DISCUSSION PAPER SERIES

No. 10271

RISKY SEXUAL BEHAVIOR: BIOLOGICAL MARKERS AND SELF-REPORTED DATA

Lucia Corno and Áureo De Paula

DEVELOPMENT ECONOMICS

Centre for Economic Policy Research

ISSN 0265-8003

RISKY SEXUAL BEHAVIOR: BIOLOGICAL MARKERS AND SELF-REPORTED DATA

Lucia Corno and Áureo De Paula

Discussion Paper No. 10271
November 2014
Submitted 19 November 2014

Centre for Economic Policy Research
77 Bastwick Street, London EC1V 3PZ, UK
Tel: (44 20) 7183 8801
www.cepr.org

This Discussion Paper is issued under the auspices of the Centre's research programme in **DEVELOPMENT ECONOMICS**. Any opinions expressed here are those of the author(s) and not those of the Centre for Economic Policy Research. Research disseminated by CEPR may include views on policy, but the Centre itself takes no institutional policy positions.

The Centre for Economic Policy Research was established in 1983 as an educational charity, to promote independent analysis and public discussion of open economies and the relations among them. It is pluralist and non-partisan, bringing economic research to bear on the analysis of medium- and long-run policy questions.

These Discussion Papers often represent preliminary or incomplete work, circulated to encourage discussion and comment. Citation and use of such a paper should take account of its provisional character.

Copyright: Lucia Corno and Áureo De Paula

RISKY SEXUAL BEHAVIOR: BIOLOGICAL MARKERS AND SELF-REPORTED DATA

Abstract

Self-reported data on sexual behaviors have been criticized to be unreliable. In recent studies, risky sexual behaviors have therefore been measured using biomarkers for curable sexually transmitted infections (STIs). Nevertheless, no previous research have tested how reliable such data are. In this paper, we first build an epidemiological model to assess the relative performance of biomarkers versus self-reported data. We then suggest an econometric strategy that combines both types of measures, biomarkers and self-reported data, to improve the estimation of correlates of risky sexual behaviors. Using the Demographic and Health Survey from Zambia, we calibrate the model and provide conditions under which self-reported data are a better proxy for risky sexual behaviors than biomarkers. In countries with low STIs prevalence, the biomarker has a higher probability of misclassification of risky behaviors than self-reported answers. Finally, we apply our estimation strategy to these data.

JEL Classification: C25, I12 and I15
Keywords: biomarker, misclassification, risky behaviour and self-reported

Lucia Corno lucia.corno@phd.unibocconi.it
Queen Mary University of London and Institute of Fiscal Studies

Áureo De Paula a.paula@ucl.ac.uk
University College London, São Paulo School of Economics and CEPR

Risky sexual behavior: biological markers and self-reported data

Lucia Corno, Áureo de Paula*

November 2014

Abstract

Self-reported data on sexual behaviors have been criticized to be unreliable. In recent studies, risky sexual behaviors have therefore been measured using biomarkers for curable sexually transmitted infections (STIs). Nevertheless, no previous research have tested how reliable such data are. In this paper, we first build an epidemiological model to assess the relative performance of biomarkers versus self-reported data. We then suggest an econometric strategy that combines both types of measures, biomarkers and self-reported data, to improve the estimation of correlates of risky sexual behaviors. Using the Demographic and Health Survey from Zambia, we calibrate the model and provide conditions under which self-reported data are a better proxy for risky sexual behaviors than biomarkers. In countries with low STIs prevalence, the biomarker has a higher probability of misclassification of risky behaviors than self-reported answers. Finally, we apply our estimation strategy to these data.

1 Introduction

An estimated 34 million people are living with HIV/AIDS worldwide, and 2.5 million is the number of people who become infected every year (UNAIDS (2012)). Risky sexual behaviors are the main conduit for the spread of the disease and understanding how they change over time is important to design and evaluate potential interventions to address the epidemic. Nevertheless, given that sexual activities are largely private, rigorously measuring those is a difficult task.

Early studies on the HIV/AIDS epidemic mainly rely on self-reported data, such as the frequency of sexual contacts, the use of condoms and the number of partners, as a proxy for risky sexual behavior and derive implications for policies. However, the use of self-reported data has been criticized because

*We thank Orazio Attanasio, Richard Blundell, Erick Gong, Bo Honoré and Imran Rasul for useful comments and seminar participants at the Institute of Fiscal Studies. Correspondence: Lucia Corno - Queen Mary, University of London and Institute of Fiscal Studies. E-mail: l.corno@qmul.ac.uk; Aureo de Paula - University College London, São Paulo School of Economics-FGV, CeMMAP, and Institute of Fiscal Studies. E-mail: a.paula@ucl.ac.uk. De Paula acknowledges financial support from National Science Foundation through award SES-1123990, the European Research Council through Starting Grant 338187 and the Economic and Social Research Council through the ESRC Centre for Microdata Methods and Practice grant RES-589-28-0001.

respondents may misreport their activities or give socially desirable answers, especially in contexts where their responses are linked to some rewards or incentives (Palen, Smith, Caldwell, and Flisher (2008); De Paula, Shapira, and Todd (2014); Fenton, Johnson, McManus, and Erens (2001); Turner, Ku, Rogers, Pleck, and et al. (1998); Ozler (2013)).[1,2] More recent works have focused on collecting biological markers (biomarkers) on the incidence of curable sexually transmitted infections (STIs) (e.g., chlamydia, gonorrhea, syphilis) as objective measures of risky sexual behaviors, providing some evidence that self-reported answers might indeed underestimate risky sexual activity (Minnis, Steiner, Gallo, and et al. (2009); Gregson, Zhuwau, and Ndlovu (2002); Gallo, Behets, Steiner, and et al. (2006), Zenilman, Weisman, and Rompalo (1995), Mauck and Straten (2008), Cleland, Boerma, and Weir (2004)). Biomarkers on curable STIs are therefore quickly becoming a popular method to measure risky sexual behaviors in large-scale behavioral change interventions aimed at promoting safer sexual practices and reducing HIV prevalence (Gong (forthcoming), de Walque, Dow, Nathan, and et al. (2012)).[3] As with other indirect records of sexual behavior, biomarkers themselves are not perfect: individuals who adopt risky sexual behaviors may remain uninfected and hence be misclassified as not having behaved in a risky manner. However, to our knowledge, no previous research have so far evaluated the reliability of biomarkers: whether risky sexual behaviors are better measured by self-reported data or biomarkers or a combination of both remains an open question.

In this paper, we first illustrate a variant of standard epidemiological models to characterize the probability of misclassification of risky sexual behaviors with biomarkers. Risky sexual behaviors are misclassified by biomarkers when they are not detected by the presence of a sexually transmitted disease (i.e., being tested STI negative but having behaved in a risky manner). We use the model to compare the misclassification rate of biomarkers with self-reported data on risky sexual behavior. The model builds upon the work of Hyman, Li, and Stanley (2001), where the probability of transmission of a sexually transmitted disease from an infected to an uninfected person depends on the proportion of

[1]To address the issue of misreporting risky behavior during face-to-face interviews, audio computer assisted self-interviewing (ACASI) techniques have been used to collect data on sexual behavior in the United States (Tourangeau and Smith (1996); Hewitt (2002)). Some randomized evaluations have found that there is a higher probability of reporting risky sexual behavior with computerized mode of interviews than with face-to-face interviews mode (Hewett, Mensch, and Ribeiro (2008)). A computerized interviews method would be however difficult to implement in developing countries, where HIV/AIDS is widely spread. Gregson, Zhuwau, and Ndlovu (2002) described an informal, confidential and low-technology method, called Informal Confidential Voting Interview, to collect data on sexual behavior in low income countries.

[2]The issue of social desirability bias in self-reported answers has been discussed in other settings. For example, Baird and Ozler (2010), using data from a randomized Cash Transfer Program in Malawi, compare self-reported data on school attendance with administrative records. They show that participants significantly overstate their school participation and this overreporting is higher in the control group, thus producing biased impact estimates. In a more recent paper, Karlan and Zinman (2012) test the validity of self-reported data on loan expenditure, for consumption or for investment purpose. They find that respondents were more likely to admit using their loan for household items and medical/educational expenses on an anonymous survey than they were in response to direct questioning.

