.3 times higher likelihood of a health facility visit for sexually active women compare to virgins. Results are not reported but are available from the author on request.

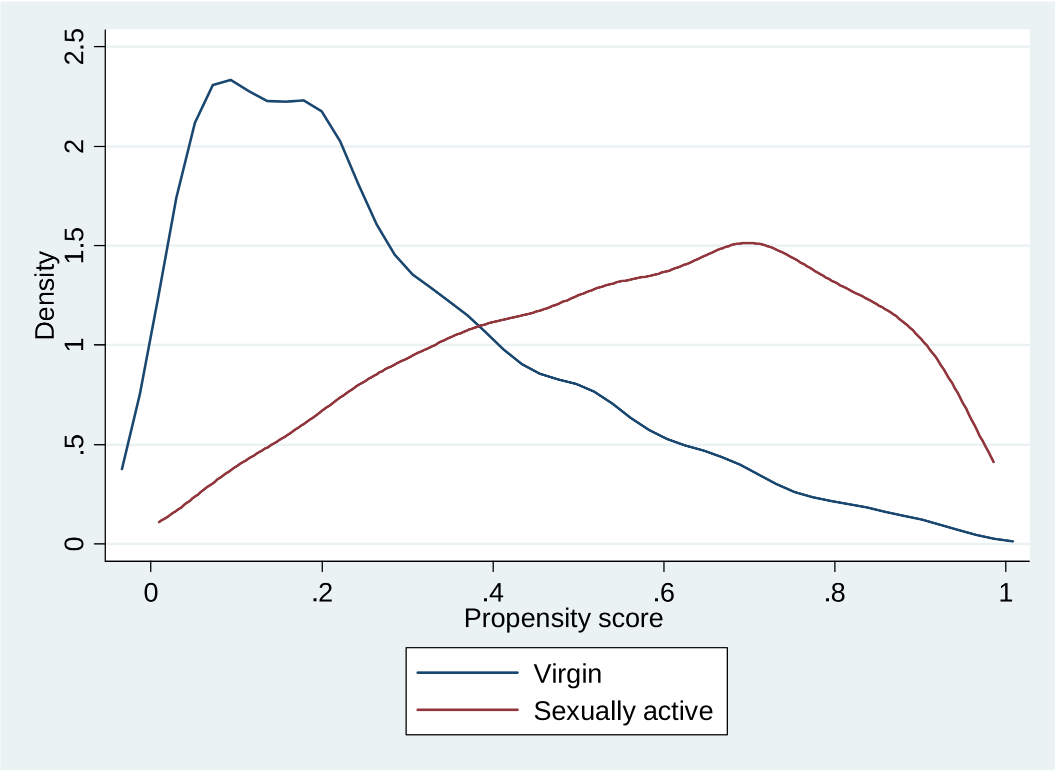
Appendix

Table A1: Descriptive statistics

	Total	Self-reported virgins	Sexually active
N	3085	1873	1212
Sexually active	0.39	0.00	1.00
HIV positive	0.06	0.04	0.09
Age	17.04	16.58	17.75
Self reported barriers to access health care (big problem vs. small problem)			
Getting permission to go	0.22	0.17	0.28
Getting money needed	0.49	0.49	0.50
Distance to health facility	0.39	0.38	0.41
Having to take transport	0.41	0.40	0.44
Not wanting to go alone	0.32	0.30	0.36
No female health provider	0.25	0.22	0.30
Education			
None	0.01	0.01	0.02
Primary	0.44	0.37	0.56
Secondary, higher	0.55	0.63	0.42
Wealth			
Poorest	0.18	0.15	0.21
Poorer	0.18	0.16	0.21
Middle	0.21	0.21	0.20
Richer	0.20	0.20	0.21
Richest	0.23	0.27	0.16
Occupation			
Not working	0.74	0.79	0.65
Sales	0.04	0.03	0.05
Agricaltural	0.13	0.09	0.20
Household & domestic	0.05	0.05	0.05
Services	0.01	0.01	0.01
Skilled & unskilled manual	0.02	0.02	0.03
Clerical	0.01	0.01	0.01
Residence: Rural	1.72	1.68	1.79
Religion			
Roman-Catholic	0.22	0.22	0.22
Other Christian	0.70	0.72	0.67
Muslim	0.03	0.02	0.04
No Religion, Traditional	0.05	0.04	0.06
Country			
Lesotho	0.25	0.22	0.28
Malawi	0.18	0.12	0.26
Zimbabwe	0.58	0.66	0.46

Note: Unweighted average values are reported; descriptive statistics for regions are not reported but are provided from the author on request

Figure A1: Kernel density estimate for propensity scores



Note: The Gaussian kernel is used to calculate the kernel density estimate. The optimal width is used which minimizes the mean integrated squared error if the data were Gaussian. The density estimate is evaluated at N=50.

