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

University of St. Gallen, Empirical Economic Research (SEW)

Abstract
The Demographic and Health Surveys from Lesotho, Zimbabwe, and Swaziland reveal that a significant proportion of HIV infections occurred in adolescent women who claim to be virgins. Two explanations for this observation have been proposed: adolescent women misreport sexual status or non-sexual risks are more prevalent than previously asserted. This paper analyses this puzzle by proposing a method to estimate bounds for the proportion of HIV prevalence attributable to sexual transmission, and applies this method to data from unmarried adolescent women. Without having detailed information on the extend of misreporting, predicted bounds for the proportion of sexual HIV transmissions are between 0.3 and 1. With moderate misreporting (between 0 and 10%), the proportion of sexual HIV transmissions would be bounded between 0.3 and 0.4. The assumption that more than 95% of HIV cases are due to sexual transmission is only valid if more than 70% of sexually active women misreport sexual activity status.

Keywords: HIV, population attributable fraction, response bias, transmission modes, bounds
JEL – Classification : C14, I10, I12

Address for correspondence: Eva Deuchert, Empirical Economic Research (SEW), University of St. Gallen, Varnbühlstrasse 14, St. Gallen, Switzerland, eva.deuchert@unisg.ch

* I thank Michael Lechner, Damien de Walque, the participants of the COST A23 Conference, and the participants of the Annual Conference of the Verein für Socialpolitik Research Committee Development Economics for helpful comments on the previous version of this manuscript. The usual disclaimer applies.
1. Introduction

It is commonly 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; Garnett et al. 2006). This belief seems inconsistent with the results of a study on HIV infections in adolescent women in Zimbabwe which shows that 41% of HIV positive women report themselves to be virgins (Gavin et al. 2006). Recent Demographic and Health Surveys from Zimbabwe, Lesotho, and Swaziland also document high infection rates among self-reported virgins (CSO 2007, 2008; MOH 2005).

Two alternative explanations for the high infection rates in self-reported virgins 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 modes (Brewer et al. 2007). The issue whether "virgin" HIV cases are explained by misreporting or non-sexual HIV transmission has yet not been solved. However, this question 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. Alternatively, a high proportion of non-sexual HIV transmission may bring up the issue whether the prevention of non-sexual HIV risks is sufficiently covered by the current prevention paradigm. The present paper addresses the "virgin" HIV puzzle by proposing a method for estimating bounds for the proportion of HIV prevalence attributable to sexual HIV transmission in case that misreporting is relevant.

With the availability of large nationally representative surveys (such as the Demographic and Health Surveys used in this paper), HIV/AIDS became focus of economic research aimed at understanding how individual behaviours and risk-factors facilitate the spread of HIV in Sub Saharan Africa. Recent examples for this research are the papers by Oster (2005; 2007), Glick and Sahn (2008), de Walque (2007a), or Cogneau and Grimm (2007). The availability of better data has greatly enhanced our understanding of Sub Saharan African's HIV/AIDS epidemic but has also revealed several anomalies that are unexplained by our current knowledge of 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) In discordant partnerships (one partner is infected and the other is not infected), the female partner is often the one who is infected, which is at odds with reported behaviour, because most infected women do not report extramarital sexual relationships (de Walque 2002b). (3) It was widely believed that the HIV/AIDS epidemic is associated with
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 had sexual intercourse is found by Gavin et al. (2006), and Buvé et al. (2001b).

Several researchers argue, that these anomalies are (at least partly) explained by misreporting of sexual behaviours (see for instance in de Walque 2007b; Gavin 2006). Social scientists are usually sceptical of the accuracy of self-reported sexual behaviours, particularly as various surveys find intercouple disagreements concerning certain sexual practices or condom use (de Boer et al. 1998), or a huge gender discrepancy in the number of sex partners (Catania et al 1995; Smith 1992). It is also argued that the initiation of sexual intercourse could be subjected to misreporting, because data from subsequent cross-sectional surveys document inconsistencies in within-cohort responses (Gersowitz 2005; Glick and Sahn 2008), and because studies using biomarkers (tests for sexually transmitted infections -STI) reveal high STI prevalence in self-declared virgins (Buvé et al. 2001b; Mensch et al. 2008). These studies document that misreporting is an important issue in many surveys on sexual behaviour. However, it remains unclear whether the response bias is sufficiently large to account for the anomalies discussed above, since to my knowledge no study has systematically addressed this issue.

Misreporting of sexual behaviour causes two problems when analyzing the importance of sexual transmission: First, the proportion of sexually active respondents is underestimated and second, the estimate for a causal effect of sexual activity on HIV prevalence is biased, because one cannot assume that the measurement error is independent of the true value (non-classical misclassification). While the first problem could in principle be solved by making an (educated) guess on the probability of misreporting, the second problem is less trivial.

Several empirical models have been developed to address the problem of non-classical measurement errors. Mahajan (2006) and Lewbel (2007) show that a point identification of a causal effect can be achieved if an instrument (i.e. an independent measure of the treatment) were available. For the research question of this paper however, this would require identifying a variable which is related to "true" sexual activity but not to misreporting sexual behaviour. Such a variable is not available. Battistin and Sianesi (2006) show how to construct bounds for the true effect, when an instrument is not available.
The paper demonstrate how the ideas of Battisti and Sianesi (2006) can be applied to estimate bounds for the proportion of HIV cases attributable to sexual transmission if people misreport sexual activity. It applies the method to data from adolescent women who are not in a fixed relationship to demonstrate that only 30% of HIV infections in the sample can be attributed to sexual HIV transmission if one assumes that the data is accurate. With no detailed information on misreporting, bounds are uninformative (0.3-1); more detailed information on misreporting can sharpen the bound. If for example misreporting is modest (0-10%), the proportion of sexually transmitted infections would be bounded between 0.3 and 0.4. The hypotheses that more than 95% of HIV infections are due to sexual transmission is only valid if a substantial proportion of sexually active women (more than 70%) have misreported sexual activity.