[3]The use of biomarkers on curable STIs as outcome measure is motivated by the fact that data on HIV incidence, although more appropriate, may be expensive to collect, given the sample size required to detect any effect of HIV/AIDS interventions (Fishbein and Pequegnat (2000)).

infected people in the population, the number of partners, the number of sexual contacts per partner, the probability of infection from an infected partner and, finally, on the likelihood to meet an infected partner. By using STIs as a proxy for risky sexual behaviors, individuals who become infected in a given period are tagged as having behaved in a risky manner. The probability of correct classification using biomarkers for those who engage in risky sexual behavior is therefore equal to the probability of disease transmission from an infected to an uninfected individual. On the other hand, the probability of correct classification using elicited sexual behaviors is defined as the likelihood of truthful elicitation in a survey. Since misclassification is possible in both cases, it is not ex-ante clear that either one is a superior marker for risky behavior. The model provides the conditions for when the probability of correct classification is higher using self-reported data or biomarkers for STI and vice-versa. In general, we find that the biomarkers have a higher probability of misclassification than behaviors elicited by a survey questionnaire in populations with a lower STI prevalence. The intuition behind these results is straightforward: if a sexually transmitted disease is common in the population, the probability that an individual who engaged in risky sexual intercourse will be infected is higher, compared to settings where the same STI is less common. Thus, if the probability of infection is higher, the probability of detecting the infection with biomarkers is also higher. This finding implies that exclusive reliance on biomarkers for STI or self-reported data can lead to a biased measure of risky sexual behavior. This result can be a very useful tool for researchers to infer the best proxy for risky sexual behavior among various biomarkers and/or self-reported questions on sexual activities in a given country.

We then suggest an econometric strategy that combines both types of measures - biomarkers and self-reported data - to improve the estimation of correlates of risky sexual behaviors. Our strategy uses the information from biomarkers for curable STIs (i.e., STI positive or STI negative) to estimate the probability of correct classification with self-reported data. Building on previous work by Hausman, Abrevaya, and Scott-Morton (1998), we combine this information to provide a consistent estimator for the parameters of a binary outcome econometric model for risky behavior. The estimator can be easily computed by Generalized Method of Moments (GMM), using commercially available packages. We believe that variations of this strategy can be employed to estimate similar models with more general outcomes and objective functions.

Finally, using data on STI prevalence and self-reported sexual behaviours from the Demographic and Health Survey collected in Zambia in 2007, we calibrate the parameters of the epidemiological model and show the conditions under which self-reported data are a better proxy for risky sexual behaviors than biomarkers and vice versa. We then estimate the association between several individual characteristics and (true) risky sexual behavior using our proposed empirical strategy.

The remainder of the paper is organized as follows: section 2 provides an epidemiological model to compare the probability of misclassification of risky sexual behavior using biomarkers for STI and self-reported data; section 3 combines the two in a new estimator; section 4 describes the data used for simulating the model and in section 5 we show the results. Section 6 concludes and discusses the implications of our findings.

2 An Epidemiological Model for Biomarkers Misclassification

In this section, we present a model of misclassification of risky sexual behaviors using biological markers for curable sexually transmitted infections (STIs). Risky sexual behaviors are misclassified when they are not detected if measured by the presence of an STI. We use the model to draw comparisons with the probability of self-reporting information on risky sexual behaviors in a survey and generate suggestions to improve inferences on such behaviors. Implicitly, we define a risky sexual contact as one that allows for the transmission of the STI. Protected sexual contacts that preclude the transmission of the sexually transmitted infection (STI) (by, for instance, using condoms) are not classified as risky behavior in the model.[4]

Our theoretical framework is built upon epidemiological models of disease dynamics (see for example Anderson and May (1991), Hethcote (2000), Hyman, Li, and Stanley (2001)). The key output of the model is the transmission rate of the disease, λ : the rate at which uninfected individuals are infected by infected partners.[5] We assume that the transmission rate is homogeneous across different infected individuals at different stages of the disease. We suppose that the (annual) transmission rate λ depends on the number of partners per individual (p), the total share of infected individuals in the population (I), the average number of sexual contacts per partner ($c(p)$), and the probability of infection by an infected partner ($\beta(p)$). Following Hyman, Li, and Stanley (2001), we model the average number of sexual contacts with each partner as a decreasing function in the number of partners per year and equal to:

$$c(p) = 104p^{-\eta} + 1, \tag{1}$$

[4] If a non-risky sexual behavior (e.g., the proper use of a condom during sexual intercourse) significantly reduces the the probability of transmission of the STIs but nonetheless still allows it, another type of misclassification would arise: being tagged as risky when the behavior is non-risky. This would be the case, for example, of genital herpes which can be transmitted when outbreaks occur in areas that are not protected by the condom but still under contact in the sexual act (Centers for Disease Control and Prevention (2014)).

[5] The rate λ is a measure of risky sexual behavior over the reference period of interest (e.g., a year) and is therefore seen as an incidence rate. The incidence rate for curable STIs with short cycles in a given population is usually close to the prevalence rate for curable STIs (World Health Organization (2008)). Other measurements, such as the prevalence rate for non-curable STIs (e.g., HSV2), may be used to assess risky behavior over a longer horizon and as a lasting marker for past risky behavior (Baird, Gong, McIntosh, and Ozler (2014)). Whereas we focus the exposition on the incidence of curable STIs, our results can be adapted replacing our incidence measure λ with the modeled or measured prevalence rate.

where η is a positive parameter that controls how fast the number of sexual contacts decreases with the number of partners. As in Hyman, Li, and Stanley (2001), we set η equal to one. With one partner per year ($p = 1$), the above functional form reasonably implies 105 sexual intercourses in a year: roughly two per week. Using data from the Malawi Diffusion and Ideational Change Project for 2004, for example, about one-third of unmarried female respondents report having sex at least twice a week. Table 1 reports the figures for unmarried females in Malawi.

[Insert Table 1]

With 20 partners a year, for example, the number of sexual contacts per partner would be 6.2 per year. So, as the number of partners increases, the number of encounters per partner asymptotes to one. If the number of contacts per partner is not available from the data, one can use equation (1) to bound the probability of misclassification since in the case above, $1 \leq c(p) \leq 105$.[6] Note that alternative functional forms may also be used to calibrate the number of sexual contacts per partner and could incorporate other observables (whenever those are available) such as length of partnership.

Following Hyman, Li, and Stanley (2001), if someone has p partners and a given partner is infected, the probability of infection from that partner, $\beta(p)$, depends on the average number of sexual contacts with a given partner, $c(p)$, and is given by the probability that the disease is transmitted in at least one sexual encounter:

$$\beta(p) = 1 - (1 - \xi)^{c(p)}, \tag{2}$$

where ξ is the probability of transmission from a single contact with an infected person and $(1-\xi)^{c(p)}$ is the probability that an individual will avoid infection when she has $c(p)$ contacts with an infected partner.

Lastly, we adapt the model by assuming that infected and uninfected individuals may have a different numbers of sexual partners (p_I and p_U, respectively). We model the probability that an uninfected person meets an infected partner as

$$P_{UI} = \frac{\rho p_I I}{\rho p_I I + p_U(1 - I)}, \tag{3}$$

where $\rho \geq 0$ controls the degree of sorting between infected and uninfected individuals. When $\rho = 1$, the person meets every individual with equal and independent probability regardless of their infection status. There is perfect sorting when $\rho = 0$: an uninfected individual only meets other uninfected individual. Finally, when $\rho \to \infty$, uninfected individuals only meet infected ones.[7]

[6] Indeed, standard national representative surveys eliciting information on sexual behaviors, such as the Demographic Health Surveys (DHS), do not include questions on the number of sexual intercourse per partner or per month.