References

Asiimwe-Okiror G, Opio AA, Musinguzi J, Madraa E, Tembo G, Caraël M (1997), Change in sexual behaviour and decline in HIV infection among young pregnant women in urban Uganda, *AIDS*, Vol. 11(14):1757-63.

Beckerleg S, Telfer M, Hundt GL (2005), The rise of injecting drug use in East Africa: a case study from Kenya, *Harm Reduction Journal*, Vol. 2:12.

Brewer DD, Potterat JJ, Muth SQ, Brody S (2007), Converging evidence suggests nonsexual HIV transmission among adolescents in sub-Saharan Africa, *J Adolesc Health*, Vol. 40(3): 290-291.

Buvé A, Caraël M, Hayes RJ, Auvert B, Ferry B, Robinson NJ, Anagonou S, Kanhonou L, Laourou M, Abega S, Akam E, Zekeng L, Chege J, Kahindo M, Rutenberg N, Kaona F, Musonda R, Sukwa T, Morison L, Weiss HA, Laga M (2001a), Study Group on Heterogeneity of HIV Epidemics in African Cities.The multicentre study on factors determining the differential spread of HIV in four African cities: summary and conclusions, *AIDS*, Suppl 4: S127-S131.

Buvé A, Lagarde E, Caraël M, Rutenberg N, Ferry B, Glynn JR, Laourou M, Akam E, Chege J, Sukwa T; Study Group on Heterogeneity of HIV Epidemics in African Cities (2001b), Interpreting sexual behaviour data: validity issues in the multicentre study on factors determining the differential spread of HIV in four African cities, *AIDS*, Suppl 4: S117-S126.

Clark S, Judith B, Dude A (2006), Protecting Girls from HIV/AIDS: The Case Against Child and Adolescent Marriage, I*nternational Family Planning Perspectives*, Vol. 32(2): 79-88.

Cogneau D and Grimm M (2007), Socioeconomic Status, Sexual Behavior, and Differential AIDS Mortality: Evidence from Cote D'Ivoire, *Population Research and Policy Review*, Vol. 25(4): 393-407.

de Oliveira T, Pybus OG, Rambaut A, Salemi M, Cassol S, Ciccozzi M, Rezza G, Gattinara GC, D'Arrigo R, Amicosante M, Perrin L, Colizzi V, Perno CF; Benghazi Study Group (2006), Molecular epidemiology: HIV-1 and HCV sequences from Libyan outbreak, *Nature*, Vol. 444(7121):836-7.

de Walque D (2007a), How Does the Impact of an HIV/AIDS Information Campaign Vary with Educational Attainment? Evidence from Rural Uganda, *Journal of Development Economics*, Vol. 84(2): 686-714.

de Walque D (2007b), Sero-discordant Couples in Five African Countries: Implications for Prevention Strategies de Walque, Damien, *Population and Development Review*, Vol. 33(3): 501-523.

Deuchert E and Brody S (2007), Plausible and implausible parameters for mathematical modeling of nominal heterosexual HIV transmission, *Ann Epidemiol,* Vol. 17(3): 237-244.

Gavin L, Galavotti C, Dube H, McNaghten AD, Murwirwa M, Khan R, St Louis M (2006), Factors associated with HIV infection in adolescent females in Zimbabwe, *J Adolesc Health*, Vol. 39(4): 596.e11-8.

Gavin L, Galavotti C, Dube H, McNaghten AD, Murwirwa M, Khan R, St Louis M (2007), Converging evidence suggests nonsexual HIV transmission among adolescents in sub-Saharan Africa: author reply, *J Adolesc Health*, Vol. 40(3): 291-293.

Gregson S, Garnett GP, Nyamukapa CA, Hallett TB, Lewis JJ, Mason PR, Chandiwana SK, Anderson RM (2006), HIV decline associated with behavior change in eastern Zimbabwe, *Science*, Vol. 311(5761):664-666.

GSS [Ghana Statistical Service], Health Research Unit, Ministry of Health, and ORC Macro (2003), *Ghana Service Assessment Survey 2002*, Calverton, Maryland: Ghana Statistical Service, and ORC Macro.

Hjortsberg C (2003), Why Do the Sick Not Utilise Health Care? The Case of Zambia, *Health Economics*, Vol. 12(9): 755-770.

Hutin YJ, Hauri AM, Armstrong GL (2003) Use of injections in healthcare settings worldwide, 2000: literature review and regional estimates, *BMJ*, Vol. 327(7423):1075.

Johnson, E. O., Schntz, C. G., Anthony, J. C. et al. (1995) Inhalants to heroin: A prospective analysis from adolescence to adulthood, *Drug and Alcohol Dependence*, Vol. 40: 159-164.

Jordan D (2004), Premarital Sex, Procreation, and HIV Risk in Nigeria Smith, *Studies in Family Planning*, Vol. 35(4): 223-235.