The rest of the paper is organized as follows: The next section outlines the model that is used to estimate the proportion of sexual HIV transmission. Section 3 provides an empirical application using data from Lesotho, Swaziland, and Zimbabwe. The conclusion can be found in the final section.

2. An empirical model of sexual HIV transmission

Sexual HIV transmission is typically modelled in the context of a Bernoulli-risk simulation model (see for instance in Oster 2005). In these models each sexual episode or sexual partnership is treated as an independent trial and the probability of a "success" (HIV infection) 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 must be based on assumptions, are measured with a great uncertainty, or are constrained by misreporting. Simulation outcomes are therefore constrained by great uncertainty (a discussion on this issue can be found in Deuchert and Brody 2007).

An alternative to risk simulation models is to estimate population attributable fractions \(-PAF\), which are defined as the "proportion of disease cases over a specified time that would be prevented following elimination of the exposures, assuming the exposures are causal" (Rockhill et al. 1998, p. 15). This can be formulated as

\[
PAF = \frac{E(Y) - E(Y \mid NS)}{E(Y)}
\]  

(1)
where $E(Y)$ is the proportion of HIV (prevalence) in the population, and $E(Y|NS)$ represents the proportion of non-sexually acquired HIV infections in the population.

To estimate equation (1), one need to consistently estimate the prevalence of HIV from non-sexual transmission. By definition, all HIV infections in true virgins ($A^* = 0$) have acquired HIV by non-sexual transmission routes. The proportion of HIV infections in virgins however, reflect only the lower bound for $E(Y|NS)$ since sexually active individuals ($A^* = 1$) may have been also exposed to non-sexual risks. A consistent estimator for the total prevalence of HIV from non-sexual transmission therefore need to take potential nonsexual transmission in sexually active respondents into account.

To estimate the probability of non-sexual HIV transmission, the potential outcome notation is used. Let $Y^i$ denote the outcome for an individual that is sexually active and let $Y^0$ be the outcome for the same individual if she was not sexually active. If $A^*$ is an indicator for sexual activity, the realized (observed) outcome $Y$ for an individual equals

$$Y = Y^0 (1 - A^*) + Y^i A^*$$

The HIV prevalence from non-sexual transmission is $E(Y^0)$, which is the HIV prevalence that would occur if no one was sexually active. To consistently estimate $E(Y^0)$, two problems need to be solved: (1) One can in principle observe a person in only one state (sexually active or inactive) so that the counterfactual outcome can never be observed. (2) Misreporting of sexual activity is not observable, so that it is unclear whether the observed outcome $Y$ is the outcome for an individual that is truly sexually active or that is inactive.

To solve these two problems, the paper applies the ideas from Battistin and Sianesi (2006), who estimate bounds for the causal effect of education on labour outcomes, given that education is misreported. This approach is based on the following assumptions:

(i) Suppose we had a prior information on the fraction of respondents who provide inaccurate answers (which could be provided by studies using biomarkers for purely sexually transmitted infections for example):

$$v_1 \leq P(Z = 1) \leq v_2$$

where $Z$ indicates whether a respondent provides inaccurate information, $v_1$ is the lower bound and $v_2$ is the upper bound for the fraction of misreporting. By varying the value of $v_1$ and $v_2$, one can consider a wide range of views characterising the debate on inaccurate self-reported sexual behaviours. The most extreme case is where it is only known that some people misreport (i.e. $v_1 = 0$ and $v_2 = 1$).
(ii) Prior information can be also used for a verification of observed subgroups -VOS-, which assumes fully accurate self-reports within particular groups. This may be particularly relevant in the current example because the traditional cultural value for young respondents to remain virgins until marriage (see for example in Cowan et al. 2002) sets an incentive for misreporting for sexually active respondents but not for virgins. Therefore, assume that:

\[
P(Z = 1 | A^* = 1) = P(A = 0 | A^* = 1) \geq 0
\]
\[
P(Z = 1 | A^* = 0) = P(A = 1 | A^* = 0) = 0
\]

If misreporting varies with observed covariates (see for example Rodgers 1982), the total proportion misreporting equals

\[
P(Z = 1) = \int \lambda(x) P(A^* = 1 | x) f(x) dx
\]

with \( \lambda(x) = P(Z = 1 | A^* = 1, x) \) being the conditional probability that a sexually active respondent misreports.

(iii) The Conditional Independence Assumption -CIA- states that the true sexual activity status \( A^* \) is independent of the potential outcomes conditional on the value of suitably chosen covariates:

\[
Y_1, Y_0 \perp \!
\!
\perp A^* | X = x
\]

Thus, assuming CIA holds, the conditional HIV prevalence from non-sexual transmission can be rewritten as

\[
E[Y_0 | A^* = 0, x] = E[Y_0 | x] = E[Y_0 | A^* = 1, x]
\]

(iv) The Non-Differential Misclassification Assumption -NDM- solves the problem that the confounder distribution of women who misreport sexual activity is unknown. NDM assumes that the reported sexual behaviour is uninformative about outcome (HIV prevalence) given the true sexual activity status and all other covariates:

\[
Y_1, Y_0 \perp \!
\!
\perp A | X = x, A^*
\]