[7] Since our focus is on the misclassification of risky sexual behavior for uninfected individuals, we abstract away from

The probability that an uninfected individual becomes infected during a given year is therefore:

$$\lambda = \left[1 - \left(1 - \beta(p_U)\frac{\rho p_I I}{\rho p_I I + p_U(1-I)}\right)^{p_U}\right]. \tag{4}$$

This probability is equal to the probability that the disease is transmitted by at least one of the sexual partners, which is obtained as one minus the probability that it is not transmitted by any of them. The probability that the disease is transmitted by each one of the sexual partners is given by the probability that this person meets an infected partner, P_{UI}, and the disease is transmitted by that partner, $\beta(p_U)$. Since there are p_U sexual partners, this is given by $\beta(p_U)\frac{\rho p_I I}{\rho p_I I + p_U(1-I)}$. One minus $\beta(p_U)\frac{\rho p_I I}{\rho p_I I + p_U(1-I)}$ equals the probability that the disease is not transmitted. Raising this term to the power of p_U gives the probability that the disease is not transmitted by any sexual partner. λ is the probability of the complementary event.

If $p_I = p_U = p$, λ is equal to

$$\lambda = \left[1 - \left(1 - \beta(p)\frac{\rho I}{\rho I + (1-I)}\right)^{p}\right]. \tag{5}$$

A limitation of equations (4) and (5) is that, since generally self-reported, one may only have imperfect data on the number of sexual partners. If individuals underreport the number of sexual partners, the calibrated value of λ will be larger than if the value were obtained with truthful reports. To see that, assume for simplicity that $p_U = p_I = p$ and that only a fraction of p is reported, say kp (where $0 < k < 1$). This would lead to a higher number of calibrated contacts per partner, $c(kp)$, and consequently a higher probability that the disease is transmitted in at least one encounter with one of those partners (from equation (2)). Because P_{UI} (from equation (3)) remains at the same value, this in turn leads to a lower probability that the disease is not transmitted with any of the person's partners (i.e., the expression in parenthesis in equation (4)). Hence, the inferred transmission rate λ would be higher if compared to the accurately calibrated one. An analogous argument delivers that, when individuals overreport the number of sexual partners, λ is lower than if the number of partners is truthfully reported.

Of course, when panel data are available, it is possible to estimate the true value of λ, by computing the rate at which individuals, who were previously STI negative, contract the infection a year later.

2.1 Comparing Biomarkers and Self-Reported Measurements

Whereas biological markers can register whether an individual adopted a risky sexual behaviors or not (i.e., the "extensive margin" of such behaviors), they are less informative about its intensity (i.e.,

the equilibrium characterization of the matching process.

the "intensive margin"). Then the presence of an STI would not provide any information about the intensity of the behavior leading to the acquisition of the STI. Hence, in our comparison between biological markers and elicited measures of risky behaviors, it seems adequate to encode those into a binary variable.

Let Y^t indicate whether an individual actually engaged in risky behavior ($Y^t = 1$) or not ($Y^t = 0$). Denote by Y^e the variable indicating whether elicited behavior (for example, in a survey) is reported to be risky ($Y^e = 1$) or not ($Y^e = 0$). Finally, α denotes the probability of correct classification (i.e., marking someone who engaged in risky behavior as having behaved in a risky way) using elicited sexual behavior: $\alpha = \mathbb{P}(Y^e = 1|Y^t = 1)$. Here, we assume (realistically) that people have no incentives to report risky sexual behaviors if they did not engage in such behaviors (i.e., $\mathbb{P}(Y^e = 0|Y^t = 0) = 1$ and, consequently, $\mathbb{P}(Y^e = 1|Y^t = 0) = 0$).

On the other hand, by using biomarkers for STIs as a proxy for risky sexual behaviours, all those who become infected in a given period are tagged as having behaved in a risky manner.[8] The probability of correct classification using biomarkers for those who engage in risky sexual behavior is then given by $\lambda \in [0, 1]$. We also assume that those who do not engage in risky sexual behavior are not infected (though see footnote 4). This means that the probability of correct classification using biomarkers for those who do not behave in a risky manner is one.

Since misclassification is possible in both cases, it is not ex-ante clear that either one is a superior marker for risky behavior. For example, when the infection rate is low enough (either because of low prevalence or because of low transmission rates), the biomarkers would misclassify risky behavior more often. This is formalized in the following result:

Proposition 1 *If $\lambda < \mathbb{P}(Y^e = 1)$, the biomarker has a higher probability of misclassification of risky behavior than behavior elicited by the survey questionnaire.*

This result is easily established by noting that $\lambda\mathbb{P}(Y^t = 1) \leq \lambda$ and $\mathbb{P}(Y^e = 1) = \mathbb{P}(Y^e = 1|Y^t = 1) \times \mathbb{P}(Y^t = 1) + \mathbb{P}(Y^e = 1|Y^t = 0) \times \mathbb{P}(Y^t = 0) = \mathbb{P}(Y^e = 1|Y^t = 1)\mathbb{P}(Y^t = 1) \equiv \alpha\mathbb{P}(Y^t = 1)$. If $\lambda < \mathbb{P}(Y^e = 1)$, we got that $\lambda < \alpha\mathbb{P}(Y^t = 1)$. This implies that $\lambda\mathbb{P}(Y^t = 1) < \alpha\mathbb{P}(Y^t = 1)$ and thus $\lambda < \alpha$.

An interesting aspect of this proposition is that we can compare data on elicited sexual behavior with reasonable values for λ to establish whether the above inequality holds. Note that Proposition 1 holds even when individuals underreport the number of partners since in this case, the calibrated λ will be higher than the one computed using the accurate number of partners.

[8] Here, we implicitly assume that all the biological markers for sexually transmitted infections are able to detect the disease in infected individuals. If "false negatives" are possible, the reliability of biomarkers would be further compromised.

In the above result, a maintained assumption is that $\mathbb{P}(Y^e = 1|Y^t = 0) = 0$. This appears to us as a realistic assumption, but circumstances where individuals misreport a non-risky behavior can be ventilated. Whereas Proposition (1) would not accommodate such environments, our econometric strategy below can be adapted to such circumstances.

3 An Econometric Model Combining Measurements

The epidemiological model developed in the previous section allows us to establish the best marker for risky sexual behavior (see Proposition (1)). We now describe how the combination of biomarkers and self-reported data can be used to address the misclassification issue and more precisely estimate risky sexual behaviors. In particular, we can combine biomarkers and self-reporting data to improve inference on the determinants of risky behavior, even when the biomarker alone is a superior measurement. Our goal is then to estimate a set of parameters θ, which characterizes the correlates of risky sexual behaviors. Our estimation strategy builds on previous work by Hausman, Abrevaya, and Scott-Morton (1998).

We start by setting the relationship between risky sexual behaviors and observable covariates of interest (i.e., the determinants of risky sexual behaviors). In particular, we assume that the relationship between risky sexual behaviour and observable covariates can be summarized by the following conditional probability specification:

$$\mathbb{P}(Y^t = 1|\mathbf{X}) = F(\mathbf{X};\theta), \tag{6}$$

where $\mathbf{X}$ are the covariates of interest. Here, $F(\cdot)$ is known up to parameters θ and differentiable in θ (e.g., $F(\mathbf{X};\theta) = \Phi(\mathbf{X}^\top\theta)$ if the model corresponds to a Probit, $F(\mathbf{X};\theta) = \Lambda(\mathbf{X}^\top\theta)$ if the model corresponds to a Logit or $F(\mathbf{X};\theta) = \mathbf{X}^\top\theta$ if the model corresponds to a Linear Probability Model). Then, the probability of reporting risky behaviour given $\mathbf{X}$ is equal to

$$\mathbb{P}(Y^e = 1|\mathbf{X}) = \mathbb{P}(Y^e = 1|Y^t = 1, \mathbf{X}) \times \mathbb{P}(Y^t = 1|\mathbf{X}) + \mathbb{P}(Y^e = 1|Y^t = 0, \mathbf{X}) \times \mathbb{P}(Y^t = 0|\mathbf{X}).$$