Kloos H and Mariam DH (2007), Some neglected and emerging factors in HIV transmission in Ethiopia, *Ethiop Med J*, Vol. 45(1):103-107.

Lachaud J (2007), HIV Prevalence and Poverty in Africa: Micro- and Macroeconometric Evidences Applied to Burkina Faso, *Journal of Health Economics*, Vol. 26 (3): 483-504.

Luke N (2006), Exchange and Condom Use in Informal Sexual Relationships in Urban Kenya Luke, *Economic Development and Cultural Change*, Vol. 54(2): 319-48.

Marston M, Zaba B, Salomon JA, Brahmbhatt H, Bagenda D (2005), Estimating the net effect of HIV on child mortality in African populations affected by generalized HIV epidemics, *J Acquir Immune Defic Syndr*, Vol. 38(2): 219-227.

McCurdy SA, Ross MW, Kilonzo GP, Leshabari MT, Williams ML (2006), HIV/AIDS and injection drug use in the neighborhoods of Dar es Salaam, Tanzania, *Drug and Alcohol Dependence*, Vol. 82 Suppl 1:S23-S27.

McCurdy S, Kilonzo GP, Williams M, Kaaya S (2007), Harm reduction in Tanzania: an urgent need for multisectoral intervention, *Int J Drug Policy,* Vol. 18(3): 155-159.

Mishra V, Vaessen M, Boerma JT, Arnold F, Way A, Barrere B, Cross A, Hong R, Sangha J (2006), HIV testing in national population-based surveys: experience from the Demographic and Health Surveys, *Bull World Health Organ,* Vol. 84(7): 537-545.

Mishra V, Assche SB, Greener R, Vaessen M, Hong R, Ghys PD, Boerma JT, Van Assche A, Khan S, Rutstein S (2007), HIV infection does not disproportionately affect the poorer in sub-Saharan Africa, *AIDS*, Suppl 7:S17-28.

MoHR [Ministry of Health Rwanda], National Population Office, and ORC Macro (2003): *Rwanda Service Assessment Survey 2004,* Calverton, Maryland: National Population Office, and ORC Macro.

Morgan D and Whitworth J (2001), The natural history of HIV-1 infection in Africa, *Nature Medicine,* Vol. 7(2): 143-145.

Oster E (2005), Sexually Transmitted Infections, Sexual Behavior, and the HIV/AIDS Epidemic, *Quarterly Journal of Economics*, Vol. 120(2): 467-515.

Oster E (2007), HIV and Sexual Behavior Change: Why Not Africa? National Bureau of Economic Research, Inc, *NBER Working Papers: 13049.*

Pantaleo G, Graziosi C, Fauci AS (1993), New concepts in the immunopathogenesis of human immunodeficiency virus infection, *New England Journal of Medicine*, Vol. 328: 327-335.

Philipson T and Posner R (1995), The Microeconomics of the AIDS Epidemic in Africa, *Population and Development Review*, Vol. 21(4): 835-848.

Poulton R, Caspi A, Milne BJ et al. (2002) Association between children's experience of socio-economic disadvantage and adult health: A life-course study, *Lancet*, Vol. 360: 1640-1645.

Reinherz HZ, Giacconia RM, Carmola Hauf AM et al. (2000), General and specific childhood risk factors for depression and drug disorders by early adulthood, *Journal of the American Academy of Child and Adolescent Psychiatry*, Vol. 39: 223-231.

Rosenbaum P and Rubin D (1983), The Central Role of the Propensity Score in Observational Studies for Causal Effects, *Biometrika*, Vol. 70(1): 41-55.

Sahn D, Younger S, Genicot G (2003), The Demand for Health Care Services in Rural Tanzania, *Oxford Bulletin of Economics and Statistics*, Vol. 65(2): 241-259.

Schmid GP, Buvé A, Mugyenyi P, Garnett GP, Hayes RJ, Williams BG, Calleja JG, De Cock KM, Whitworth JA, Kapiga SH, Ghys PD, Hankins C, Zaba B, Heimer R, Boerma JT (2004), Transmission of HIV-1 infection in sub-Saharan Africa and effect of elimination of unsafe injections, *Lancet*, Vol. 363(9407): 482-488.

Stronski SM, Ireland M, Michaud P. et al. (2000) Protective correlates of stages in adolescent substance use: A Swiss national study, *J Adolesc Health*, Vol. 26: 420-427.

WHO [World Health Organization] (2002a), *Quality of care: patient safety*, World Health Assembly Resolution WHA55.18, WHO: Geneva.

WHO [World Health Organization] (2002b), *Blood Safety - Strategy for the African Region*, Brazzaville: Regional Office for Africa (ADR/RC51/9 Rev1).

Zaba B, Pisani E, Slaymaker E, Boerma JT (2004), Age at first sex: understanding recent trends in African demographic surveys, *Sex Transm Infect*, Vol. 80 Suppl 2:ii28-35.