NDM does not require that misreporting be entirely random. Misreporting can be related to the true outcome due to its correlation with other variables in the model. The condition however holds true only if it is possible to control for all characteristics that jointly influence sexual activity, misreporting, and the considered outcome.
CIA and NDM can now be used to construct an estimator for the proportion of non-sexual HIV transmission. Note that the conditional prevalence for self-reported virgins can be rewritten as

\[
E(Y \mid A = 0, x) = E(Y^i \mid A^* = 1, A = 0, x) P(Z = 1 \mid A = 0, x) \\
+ E(Y^0 \mid A^* = 0, A = 0, x) P(Z = 0 \mid A = 0, x)
\]

If NDM hold, the equation above equals

\[
E(Y \mid A = 0, x) = E(Y^i \mid A^* = 1, x) P(Z = 1 \mid A = 0, x) + E(Y^0 \mid A^* = 0, x) P(Z = 0 \mid A = 0, x) \quad (2)
\]

The first part of equation (2) is identified by NDM and the assumption that respondents who admit sexual activity have no incentive to misreport sexual activity status. It can therefore be directly estimated from reported data: \(E(Y^i \mid A^* = 1, x) = E(Y \mid A = 1, x)\).

**Proof:**

\[
E(Y \mid A = 1, x) = E(Y \mid A^* = 1, A = 1, x) \left[ 1 - P(A^* = 0 \mid A = 1, x) \right] \\
+ E(Y \mid A^* = 0, A = 1, x) P(A^* = 0 \mid A = 1, x)
\]

\[
= E(Y \mid A^* = 1, A = 1, x) = E(Y^i \mid A^* = 1, x)
\]

as

\[
P(A^* = 0 \mid A = 1, x) = P(Z = 1 \mid A^* = 0, x) \frac{P(A^* = 0 \mid x)}{P(A = 1 \mid x)} = 0
\]

This can be introduced into equation (2), which can be reformulated to get the conditional HIV prevalence from non-sexual transmission:

\[
E(Y^0 \mid x) = E(Y \mid A^* = 0, x) = \frac{E(Y \mid A = 0, x) - P(Z = 1 \mid A = 0, x) E(Y \mid A = 1, x)}{1 - P(Z = 1 \mid A = 0, x)} \quad (3)
\]

with

\[
P(Z = 1 \mid A = 0, x) = \frac{P(A = 1 \mid x)}{1 - P(A = 1 \mid x)} \frac{\lambda(x)}{1 - \lambda(x)}
\]

Note that equation (3) need to be strictly between zero and one, which sets a binding constraint for the probability of misreporting \(\lambda(x)\):

\[
(C1) \quad P(Z = 1 \mid A = 0, x) \leq 1 \quad \Rightarrow \lambda(x) \leq 1 - P(A = 1 \mid x)
\]

\[\text{Note that } P(A^* = 1 \mid A = 0, x) = P(A = 0, A^* = 1, x) \times P(A^* = 1 \mid x) / P(A = 0 \mid x) \]

and \(P(A^* = 1 \mid x) = P(A = 1 \mid x) + P(A^* = 1 \mid A = 0, x) \times P(A = 0 \mid x)\)
\[
(C2) \quad E(Y^0 | x) \leq 1 \quad \Rightarrow \lambda(x) \leq \frac{1 - E(Y | A = 0, x)}{1 - E(Y | A = 1, x)} \frac{1 - E(Y | A = 0, x) + P(A = 1 | x)}{1 - E(Y | A = 1, x) + 1 - P(A = 1 | x)}
\]

The total HIV prevalence on non-sexual transmission can be estimated by integrating the conditional HIV prevalence over the full distribution of \(X\):

\[
E(Y^0) = \int E(Y^0 | x) f(x) dx
\]  

Thus, an estimator for HIV prevalence from non-sexual HIV transmission can be derived from a consistent estimator for HIV prevalence conditional on self-reported sexual behaviour \(E(Y|A=i,x)\), the probability to report sexual activity \(P(A=1|x)\), and some prior information regarding the probability of misreporting \(\lambda(x)\).

3. Empirical application

3.1. Data

The data is from the Demographic and Health Surveys (DHS) in Lesotho, Zimbabwe, and Swaziland. DHS conduct interviews on maternal and child health, family planning, nutrition and related issues, and also collect biological and clinical data such as HIV testing (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%.

The analysis is restricted to adolescent women (age 15-19) who declare that they have never been married. The reason, why women who have been in fixed relationships are disregarded in the present analysis lies in the nature of the cross-sectional data. Most data is available on the household level (such as wealth, or the place of residence). Women in fixed relationships usually move together with their partner so that the household characteristics are not the same than when they have initially decided to be become sexually active. The age restriction is necessary because most adult women (age>19) declare to be sexually active. In total, 3,124 women are considered.

The surveys are conducted by face-to-face interviews, where the individual (woman) questionnaire is administered by a female interviewer. The interviewer strives to interview the respondent alone, since a third person during an interview can prevent them from getting hon-
est answers. If this was not possible, the presence of another person was coded. 12% of respondents in the considered sub-sample was interviewed with another person being present.

The surveys are designed such that the questionnaire on socio-economic variables and individual behaviour is independent from HIV testing. After finishing the questionnaire, all men and women who live in a randomly selected sub-sample of households (usually one third of survey households) are asked whether they accept a HIV test. Rates of consent for HIV testing are very high, where more than 95% of women in the considered sub-group accepted being tested for HIV.