Assume that $\mathbb{P}(Y^e = 1|Y^t = 0, \mathbf{X}) = 0$ and, for simplicity, that $\mathbb{P}(Y^e = 1|Y^t = 1, \mathbf{X}) = \mathbb{P}(Y^e = 1|Y^t = 1)$. Using (6), we get:

$$\mathbb{P}(Y^e = 1|\mathbf{X}) = \mathbb{P}(Y^e = 1|Y^t = 1)F(\mathbf{X};\theta) = \alpha F(\mathbf{X};\theta). \tag{7}$$

The above relationship implies that a model using elicited behavior alone as a proxy for risky behavior would lead to bias in the estimation of θ (see Hausman, Abrevaya, and Scott-Morton (1998)). Similar

arguments establish that using biomarkers alone would also lead to bias in the estimation of θ.[9]

For example, in a linear probability model where $F(\mathbf{X};\theta) = \mathbf{X}^\top\theta$, the probability that $Y^e = 1$ given $\mathbf{X}$ is

$$\mathbb{P}(Y^e = 1|\mathbf{X}) = \mathbb{P}(Y^e = 1|Y^t = 1)\mathbf{X}^\top\theta = \alpha\theta^\top\mathbf{X}$$

and a linear probability model would estimate $\alpha\theta$, which is smaller in absolute value than θ (since $\alpha < 1$). Similarly, with biomarkers:

$$\mathbb{P}(Y^{STI} = 1|\mathbf{X}) = \mathbb{P}(Y^{STI} = 1|Y^t = 1)\mathbf{X}^\top\theta = \lambda\theta^\top\mathbf{X}$$

where $\|\lambda\theta\| < \|\theta\|$). Consequently, estimation of θ based solely on elicited behaviour or biomarkers will suffer from attenuation bias (see Hausman, Abrevaya, and Scott-Morton (1998))).

Assume now that the probability of correct classification with elicited sexual behaviors can be inferred from the proportion of individuals who admit to having engaged in risky behaviors among those tested positive for the STI:

$$P(Y^e = 1|Y^t = 1) = P(Y^e = 1|Y^{STI} = 1). \tag{8}$$

Substituting equation (8) into (7), we obtain

$$\mathbb{P}(Y^e = 1|\mathbf{X}) = \mathbb{P}(Y^e = 1|Y^{STI} = 1)F(\mathbf{X};\theta).$$

Since both $\mathbb{P}(Y^e = 1|\mathbf{X})$ and $\mathbb{P}(Y^e = 1|Y^{STI} = 1)$ are estimable, the above relationship can be used to consistently estimate θ. We also note that the probability of misclassification by self-reported risky behaviour can be made dependent on covariates, yielding a straightforward generalization of the equations above.

Our approach is based on the following regression model:

$$Y^e = \mathbb{P}(Y^e = 1|Y^{STI} = 1)F(\mathbf{X};\theta) + u, \tag{9}$$

and uses the moment condition:

$$\mathbb{E}[Y^e - \mathbb{P}(Y^e = 1|Y^{STI} = 1)F(\mathbf{X};\theta)|\mathbf{X}] = 0.$$

[9]If misreporting of non-risky behavior is possible, the expression in (7) becomes $P(Y^e = 1|X) = \beta + (\alpha - \beta)F(X;\theta)$, where $\beta = P(Y^e = 1|Y^t = 0)$. As pointed out previously, this can be accommodated in the estimation strategy enunciated below by having β as an additional parameter to be estimated (this identification strategy is analogous to the one in Hausman, Abrevaya, and Scott-Morton (1998)).

Let $\hat{\alpha}$ be the estimator for $\mathbb{P}(Y^e = 1|Y^{STI} = 1)$ based on the frequency of individuals for whom $Y^e = 1$ among STI positive individuals. Then, one estimator for θ is the Nonlinear Least Squares Estimator, defined by the minimizer of the quadratic function

$$Q_N(\theta) \equiv \sum_{i=1}^{N} \left[y_i^e - \hat{\alpha} F(\mathbf{x}_i; \theta) \right]^2,$$

where i indexes the individual observations and N is the sample size. The estimator can be framed as a GMM estimator and computed in standard packages (e.g., STATA). Standard textbook arguments deliver consistency and asymptotic normality for the above estimator. We summarize those results in the following proposition:

Proposition 2 *Under random sampling, assuming that the parameter space for θ is compact, $\mathbb{E}[\sup_\theta \|\partial_\theta F(\mathbf{X};\theta)\|] < \infty$ and $\mathbb{E}[\sup_\theta \|\partial^2_{\theta\theta^\top} F(\mathbf{X};\theta)\|] < \infty$, the estimator $\hat{\theta} \equiv \arg\min_\theta Q_n(\theta)$ is ($\sqrt{n}$-) consistent for θ and asymptotically normal:*

$$\sqrt{n}\left(\begin{bmatrix} \hat{\theta} \\ \hat{\alpha} \end{bmatrix} - \begin{bmatrix} \theta_0 \\ \alpha_0 \end{bmatrix} \right) \longrightarrow_d \mathcal{N}(0, \mathbf{G}^{-1}\mathbf{S}\mathbf{G}^{\top -1}) \tag{10}$$

where $\mathbf{G}$ *and* $\mathbf{S}$ *are given by*

$$\mathbf{G} = \mathbb{E}\begin{bmatrix} \alpha \partial_\theta F(\mathbf{X};\theta)\partial_\theta F(\mathbf{X};\theta)^\top & F(\mathbf{X};\theta)\partial_\theta F(\mathbf{X};\theta) \\ 0 & Y^{STI} \end{bmatrix}$$

and

$$\mathbf{S} = \mathbb{E}\begin{bmatrix} (\alpha F(\mathbf{X};\theta) - Y^e)^2 \partial_\theta F(\mathbf{X};\theta)\partial_\theta F(\mathbf{X};\theta)^\top & (\alpha F(\mathbf{X};\theta) - Y^e) Y^{STI} (\alpha - Y^e)\partial_\theta F(\mathbf{X};\theta) \\ (\alpha F(\mathbf{X};\theta) - Y^e) Y^{STI} (\alpha - Y^e)\partial_\theta F(\mathbf{X};\theta)^\top & Y^{STI}(\alpha - Y^e)^2 \end{bmatrix}$$

Proof. The estimator can be cast as the GMM estimator minimizing $\| \sum_{i=1}^N g(\mathbf{z}_i, (\theta, \alpha))_{\dim(\theta)+1\times 1} \|^2$ where $g(\mathbf{Z}_i, (\theta, \alpha)) = [(\alpha F(\mathbf{X};\theta) - Y^e)\partial_\theta F(\mathbf{X};\theta) \quad Y^{STI}(\alpha - Y^e)]^\top$. Using standard textbook results (see, for example, Hayashi (2001)), this estimator is consistent and asymptotically normal with displayed asymptotic variance where $\mathbf{G} = \mathbb{E}[\partial_{(\theta,\alpha)} g(\mathbf{Z}_i, (\theta, \alpha)]$ and $\mathbf{S} = \text{var}(g(\mathbf{Z}_i, (\theta, \alpha))$. The dominance condition $\mathbb{E}[\sup_\theta \|\partial_\theta F(\mathbf{X};\theta)\|] < \infty$ guarantees the dominance condition for consistency (see Hayashi, Prop. 7.7) and both dominance conditions guarantee the dominance condition for asymptotic normality (see Prop. 7.10 in Hayashi).

For a linear probability model, the estimator for $\hat{\theta}$ can be simply computed as the ratio between the OLS estimates, where the dependent variable is Y^e, and $\hat{\alpha}$, the proportion of people reporting

risky sex among the STI positive respondents. In the linear model, standard errors need nevertheless to be adjusted as indicated in expression (10). More generally, this estimator can be estimated in most statistical packages (using, for example, the command `gmm` in STATA).

4 Data and Descriptive Statistics

We use data from the Demographic and Health Surveys (DHS) conducted in Zambia in 2007 to calibrate the epidemiological model described in section 2 and to estimate the correlates of risky sexual behaviors using the Generalized Method of Moments reported in section 3.

The DHS are nationally representative and data are collected using a two-stage sampling design (Zambia Central Statistical Office (2009)). In the first stage, a sample of clusters is selected from a list of enumeration areas from the latest national census. In the second stage, a complete list of households is created in each cluster. In each randomly selected household, all women aged 15-49 and all men aged 15-59 who were either permanent residents or visitors present in the household on the night before the survey were eligible to be interviewed.

Crucially for our purposes, the Zambia DHS, besides eliciting detailed information on the respondents' sexual behaviour, includes data on biomarkers for STIs: all women and men eligible to be interviewed were asked to voluntarily provide a blood sample for HIV and syphilis testing in order to determine national prevalence rates. The STIs tests were conducted after the survey. For this reason, the dataset is particularly suitable to test the predictions of the epidemiological model. Our final sample with non-missing information for self-reported sexual behaviors and biological marker for syphilis includes 2,414 individuals, 52.5% of women and 47.5% of men.

[Insert Table 2]

Table 2 reports the characteristics of our sample. In Panel A, we report sociodemographic characteristics. The average age of the individual in the sample is 29 years old and approximately 60% of them are currently married. Looking at the highest educational level, 7.3% of the respondents did not have any formal education, almost half of them have primary education and about 43% have achieved secondary education or a college degree. Panel B reports the fraction of individuals tested positive for HIV and syphilis. The prevalence of syphilis is equal to 4.3% and it is very similar across genders. Nearly 15% of the respondents have been tested positive for HIV, 17% among women and about 13% among men. Looking at self-reported data on sexual behavior (Panel C), 17.5% of the respondents reported to have used a condom during their last intercourse and the average number of sexual partners in the last 12 months is approximately equal to 0.87 (including those who never had a partner). Among the subsample of married or cohabiting respondents, the average number of

sexual partners different from spouses is 0.096 and 1.8% of them declare that the last sexual intercourse was not with their spouse/cohabiting partner. We next construct two aggregate indicators of risky sexual behavior. "Risky sex 1" is a indicator equal to 1 if respondent did not use a condom during his/her last intercourse or if a married respondent declared that his/her last intercourse was not with his/her spouse/cohabiting partner and zero otherwise. Given that not using a condom during sexual intercourse with the spouse is often not considered a risky behavior, we restrict our definition further. "Risky sex 2" is equal to 1 if an unmarried respondent reported that a condom was not used the last time he/she had sexual intercourse or a married respondent reported that his/her last sexual intercourse was not with spouse/cohabiting partner and no condom was used, and equal to zero if an unmarried respondent reported using a condom during his/her last intercourse or a married respondent reported not having had extramarital sex or extramarital sex with condom.

[Insert Table 3]

In Table 3, we show the share of respondents reporting risky sexual behaviors by syphilis status to investigate any sort of relationship between self-reported behavior and biomarkers data. The fraction of individuals reporting risky sexual behaviors is higher among those tested positive for syphilis. In particular, a lower fraction of respondents reported to have used a condom in their last intercourse among those tested positive for syphilis (15.7%) compared to those who declared to have used a condom in their last intercourse among the STI negative respondents (17.6%) and the difference is not statistically significant (p-value equal to 0.657). Looking at the number of partners in the last 12 months, we observe a statistically significant difference among the number of partners declared by STI positive respondents (1.07 partners) compared to the number reported by STI negative individuals (0.86 partners) while the difference in the number of extramarital partners between STI positive and negative individuals is not statistically significant. Approximately 5.3% of syphilis positive married individuals reported that their last sexual intercourse was not with spouse compared to about 1.4% of the STI negative respondent (p-value equal to 0.008). By looking at the aggregate indicators for risky sexual behaviors - "Risky sex 1" - it is interesting to note that only 14.3% of the respondents tested positive for syphilis reported risky behaviors, suggesting that underreporting can be in place. There is no difference in the fraction of individuals reporting risky sexual behavior between syphilis positive and syphilis negative respondents either, as indicated by the aggregate variable "Risky sex 2".

We should note that if the horizon over which the risky behavior is elicited, for example last 12 months, is longer than the cycle of the STI, the STI test at the end of the questionnaire may not be able to detect the risky behavior. If this is the case, the individual may correctly report risky behavior in the last 12 month but be tested negative. This happens not because biomarkers missclassify

risky sex, but because the STI might have spontaneously resolved prior to the test. In the case of syphilis, for example, the positive individual will usually heal within 1-2 months of onset whereas risky behavior measures are usually defined over a longer spell. In practice, this will only exacerbate the mismeasurement by biomarkers (it also highlights the need to carefully define the horizon over which the risky behavior is elicited by the survey or revealed by a biological marker).

5 Results

5.1 Comparing Biomarkers and Self-Reported Measurements

We now use the Zambia DHS to fix a set of parameters to simulate the model described in section 2. More precisely, we simulate λ, the probability of correct classification of risky sexual behaviors using biomarkers with the parameters' values reported in Table 4.

[Insert Table 4]

Recall that ξ is the transmission probability of syphilis from an infected to an uninfected individual during a given unprotected sexual contact. Following Hyman, Li, and Stanley (2001), we set η, the parameter that controls for how the number of contacts varies with number of partners, equal to one. Finally, we also set $\rho = 1$, allowing for a person to meet every individual with an equal and independent probability regardless of their infection status.

Next, we apply Proposition 1 by comparing the simulated value of λ with self-reported risky behaviors, measured by the aggregate variables "Risky sex 1" and "Risky sex 2" reported in table 2.

[Insert Figure 1]

Figure 1 shows the theoretical combinations of I - the STI prevalence rate in a given country in a given year - and λ, the probability of correct classification of risky sexual behaviours using biomarkers. The vertical dotted line indicates the value of syphilis prevalence, equal to 0.043, used to simulate λ. When $I = 0.043$, $p = 1.14$, $\rho = 1$ and $\eta = 1$, the simulated λ is equal to 0.047. From figure 1, a positive relationship emerges between the probability to correctly measure risky behaviors with biomarkers (λ) and STIs prevalence in a given country. Hence, countries with a high STI prevalence are more likely to get a good proxy of risky sexual behaviors using biomarkers compared to countries where the STI prevalence is low. The intuition behind this result is very simple: if a sexually transmitted disease is common in the population, the probability that an individual who engaged in risky sexual intercourse will be infected is higher, compared to settings where the same STI is less common. Thus, if the probability of infection is higher, the probability of detecting the infection with biomarkers is also higher.

[Insert Figure 2]

Given that the number of partners used to estimate λ is self-reported, in figure 2, we plot again the relationship between λ and I but for two different values of p. In particular, we consider the number of partners declared by married women, who are more likely to underreport them, as a lower bound, and the number of partners mentioned by single men, who are more likely to overreport them, as an upper bound for p (Oster, 2004).[10] The average number of partners in the last 12 months for single men is equal to 1.32, while the average number of partners for married women is equal to 1.005. We then compared the estimated value of λ with the horizontal lines in the graph, that indicate the fraction of people reporting risky sexual behaviors $\mathbb{P}(Y^e = 1)$ in the DHS, using the aggregate measures of risky sex reported in table 2. We show that as the number of reported sexual partners increases for a given I, the probability of correct classification using biomarkers also increases. Thus, in countries with low STI prevalence and low average number of sexual partners per individual, λ is lower than $\mathbb{P}(Y^e = 1)$ and the biomarker has a higher probability of misclassification of risky behaviors than behaviors elicited by a survey questionnaire. On the other hand, in countries where both STIs prevalence and the number of sexual partners per individual are high, λ is greater than $\mathbb{P}(Y^e = 1)$, meaning that the biomarkers have a lower probability of misclassification of risky behaviours than behaviours elicited by survey questionnaire. In other words, the intersection between the horizontal line indicating self-reported behavior and the upward sloping curve describing the relationship between λ and I provides the threshold above which biological markers and below which self-reported answers are a better proxy for risky sexual behaviors, respectively. For values of I above this threshold, biomarkers are have a smaller probability of misclassification compared to self-reported data whereas for values of I below the threshold self-reported data are more accurate. Figure 2 shows that, in the case of Zambia, where syphilis prevalence is equal to 4.3%, if the person meets every individual with equal and independent probability regardless of their infection status ($\rho = 1$), biomarkers have a lower probability to correctly measure risky sexual behaviors than self-reported responses. Positive assortativeness ($\rho < 1$), a totally plausible scenario, leads to an even more favorable picture for elicited behavior.[11]

These results can be an important tool to find the best measure of risky behaviors in interventions or programmes aiming at reducing HIV prevalence in a given country. Let's suppose that researchers need to understand the best measure of risky behavior in a country where syphilis prevalence is equal to 10%, the number of partners is 1.3 per year, and the self-reported risky indicator suggests that

[10] As indicated in Section 2, because λ is overestimated when individuals underreport the number of sexual partners and underestimated when individuals overreport them, these bounds would be wider if those reports are not truthful.