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

<table>
<thead>
<tr>
<th></th>
<th>Sexually active</th>
<th>Virgin</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total</td>
<td>25.1%</td>
<td>74.9%</td>
</tr>
<tr>
<td>HIV prevalence</td>
<td>12.7%</td>
<td>4.0%</td>
</tr>
<tr>
<td>Proportion of total HIV infection</td>
<td>51.8%</td>
<td>48.2%</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>Lesotho</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total</td>
<td>33.7%</td>
</tr>
<tr>
<td>HIV prevalence</td>
<td>10.0%</td>
</tr>
<tr>
<td>Proportion of total HIV infection</td>
<td>48.8%</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>Swaziland</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total</td>
<td>37.4%</td>
</tr>
<tr>
<td>HIV prevalence</td>
<td>14.6%</td>
</tr>
<tr>
<td>Proportion of total HIV infection</td>
<td>66.7%</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>Zimbabwe</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total</td>
<td>11.7%</td>
</tr>
<tr>
<td>HIV prevalence</td>
<td>11.6%</td>
</tr>
<tr>
<td>Proportion of total HIV infection</td>
<td>31.7%</td>
</tr>
</tbody>
</table>

Table 1 shows that approximately 75% of adolescent women in the considered sample report that they are virgins. The highest proportion of self-reported virgins can be found in Zimbabwe (88%); the lowest proportion can be found in Swaziland (63%). Approximately 6.2% of women in the sample considered is HIV positive, with HIV prevalence rates highest in Swaziland and lowest in Zimbabwe. A significant share of HIV infections (48%) occurred in women, who reported themselves virgins. This proportion is highest in Zimbabwe (68%) and lowest in Swaziland (33%). This demonstrates that a significant share of HIV infections in this sample is unexplained by reported sexual behaviour. This could be either a sign for mis-

---

2 The remaining descriptive statistics for the sample are presented in the appendix (Table A1).
reporting or for non-sexual HIV transmission. A detailed analysis of this issue is therefore necessary.

3.2. Identification strategy

The Conditional Independence Assumption and the Non-Differential Misclassification Assumption cannot be directly tested. To make these assumptions creditable, it is therefore essential to understand how HIV is transmitted and which factors determine the reliability of the data concerning sexual behaviour.

HIV transmission modes

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, (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 introduced in 1997, so that effective HIV treatment had not been available to women belonging to this sample at infancy. Because average 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 sample are caused by mother-to-child transmission.

It is of interest to estimate the HIV prevalence by non-sexual transmission. Therefore, it is necessary to control for all confounders that potentially influence blood-to-blood and sexual HIV transmission.

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).1 Relatively little is known about injection drug use in Sub Saharan Africa but it is suggested that

---

1 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).
injection drug use is an emerging HIV risk as a result of the expansion of international drug trafficking (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 (McCurdy 2005; McCurdy et al. 2007). To my knowledge, risk factors that lead to injection 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 for public health authorities (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). The individual demand for health care is typically influenced by 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).

Various studies document that HIV infections are associated with sexual activity. For example a systematic overview of 68 Epidemiological Studies conducted in Sub Saharan Africa identified sexual risk factors such as multi-partner sex, paid sex, and STIs infections as important risk factors for HIV transmission (Chen et al. 2007). These observations have led to the conclusion that HIV is predominantly sexually transmitted in Sub Saharan Africa accounting for more than 95% of HIV infections (Schmid et al. 2004; Garnett et al. 2006). Since the current paper considers unmarried women, the risk factor of interest is premarital sexual activity. Representative surveys reveal that reported median age of first sex at less than 20 in almost all Sub Saharan African countries (Zaba et al. 2004). Since self-reported age of first sexual activity has either remained constant or increased slightly, while the age at first marriage increased dramatically over the past 15 years, the proportion of young women who engage in premarital sex has increased substantially (Mensch et al. 2006). A higher likelihood to engage in premarital sexual relations is related to 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).
Reliability of self-reported data

The information of the initiation of sexual activity cannot be externally validated. Researchers have used three different strategies to analyze the consistency and reliability of self-reported sexual behavior: (1) The reliability of data is examined within data reported by the same individual or the same age-cohorts at different times, which demonstrated inconsistencies in self-reported age of first sexual intercourse (Gersowitz 2005; Glick and Sahn 2008; Rodgers 1982). Inconsistencies in answers are correlated with observable characteristics (such as race and gender), which suggest that errors in the measurement of sexual activity are not random (Rodgers 1982). It was argued that misreporting could be associated with age as young women may find great difficulty admitting to an interviewer that they have had sex outside of a socially sanctioned relationship (Dare and Cleland 1994), or with religious beliefs as the Catholic Church and some evangelical Protestant churches heavily promote pre-marital abstinence (Mensch et al. 2008). (2) Examining the impact of the data collection mode, it was noticed that women provide more accurate information about sexual behaviour if the interviewer is female (Wilson et al. 2002), or that the presence of other people may affects the respondent’s ability to respond honestly (Aquilino et al. 2000). (3) Several studies use biomarkers (i.e. tests for various sexually transmitted infections) to understand whether respondents who reported that they had never had sexual intercourse had misreported sexual activity status (Buvé et al. 2001b; Mensch et al. 2006).

All studies mentioned above document that misreporting is an issue in surveys on sexual behaviour. However, these studied cannot quantify the extend of misreporting: the first two methods can only reveal inconsistencies in reports, but cannot predict the accuracy of the responses; the third method crucially relies on the assumption that sexual transmission is the only transmission mode, which is can lead to misleading results if non-sexual transmission modes are also relevant (this is particularly the case since the studies by Mensch et al. 2008 and Buvé et al. 2001b use positive HIV tests to detect ”inaccurate reports”).

It follows from the discussion above that blood-to-blood transmission, sexual behaviour and data inaccuracy are likely to be influenced by the same factors. Thus, the analysis needs to control for factors that jointly influence non-sexual risks, sexual activity, and misreporting.