[11] For example, Dow and Philipson (1993) find that HIV positive individuals in San Francisco are twice more likely to have an HIV positive partner than an HIV negative one.

the share of respondents having risky sex is equal to 1%. Looking again at figure 2, in this particular scenario, the probability of misclassification of risky behavior using self-reported questions is higher than the one using biomarkers.

5.2 Combining Biomarkers and Self-Reported Measurements

[Insert Table 5]

Table 5 compares the parameters on the determinants of risky sexual behaviors using as dependent variable self-reported indicators for risky sex (columns 1-3), the biological marker for syphilis (columns 4-6) and the probability of true risky sexual behaviour estimated using the GMM procedure described in section 3 (columns 7-9). For each dependent variable, we estimate a linear probability model, a probit and a logit model.

As emerged already from table 2, the probability of correct elicitation among those that engage in risky behavior (α) is estimated at 14.3%. We note that even in cases when this rate is less favorable compared to the probability of correct classification by the biological markers, the combination of both sources allows one to more accurately estimate the correlates of risky sexual behavior.

By comparing the determinants of risky sexual behavior using self-reported questions and biomarkers, we note a huge discrepancy both in term of magnitude and sign of the coefficients. For example, being a woman is negatively and significantly correlated with the probability of reporting risky sex (columns 1-3), but is positively associated with the probability of being tested positive for syphilis (although the coefficient is not statistically significant) (columns 4-6). Looking at the GMM specifications in columns 7-9, the coefficient on female is negatively and significantly correlated with the dependent variable but it is statistically significant only in the OLS specification. As expected, the magnitude of the coefficients estimated with the GMM is lower compared to those estimated in columns 1-3. Age is negatively and statistically significant correlated with the probability of reporting risky sexual behavior (columns 1-3) and with the likelihood of true risky sexual behaviours (columns 7-9), suggesting that older people are less likely to behave in a risky manner. On the other hand, the coefficient on age turns out to be positive when looking at the probability of being tested positive for syphilis. Both primary and secondary education are negatively correlated with the dependent variables (although statistically significant only for the probability of reporting risky sex), suggesting that more educated individuals are less likely to behave risky. The coefficients on urban areas are not statistically significant.

[Insert Table 6]

Using the GMM coefficients estimated in the previous table, in table 6, we compute the probabilities of true risky sexual behavior for individuals with different sociodemographic characteristics. For example, a 30-year-old men, with secondary education and who lives in an urban area has a 96% chance to adopt risky sexual behaviors. On the other hand, a 40-year-old woman with primary education, living in a rural area, has only 16% probability of behaving in a risky manner.

[Insert Table 7]

In table 7, we report summary statistics for the predicted probability of self-reported risky behavior and of true risky behavior estimated with the GMM. The probability of true risky behavior is both higher and more disperse than the predicted probability of elicited risky behavior.

[Insert Table 8]

In table 8, we show the covariates' means of predicted probability of risky behavior, by quartiles. The prevalence of females is relatively stable across quartiles of predicted probability. On the other hand, a higher probability is predicted for those that are younger, live in urban areas and more educated.

6 Conclusions

In this paper, we build an epidemiological model to show that, as it happens with self-reported data, misclassification of risky sexual behaviors is also possible when using biomarkers for curable sexually transmitted disease (STIs). We thus propose a new GMM estimator to precisely estimate correlates of risky sexual behaviors, by combining biomarkers and self-reported data. Using DHS data from Zambia, we simulate the model and we find that in countries with low STIs prevalence and low average number of sexual partners per individual, the biomarkers have a higher probability of misclassification than behaviors elicited by a survey questionnaire. These results have important implications for policies, given the number of growing studies on the HIV/AIDS epidemic which rely on the results from STIs tests to infer risky sexual behaviors.

References

Anderson, R., and R. May (1991): Infectious Diseases of Humans: Dynamics and Control. Oxford University Press, Oxford, UK.

Baird, S., E. Gong, C. McIntosh, and B. Ozler (2014): "The Heterogeneous Effects of HIV Testing," Journal of Health Economics, 37, 98–112.

Baird, S., and B. Ozler (2010): "Examining the reliability of self-reported data on school participation," Journal of Development Economics, 98(1), 89–93.

Centers for Disease Control and Prevention (2014): "Genital Herpes - STD fact sheet," http://www.cdc.gov/std/herpes/stdfact-herpes.htm.

Cleland, J., M. Boerma, and S. Weir (2004): "Monitoring sexual behaviour in general populations: a synthesis of lessons of the past decade," Sexually Transmitted Infections, 80, 84–92.

De Paula, A., G. Shapira, and P. Todd (2014): "How beliefs about HIV Affect Risky Behaviors: Evidence from Malawi," Journal of Applied Econometrics, 29(6), 944–964.

de Walque, D., W. H. Dow, R. Nathan, and et al. (2012): "Incentivising safe sex: a randomised trial of conditional cash transfers for HIV and sexually transmitted prevention in rural Tanzania," BMJ Open, 2(1), 1–10.

Dow, W., and T. Philipson (1993): "An Empirical Examination of the Implications of Assortative Matching on the Incidence of HIV," Journal of Health Economics, 15, 735–749.

Fenton, K., A. Johnson, S. McManus, and B. Erens (2001): "Measuring sexual behaviour: methodological challenges in survey research," Sexually Transmitted Infections, 77, 84–92.

Gallo, M. F., F. M. Behets, M. J. Steiner, and et al. (2006): "Prostate-Specific Antigen Ascertain Reliability of Self-Reported Coital Exposure to Semen," Sexually Transmitted Infections, 33(8), 476.

Gong, E. (forthcoming): "HIV Testing and Risky Sexual Behaviour," Economic Journal.

Gregson, S., T. Zhuwau, and J. Ndlovu (2002): "Methods to reduce social desirability bias in sex surveys in low-development settings - Experience in Zimbabwe," Sexually Transmitted Diseases, 29, 568–575.

Hausman, J., J. Abrevaya, and F. Scott-Morton (1998): "Misclassification of the dependent variable in a discrete-response setting," Journal of Econometrics, 87(2), 239–269.

Hayashi, F. (2001): Econometrics. Princeton University Press.

Hethcote, H. (2000): "The Mathematics of Infections Diseases," SIAM Review, 42(4), 599–653.

Hewett, P., B. Mensch, and M. Ribeiro (2008): "Using sexually transmitted infection biomarkers to validate reporting of sexual behavior within a randomized, experimental evaluation of interviewing methods," American Journal of Epidemiology, 168.

Hewitt, M. (2002): "Attitudes toward interview mode and comparability of reporting sexual behavior by personal interview and audio computer assisted self-interviewing: analyses of the 1995 National Survey of Family Growth," Sociological Methods Res, 31, 3–26.

Hyman, J., J. Li, and E. Stanley (2001): "The Initialization and Sensitivity of Multigroup Models for the Transmission of HIV," Journal of Theoretical Biology, 208, 227–248.

Karlan, D., and J. Zinman (2012): "List randomization for sensitive behavior: An application for measuring use of loan proceeds," Journal of Development Economics, 98(1), 71–75.

Mauck, C., and A. Straten (2008): "Using objective markers to assess participant behavior in HIV prevention trials of vaginal microbicides," Journal of Acquired Immune Deficiency Syndromes, 1, 49–64.