---

4 Virginity tests (medical test that verifies whether or not a woman's hymen is intact) cannot provide a precise answer because the hymen can be intact even when the woman started having sex, and there could be other reasons which causes the hymen to break.
These factors can be categorized into the following groups: (1) socio-economic variables, (2) socio-cultural variables, (3) socio-geographic variables, (4) bargaining power, and (5) survey characteristics.

The Demographic and Health Surveys provide a large set of covariates that can be used to approximate these categories such as age, education, household wealth, occupation, religion, regional indicators, a variable that indicates the presence of other people during the interview, and a set of self-reported barriers to health care, which approximate geographic accessibility, individual bargaining power, and moral values.

Unfortunately, Demographic and Health Surveys are cross-sectional and it is not possible to control for confounding variables before women decide to initiate in sexual activity (which would rule out that variables are influenced by sexual activity or resulting HIV infections). For two reasons I believe that this does not cause a major bias: (1) The period of clinical latency (where patients have no symptoms) is about 10 years (Pantaleo et al. 1993), and is therefore much longer than the time span of reported sexual activity in this age cohort (max 8 years, median 1 year). Therefore, it is unlikely that education, occupation choice, productivity etc. is influenced by resulting HIV infections, because women who acquired HIV by sexual transmission should not show symptoms of AIDS yet. (2) Other variables are measured on the household level (such as wealth, residence, etc.) which is unlikely to be affected by adolescent women's decision to initiate sexual activity or subsequent HIV infections.

A second problem may arise because important information, such as current income and ethnicity is not available from the Demographic and Health Surveys. 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 most likely approximate income; cultural differences are likely to be captured by the indicators for regional differences,\(^5\) which are closely related to differences in ethnic composition.

A further limitation arises from this identification strategy, because this analysis implicitly assumes that the uncomfounded risk increase in HIV 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). It is not possible to adjust for variables that approximate exposure to health care, because available variables on access to

\(^5\) The analysis includes dummy variables for 24 regions.
health care (injections in the previous months, visits of health facilities) are constrained by reverse causality. The paper controls for self-reported barriers to access health care, which is likely to partly capture this effect.

3.3. Predicting the probability of self-reported sexual activity and HIV prevalence

To predict the population attributable fraction (equation 1), one need to consistently estimate the conditional HIV probability to report sexual activity \( P(A=0|x) \) and the conditional HIV prevalence \( E(Y|A,x) \). This is done by using two Probit models, where the dependent variables are variables indicating that a person reported sexual activity \((A)\), that a person is HIV positive \((Y)\). The control variables are all variables presented in table A2, dummy variables for 24 regions, and interactions terms between wealth and country dummy variables (not presented). The interaction terms are included because the wealth indicator is constructed for each country separately (for a construction of this indicator see Rutstein and Johnson 2004), and thus measures the relative wealth distribution within a country and not across the three countries.

Heteroskedasticity is likely to be a problem because pooled data from three different countries is used. I therefore assume that the variance can vary as a function of dummy variables for the different countries (Swaziland SWA and Zimbabwe ZIM):

\[
\sigma_i = \exp(SWA\gamma_1 + ZIM\gamma_2) 
\]

The regression results are presented in the appendix (Table A2). The first column reveals that older age, working (compare to not-working), and not wanting to go alone to health providers is positively associated with self-reported sexual activity; higher education, being Christian (Roman Catholic and Protestant), and rural residence are negatively associated with self-reported sexual activity.

The second column shows that HIV prevalence is significantly higher if a person reports sexual activity but does not does significantly vary with most socio-economic variables; only the coefficients for the second lowest wealth quintile, money as a barrier for accessing health care, the presence of other person during interview, and rural residence is are significantly different to zero. It is likely that the effects of other variables is too small to get detected in this sample, since only 193 women in the considered sample are HIV positive.

\[\text{Battistin and Sianesi (2006) use a propensity score matching estimator to construct conditional treatment effects. This paper does not use this method, because conditional prevalence rates from propensity score matching will not be consistent.}\]
3.4. Estimating the proportion of sexual HIV transmission

Estimates from the Probit model discussed above are used to construct an estimator for the proportion of sexual HIV transmission. If one is willing to assume that all respondents have correctly specified their sexual activity status ($\lambda(x)=0$ for all $x$), total HIV prevalence by non-sexual transmission (equation 4) is equal to 4.4%. The corresponding population attributable fraction (equation 1) can then be estimated by setting the HIV prevalence by sexual transmission into relation with the total HIV prevalence: $(6.2-4.4)/6.2=0.3$ [95% CI: 0.20-0.39]. Thus, the scenario of no misreporting would predict that only 30% of HIV cases in the sample are attributable to sexual transmission, 70% are attributable to non-sexual routes (either before or after the onset of sexual activity).

Figure 1: Range of possible bounds for the proportion of sexually transmitted HIV infections

Note: Bounds are estimated based on the Probit model coefficients from Table A2; confidence intervals are based on an approximate variance of the sampling distribution of the liberalized ratio of sample totals for HIV prevalence $Y$ and HIV prevalence from non-sexual transmission $Y_1$: $R = Y_1/Y$; $\text{Var}(R) = (Y)^2 \left[ \text{Var}(Y_1) - 2R\text{Cov}(Y,Y_1) + 2R^2\text{Var}(Y) \right]$. 

The scenario that women do not misreport sexual behaviour however, may not be realistic and a population attributable fraction of 30% probably reflects the lower bound for the proportion of HIV cases attributable to sexual transmission. More realistic scenarios, that allow for misreporting [$\lambda(x)>0$] may provide a more precise picture on the importance of sexual
HIV transmission. Figure 1 presents the full range of possible bounds for the population attributable fractions. Suppose that no detailed information would be available on the proportion of misreporting \((v_1=0; v_2=1)\), the proportion of HIV cases due to sexual transmission is bounded between 0.3 and 1. With more specific priors regarding the probability of misreporting, sharper bounds could be derived: In a scenario where misreporting is assumed to be moderate \((v_1=0; v_2=0.1)\), the proportion of sexually transmitted HIV cases would be bounded between 0.3 and 0.4; in a medium scenario with the probability of misreporting among sexually active individuals between 20 and 30\%, the proportion of sexually transmitted HIV cases would be bounded between 0.5 and 0.6. To be able to conclude that more than 95\% of HIV cases in the considered population is due to sexual transmission (bounds: 0.95-1), one need to assume that more than 70\% of the sexually active population misreports sexual activity.

The results for the different countries are presented in the Table 3, which demonstrates that bounds for the proportion of sexual HIV transmission are not constant across the three countries. With moderate misreporting, the percentage of sexually transmitted HIV cases is bounded between 0.37 and 0.54 in Swaziland, but only between 0.2 and 0.24 in Zimbabwe. In a scenario, where more than half of the sexually active population misreports sexual activity, PAFs are bounded between 0.51 and 1 in Zimbabwe and 0.9 and 1 in Swaziland.

<table>
<thead>
<tr>
<th>Country</th>
<th>0-10%</th>
<th>20-30%</th>
<th>&gt;50%</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>lower</td>
<td>upper</td>
<td>lower</td>
</tr>
<tr>
<td>Lesotho</td>
<td>0.26</td>
<td>0.35</td>
<td>0.47</td>
</tr>
<tr>
<td>Swaziland</td>
<td>0.37</td>
<td>0.54</td>
<td>0.67</td>
</tr>
<tr>
<td>Zimbabwe</td>
<td>0.20</td>
<td>0.24</td>
<td>0.29</td>
</tr>
</tbody>
</table>

*Note: Bounds are estimated based on the Probit model coefficients from Table A2*

The results presented so far rely on the assumption, that all sexually active women have identical probability of misreporting. This may not be realistic, since empirical studies have demonstrated that inconsistencies in reports are correlated with observable characteristics (Rodgers 1982). To investigate the sensitivity of the results to different assumptions regarding factors associated with misreporting, the paper uses different scenarios: (1) Misreporting could be associated with religions or cultural values, particularly if the tradition places a high value on virginity. The Catholic Church and certain evangelical Protestant churches for example, heavily promote premarital abstinence, which may not only have an effect on the decision to initiate a premarital partnership, but also on the likelihood to admit any premarital
sexual relations. Bounds for population attributable fractions are therefore estimated assuming the probability of misreporting to be twice as high if women report to be Christian (Catholic or Protestant). (2) Young women may be less mature about their own sexual behaviours and thus, may be more likely to misreport sexual activity. Thus, bounds for PAFs are estimated assuming the probability of misreporting to be twice as high if women is younger than 17. (3) The presence of a third person during an interview may prevent respondents getting honest answers. The third scenario therefore assumes that women who were interviewed while another person was present misreport twice as often sexual activity. The results of these sensitivity checks are presented in the appendix (Figure A1). It can be seen that the results are not sensitive to the chosen scenario, as bounds are almost identical.

As a further sensitivity check, the Probit models are re-estimated using different sets of control variables. This analysis helps to understand, if the failure to control for all relevant factors causes a major bias in the result. Figure A2 demonstrates that the results presented in this paper are driven by controlling for age and regional characteristics. Bounds for the population attributable fractions from a restricted model that controls only for age and regional characteristics are almost identical to the full model, but differ to the naïve model that does not control for any confounding variables. This shows that other variables have only small impact, and the failure to control for all possible confounders is likely to be small.

The results presented in this paper demonstrate that there is a positive association between misreporting and the proportion of sexually transmitted HIV cases. With no detailed prior on misreporting \( 0<P(Z=1)<1 \) however, no sharp bounds can be derived. From this study, one can only conclude that the fraction of sexually transmitted HIV infections is bounded between 0.3 and 1. However, this study documents that a substantial proportion of the sexually active population need to have misreported sexual activity status (>70%) to be able to conclude that HIV is predominantly sexually transmitted (bounds 0.95-1). With more detailed information on the extend of misreporting, sharper bounds could be derived. This is was not possible in the current study and is left for future research.

4. Conclusions

This paper proposes a method to estimate the importance of sexual HIV transmission if women systematically misreport sexual activity status. This method is applied to data from Lesotho, Swaziland, and Zimbabwe to derive bounds for the relative importance of sexual HIV transmission. The results demonstrate that a large share of HIV infections remains unex-
plained 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, where a substantial proportion of sexually active women need to have mis-reported sexual status.

Two possible implications for public health research and policy arise from this study: (2) If one assumes that misreporting is 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 intensified. (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 an inappropriate measure for monitoring the epidemic.