Minnis, A., M. Steiner, M. Gallo, and et al. (2009): "Biomarker validation of reports of recent sexual activity: Results of a randomized controlled study in Zimbabwe," Journal of Epidemiology, 170.

Ozler, B. (2013): "Economists have experiments figured out. What's next? (Hint: It's Measurement)," Development Impact Blog, `http://blogs.worldbank.org/impactevaluations/economists-have-experiments-figured-out-what-s-next-hint-it-s-measurement`.

Palen, L., E. A. Smith, L. L. Caldwell, and A. e. a. Flisher (2008): "Inconsistent Reports of Sexual Intercourse Among South African High School Students," Journal of Adolescent Health, 43(3), 221–227.

Tourangeau, R., and T. Smith (1996): "Asking sensitive questions: the impact of data collection mode, question format, and question context," Public Opinion Quarterly, 60, 275–304.

Turner, C., L. Ku, S. Rogers, J. Pleck, and et al. (1998): "Adolescent sexual behavior, drug use, and violence: Increased reporting with computer survey technology," Science, 280, 867–873.

UNAIDS (2012): "Global Report. Unaids report on the global AIDS epidemic," Discussion paper, Joint United Nations Programme on HIV/AIDS, Geneva.

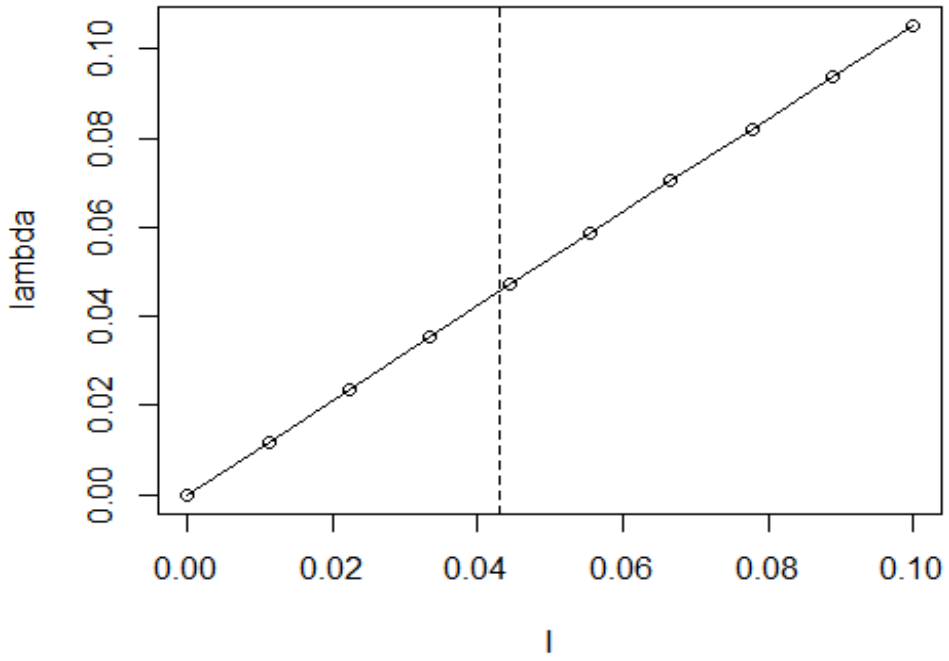
World Health Organization, Dept. of Reproductive Health and Research (2008): "Global incidence and prevalence of selected curable sexually transmitted infections," `http://www.who.int/reproductivehealth/publications/rtis/stisestimates/en/`.

Zambia Central Statistical Office (2009): "Zambia Demographic and Health Survey 2009," Discussion Paper 2, Ministry of Health, Tropical Diseases Research Centre, University of Zambia.

Zenilman, J. M., C. S. Weisman, and A. M. e. a. Rompalo (1995): "Condom use to prevent incident STDs: the validity of self-reported comdom use," Sexually Transmitted Disease, 22, 15–21.

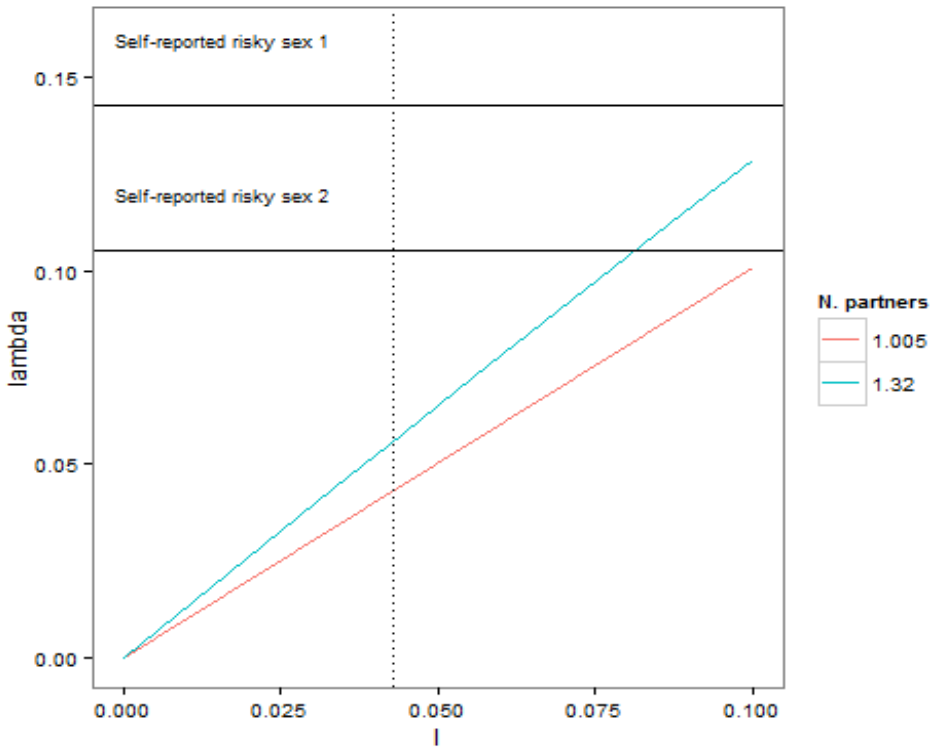
Figures

Figure 1: STI prevalence and probability of correct classification using biomakers



Notes: Lambda is the probability of correct classification of risky sexual behaviors using biomarkers for STIs and I is the syphilis prevalence rate. Source: Demographic and Health Survey, Zambia 2007

Figure 2: Comparison between biomarkers and self-reported risky behavior for lower and upper bound in the number of partners



Notes: Lambda is the probability of correct classification of risky sexual behaviors using biomarkers for STIs and I is the syphilis prevalence rate. "Self-reported risky sex 1" is a binary variable taking value 1 if the respondent reported that a condom was not used the last time he/she had sexual intercourse or if a married respondent reported to have extramarital sex in the last 12 months. "Self-reported risky sex 2" is a binary variable taking value 1 if a non-married respondent reported that a condom was not used the last time he/she had sexual intercourse or a married respondent reported that the last sexual intercourse was not with spouse/cohabiting partner and no condom was used, and 0 if a non-married respondent reported using condom during last intercourse or a married respondent reported not have extramarital sex or extramarital sex with condom. The number of partners equal to 1.005 is the average number of partners for married women in the last 12 months. The number of partners equal to 1.32 is the average number of partners for single men in the last 12 months. Source: Zambia Demographic and Health Survey 2007.

Tables

Table 1: Frequency of sexual contacts

Sexual contacts	*Freq.*	*%*
More than 3 per week	20	9.57
2 per week	46	22.01
2 per month	73	34.93
<2 per month	63	30.14
Missing	7	3.35

Source: Malawi Diffusion and Ideational Change Project (MDICP), 2004. Sample of unmarried women.