Sharper bounds cannot be provided since relevant information concerning the importance of misreporting was not available. Better data can greatly improve our current knowledge on the importance of different HIV transmission modes. Future surveys on HIV/AIDS in Sub-Saharan Africa should therefore collect detailed information on all relevant transmission modes, such as sexual behaviours, health care, and illicit drug use. Biomarker studies aimed to detect response bias should make sure that positive test result cannot be due to non-sexual transmission, which excludes HIV as a possible biomarker.
### Appendix

**Table A1: Descriptive statistics (sample means)**

<table>
<thead>
<tr>
<th></th>
<th>Sexually active</th>
<th>Virgin</th>
<th>Pearson chi2(1)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Age</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Age = 15</td>
<td>0.08</td>
<td>0.26</td>
<td>108.50 ***</td>
</tr>
<tr>
<td>Age = 16</td>
<td>0.14</td>
<td>0.29</td>
<td>68.05 ***</td>
</tr>
<tr>
<td>Age = 17</td>
<td>0.23</td>
<td>0.20</td>
<td>3.55 *</td>
</tr>
<tr>
<td>Age = 18</td>
<td>0.27</td>
<td>0.15</td>
<td>52.14 ***</td>
</tr>
<tr>
<td>Age = 19</td>
<td>0.28</td>
<td>0.11</td>
<td>143.50 ***</td>
</tr>
<tr>
<td><strong>Wealth</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Poorest</td>
<td>0.16</td>
<td>0.15</td>
<td>0.00</td>
</tr>
<tr>
<td>Poorer</td>
<td>0.19</td>
<td>0.17</td>
<td>0.90</td>
</tr>
<tr>
<td>Middle</td>
<td>0.21</td>
<td>0.21</td>
<td>0.00</td>
</tr>
<tr>
<td>Richer</td>
<td>0.22</td>
<td>0.21</td>
<td>0.32</td>
</tr>
<tr>
<td>Richest</td>
<td>0.23</td>
<td>0.26</td>
<td>2.13</td>
</tr>
<tr>
<td><strong>Education</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>No, incomplete primary</td>
<td>0.32</td>
<td>0.24</td>
<td>19.07 ***</td>
</tr>
<tr>
<td>Complete primary</td>
<td>0.14</td>
<td>0.09</td>
<td>11.82 ***</td>
</tr>
<tr>
<td>Incomplete secondary</td>
<td>0.49</td>
<td>0.64</td>
<td>53.34 ***</td>
</tr>
<tr>
<td>Complete secondary, higher</td>
<td>0.05</td>
<td>0.02</td>
<td>12.08 ***</td>
</tr>
<tr>
<td><strong>Occupation</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Not working</td>
<td>0.77</td>
<td>0.85</td>
<td>28.09 ***</td>
</tr>
<tr>
<td>Agricultural</td>
<td>0.04</td>
<td>0.05</td>
<td>1.46</td>
</tr>
<tr>
<td>Household &amp; Domestic</td>
<td>0.04</td>
<td>0.04</td>
<td>0.04</td>
</tr>
<tr>
<td>Other</td>
<td>0.15</td>
<td>0.06</td>
<td>63.70 ***</td>
</tr>
<tr>
<td><strong>Religion</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>None</td>
<td>0.04</td>
<td>0.03</td>
<td>0.03</td>
</tr>
<tr>
<td>Roman Catholic</td>
<td>0.14</td>
<td>0.10</td>
<td>9.50 ***</td>
</tr>
<tr>
<td>Protestant Churches</td>
<td>0.60</td>
<td>0.77</td>
<td>84.43 ***</td>
</tr>
<tr>
<td>Other</td>
<td>0.23</td>
<td>0.10</td>
<td>84.99 ***</td>
</tr>
<tr>
<td><strong>Self reported barriers to access health care (big vs. small problem)</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Getting permission to go</td>
<td>0.97</td>
<td>0.94</td>
<td>8.95 ***</td>
</tr>
<tr>
<td>Getting money needed</td>
<td>0.64</td>
<td>0.59</td>
<td>7.97 ***</td>
</tr>
<tr>
<td>Distance to health facility</td>
<td>0.71</td>
<td>0.66</td>
<td>6.56 ***</td>
</tr>
<tr>
<td>Having to take transport</td>
<td>0.73</td>
<td>0.67</td>
<td>10.20 ***</td>
</tr>
<tr>
<td>Not wanting to go alone</td>
<td>0.86</td>
<td>0.78</td>
<td>25.73 ***</td>
</tr>
<tr>
<td>No female health provider</td>
<td>0.87</td>
<td>0.87</td>
<td>0.00</td>
</tr>
<tr>
<td>Residence: Rural</td>
<td>0.71</td>
<td>0.71</td>
<td>0.07</td>
</tr>
<tr>
<td>Presence of other person during interview</td>
<td>0.12</td>
<td>0.12</td>
<td>0.00</td>
</tr>
</tbody>
</table>

*Note: Pearson's chi-squared for the hypothesis that the rows and columns in a two-way table are independent is presented; significance levels: *** p<0.01, ** p<0.05, * p<0.1*
Table A2: Regression results for the heteroskedastic Probit models

<table>
<thead>
<tr>
<th></th>
<th>Coef.</th>
<th>z</th>
<th>Coef.</th>
<th>z</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Sexual activity</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Age</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Age = 16</td>
<td>0.32</td>
<td>3.39</td>
<td>-0.11</td>
<td>-1.14</td>
</tr>
<tr>
<td>Age = 17</td>
<td>0.83</td>
<td>6.71</td>
<td>-0.02</td>
<td>-0.23</td>
</tr>
<tr>
<td>Age = 18</td>
<td>1.10</td>
<td>7.54</td>
<td>0.09</td>
<td>0.79</td>
</tr>
<tr>
<td>Age = 19</td>
<td>1.34</td>
<td>8.02</td>
<td>0.07</td>
<td>0.63</td>
</tr>
<tr>
<td><strong>Wealth</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Poorer</td>
<td>0.31</td>
<td>1.53</td>
<td>0.58</td>
<td>*</td>
</tr>
<tr>
<td>Middle</td>
<td>0.04</td>
<td>0.20</td>
<td>0.31</td>
<td>0.91</td>
</tr>
<tr>
<td>Richer</td>
<td>0.13</td>
<td>0.62</td>
<td>0.13</td>
<td>0.35</td>
</tr>
<tr>
<td>Richest</td>
<td>0.07</td>
<td>0.34</td>
<td>0.37</td>
<td>1.06</td>
</tr>
<tr>
<td><strong>Education</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Complete primary</td>
<td>-0.14</td>
<td>-1.58</td>
<td>0.02</td>
<td>0.19</td>
</tr>
<tr>
<td>Incomplete secondary</td>
<td>-0.36</td>
<td>-4.75</td>
<td>-0.14</td>
<td>-1.48</td>
</tr>
<tr>
<td>Complete secondary, higher</td>
<td>-0.44</td>
<td>-3.10</td>
<td>-0.33</td>
<td>-1.52</td>
</tr>
<tr>
<td><strong>Occupation</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Agricultural</td>
<td>0.45</td>
<td>2.69</td>
<td>0.20</td>
<td>1.10</td>
</tr>
<tr>
<td>Household &amp; Domestic</td>
<td>0.33</td>
<td>2.11</td>
<td>0.21</td>
<td>1.28</td>
</tr>
<tr>
<td>Other</td>
<td>0.48</td>
<td>4.76</td>
<td>0.13</td>
<td>1.08</td>
</tr>
<tr>
<td><strong>Religion</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Roman Catholic</td>
<td>-0.35</td>
<td>-1.97</td>
<td>-0.36</td>
<td>-1.61</td>
</tr>
<tr>
<td>Protestant Churches</td>
<td>-0.34</td>
<td>-2.25</td>
<td>-0.09</td>
<td>-0.58</td>
</tr>
<tr>
<td>Other</td>
<td>-0.13</td>
<td>-0.79</td>
<td>0.06</td>
<td>0.35</td>
</tr>
<tr>
<td><strong>Self reported barriers to access health care (big vs. small problem)</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Getting permission to go</td>
<td>0.13</td>
<td>0.84</td>
<td>-0.08</td>
<td>-0.57</td>
</tr>
<tr>
<td>Getting money needed</td>
<td>-0.09</td>
<td>-1.38</td>
<td>-0.15</td>
<td>*</td>
</tr>
<tr>
<td>Distance to health facility</td>
<td>0.08</td>
<td>0.93</td>
<td>-0.04</td>
<td>-0.37</td>
</tr>
<tr>
<td>Having to take transport</td>
<td>-0.03</td>
<td>-0.39</td>
<td>0.19</td>
<td>1.57</td>
</tr>
<tr>
<td>Not wanting to go alone</td>
<td>0.25</td>
<td>3.03</td>
<td>0.06</td>
<td>0.60</td>
</tr>
<tr>
<td>No female health provider</td>
<td>-0.07</td>
<td>-0.87</td>
<td>0.06</td>
<td>0.59</td>
</tr>
<tr>
<td>Residence: Rural</td>
<td>-0.21</td>
<td>-2.62</td>
<td>-0.15</td>
<td>-1.38</td>
</tr>
<tr>
<td>Presence of other person during interview</td>
<td>-0.06</td>
<td>-0.65</td>
<td>-0.23</td>
<td>*</td>
</tr>
<tr>
<td>_cons</td>
<td>-0.97</td>
<td>-2.78</td>
<td>-1.44</td>
<td>-3.16</td>
</tr>
<tr>
<td>Swaziland</td>
<td>-0.28</td>
<td>-2.07</td>
<td>-0.24</td>
<td>-0.67</td>
</tr>
<tr>
<td>Zimbabwe</td>
<td>0.16</td>
<td>0.96</td>
<td>-0.32</td>
<td>-0.81</td>
</tr>
<tr>
<td><strong>Likelihood-ratio test of ln(σ²)=0</strong></td>
<td></td>
<td></td>
<td>11.52</td>
<td>0.76</td>
</tr>
</tbody>
</table>

*Note:* Significance levels: ** p<0.01, * p<0.05, * p<0.1; omitted categories: Respondent's age: 15 years, Highest education level: none, primary, Wealth: Poorest, Occupation: None, Residence: Urban, Religion: Other/none; presence of other person: No; the regression controls for 24 regional dummy variables and interaction terms between wealth and country dummies (coefficients are not reported but available from the author on request)
Figure A1: Proportion of sexually transmitted HIV infections under different scenarios for factors associated with misreporting

Note: Bounds are estimated based on the Probit model coefficients from Table A2; the scenarios differ by the assumption on $\lambda$: **Constant**: $\lambda$ is constant for all $x$, **Age**: $\lambda$ is twice as high if age<17; **Someone present**: $\lambda$ is twice as high if someone is present during the interview; **Religion**: $\lambda$ is twice as high if respondent belongs to the Roman Catholic or Protestant Church.
Figure A2: Proportion of sexually transmitted HIV infections based on different set of control variable

Note: Bounds for PAF are based on the assumption that $\lambda$ is constant for all $x$; set of control variables: Full: as in Table A2, Age+region: Age and regional dummy variables only, Naïve: no control variable.
References


CSO [Central Statistical Office Zimbabwe] and Macro International Inc (2007), Zimbabwe Demographic and Health Survey 2005-06, Calverton, Maryland: CSO and Macro International Inc.

CSO [Central Statistical Office Swaziland], and Macro International Inc (2008), Swaziland Demographic and Health Survey 2006-07, Mbabane, Swaziland: Central Statistical Office and Macro International Inc.


Rutstein SO and Johnson J (2004), The DHS Wealth Index, *DHS Comparative Reports* No. 6, Calverton, MD: ORC Macro.