Table 2: Descriptive statistics

	Obs.	*Mean*	*S.d.*
Panel A: Self-reported socio-demographic characteristics			
Female	2,414	0.525	0.499
Age	2,414	29.55	10.51
Married	2,414	0.598	0.490
No education	2,414	0.073	0.261
Primary Education	2,414	0.496	0.501
Secondary Education and above	2,414	0.429	0.495
Urban	2,414	0.428	0.495
Panel B: Biomarkers			
Syphilis positive	2,414	0.043	0.203
HIV positive	2,392	0.149	0.356
Panel C: Self-reported sexual behaviors			
Condom used last intercourse	1,836	0.175	0.381
N. of partners in the last 12 months	2,410	0.869	0.669
N. of extramarital partners in the last 12 months	1,442	0.096	0.476
Extramarital sex last intercourse	1,405	0.018	0.134
Risky sex 1	2,124	0.126	0.332
Risky sex 2	2,023	0.104	0.305

Notes: "Condom used last intercourse" is a binary variable equal to 1 if the respondent reported using a condom during the last intercourse, 0 otherwise; "N. of partners in the last 12 months" is the number of sexual partners the respondent reported to have in the last 12 months; "N. of extramarital partners in the last 12 months" is the number of extramarital sexual partners for married/cohabiting individuals in the last 12 months. "Extramarital sex last intercourse" is a binary variable equal to 1 if married/cohabiting respondents reported that the last sexual intercourse was not with their spouse/cohabiting partner. "Risky sex 1" is a binary variable taking value 1 if respondent reported that a condom was not used the last time he/she had sexual intercourse or if a married respondent reported to have extramarital sex as last intercourse. "Risky sex 2" is a binary variable taking value 1 if a non-married respondent reported that a condom was not used the last time he/she had sexual intercourse or a married respondent reported that the last sexual intercourse was not with the spouse/cohabiting partner and no condom was used, and 0 if a non-married respondent reported using condom during last intercourse or a married respondent reported not having had extramarital sex or extramarital sex with condom. Source: Zambia Demographic and Health Survey 2007.

Table 3: Risky sexual behaviours by syphilis status

	Syphilis positive (P)		*Syphilis negative (N)*		*P-value (P=N)*
	Obs	%	Obs	%	
Condom used last intercourse	89	0.157	1748	0.176	0.657
N. of partners in the last 12 months	104	1.067	2306	0.860	0.002
N. of extramarital partners in the last 12 months	104	0.317	2306	0.258	0.335
Extramarital sex last intercourse	78	0.053	1614	0.014	0.008
Risky sex 1	91	0.143	2033	0.125	0.634
Risky sex 2	86	0.105	1937	0.104	0.979

Notes: The P-values tested the null hypothesis that the means between P and N are equal. Variables are defined as in table 2. Source: Zambia Demographic and Health Survey 2007

Table 4: Parameters

Share of infected individuals in the population	I	0.043
Probability of transmission from a single contact with an infected person	ξ	0.2
Parameter that controls for the degree of sorting between infected and uninfected individuals	ρ	1
Parameter that controls how the number of contacts varies with number of partners	η	1
Numbers of sexual partners in the last 12 months	p	1.14
Numbers of sexual partners in the last 12 months for single men	p	1.32
Numbers of sexual partners in the last 12 months for married women	p	1.01

Source: Zambia Demographic and Health Survey 2007 for all the parameters except from ξ that comes from Nelson, K. and Masters Williams, C., Infectious Disease Epidemiology: Theory and Practice, 2nd Ed, Jones and Bartletts Publishers, 2007, page 978.

Table 5: Correlates of risky sexual behaviors

Dependent Variable	*1 if self-reported risky sexual behaviors (Risky sex 1)*			*1 if STI positive (syphilis)*			GMM *1 if truly behaving risky*		
	OLS (1)	*Probit* (2)	*Logit* (3)	*OLS* (4)	*Probit* (5)	*Logit* (6)	*OLS* (7)	*Probit* (8)	*Logit* (9)
Alpha							0.143***	0.143***	0.143***
							(0.04)	(0.04)	(0.04)
Female	-0.044***	-0.189**	-0.386***	0.003	0.041	0.083	-0.310*	-0.120	-0.055
	(0.015)	(0.074)	(0.139)	(0.009)	(0.099)	(0.221)	(0.13)	(1.42)	(1.89)
Age	-0.006***	-0.030***	-0.058***	0.001	0.008**	0.017**	-0.039***	-0.127**	-0.205**
	(0.001)	(0.004)	(0.008)	(0.000)*	(0.004)	(0.008)	(0.01)	(0.04)	(0.07)
Primary Education	-0.056*	-0.305**	-0.541**	-0.006	-0.068	-0.145	-0.395	-4.014	-8.533
	(0.029)	(0.140)	(0.256)	(0.019)	(0.184)	(0.402)	(0.23)	(2.80)	(4.69)
Secondary Education and above	-0.035	-0.192	-0.332	-0.006	-0.058	-0.131	-0.244	-3.333	-7.366
	(0.032)	(0.148)	(0.271)	(0.020)	(0.195)	(0.425)	(0.23)	(2.77)	(4.72)
Urban	0.010	0.047	0.079	0.004	0.045	0.096	0.068	0.630	0.852
	(0.017)	(0.082)	(0.156)	(0.010)	(0.110)	(0.243)	(0.12)	(0.48)	(0.73)
Constant	0.352***	-0.005	0.224	0.025	-1.942***	-3.565***	2.463***	8.224***	15.373*
	(0.040)	(0.187)	(0.342)	(0.023)	(0.237)	(0.517)	(0.67)	(2.08)	(6.78)
Observations	2124	2124	2124	2124	2124	2124	2124	2124	2124

Notes: The dependent variable in column 1-3 is described in the footnote of table 2. Source: Zambia Demographic and Health Survey 2007.

Table 6: Probability of true risky behavior, by socio-demographic characteristics

	Female		Age		Primary Education		Secondary Education and above		Urban		*Probability of true risky behavior*
	0	1	30	40	0	1	0	1	0	1	
Panel A: Probit											
	X		X		X			X		X	96%
	X			X	X			X		X	73%
	X		X			X	X		X		65%
	X			X		X	X		X		51%
		X	X		X			X		X	94%
		X		X	X			X		X	63%
		X		X		X	X		X		16%
		X	X			X	X		X		61%
Panel B: Logit											
	X		X		X			X		X	99%
	X			X	X			X		X	66%
	X		X			X	X		X		67%
	X			X		X	X		X		20%
		X	X		X			X		X	94%
		X		X	X			X		X	66%
		X		X		X	X		X		20%
		X	X			X	X		X		67%

Table 7: Predicted probability of risky behavior, by quartile

	Predicted probability	
	Self-reported sexual behavior	Truly behaving riskly
	(1)	(2)
Mean	0.127	0.741
Standard Deviation	0.062	0.321
50%	0.120	0.910

Notes: In column 1, the predicted probability has been computed with a probit regression using "Risky sex 1" as dependent variable and "Female", "Age", "Primary Education", "Secondary Education", "Urban" as covariates. In column 2 the predicted probability comes from the GMM specification as described in equation (8) in the text. Source: Zambia Demographic and Health Survey 2007.

Table 8: Covariates means for predicted probability of risky behavior, by quartiles

Quartiles of predicted probability	*0-25%*	*25-50%*	*50-75%*	*75-99%*
	(1)	(2)	(3)	(4)
Panel A: Predicted probability of self-reported risky behavior				
Female	0.589	0.612	0.592	0.323
Age	43.205	31.667	24.141	19.105
Primary Education	0.680	0.592	0.477	0.231
Secondary Education and above	0.252	0.350	0.461	0.669
Urban	0.315	0.385	0.416	0.584
Panel B: Predicted probability of true risky behavior				
Female	0.432	0.560	0.559	0.550
Age	42.429	30.242	23.758	21.652
Primary Education	0.752	0.629	0.496	0.097
Secondary Education and above	0.248	0.367	0.465	0.685
Urban	0.263	0.360	0.446	0.675

Notes: In panel A, the covariate means of the predicted probabilities have been computed with a probit regression using "Risky sex 1" as dependent variable and "Female", "Age", "Primary Education", "Secondary Education" "Urban" as covariates. In panel B, they come from the GMM specification as described in equation (8) in the text. Source: Zambia Demographic and Health Survey 2007.