Published in final edited form as:
Epidemiology. 2015 March ; 26(2): 229–237. doi:10.1097/EDE.0000000000000218.

On the Assumption of Bivariate Normality in Selection Models: A Copula Approach Applied to Estimating HIV Prevalence

Mark E. McGovern[1,2,*], **Till Bärnighausen**[2,3], **Giampiero Marra**[4], and **Rosalba Radice**[5]

[1] Harvard Center for Population and Development Studies, University of KwaZulu-Natal

[2] Department of Global Health and Population, Harvard School of Public Health, University of KwaZulu-Natal

[3] Wellcome Trust Africa Centre for Health and Population Studies, University of KwaZulu-Natal

[4] Department of Statistical Science, University College London

[5] Department of Economics, Mathematics and Statistics, Birkbeck, University of London

Abstract

Background—Heckman-type selection models have been used to control HIV prevalence estimates for selection bias, when participation in HIV testing and HIV status are correlated after controlling for observed variables. These models typically rely on the strong assumption that the error terms in the participation and the outcome equations that comprise the model are distributed as bivariate normal.

Methods—We introduce a novel approach for relaxing the bivariate normality assumption in selection models using non-linear copula functions. We apply this method to estimating HIV prevalence and new confidence intervals (CI) in the 2007 Zambian Demographic and Health Survey (DHS), using interviewer identity as the selection variable that predicts participation (consent to test) but not the outcome (HIV status).

Results—We show in a simulation study that selection models can generate biased results when the bivariate normality assumption is violated. In the 2007 Zambia DHS, HIV prevalence estimates are similar irrespective of the structure of the association assumed between participation and outcome. For men, we estimate a population HIV prevalence of 21% (95% = CI 16% to 25%), compared with 12% (11% to 13%) among those who consented to be tested; for women, the corresponding figures are 19% (13% to 24%) and 16% (15% to 17%).

Conclusions—Copula approaches to Heckman-type selection models are a useful addition to the methodological toolkit of HIV epidemiology, and of epidemiology in general. We develop the use of this approach to systematically evaluate the robustness of HIV prevalence estimates based on selection models, both empirically and in a simulation study.

[*] Corresponding Author. Address: 9 Bow Street, Cambridge, MA 02138, USA. mcgovern@hsph.harvard.edu. Tel: +1 617-496-4573.

Conflicts of Interest: None Declared

SDC Supplemental digital content is available through direct URL citations in the HTML and PDF versions of this article (www.epidem.com). This content is not peer-reviewed or copy-edited; it is the sole responsibility of the author.

In order to address almost every aspect of the HIV epidemic, from assessing the risk factors associated with infection, to planning future resource allocation, to anti-retroviral treatment (ART) scale-up, accurate information on HIV prevalence is required.[1] Research on HIV/ AIDS often relies on nationally representative surveys,[2] but participation rates in these surveys can be low. Table 1 shows that participation rates in the HIV surveys that are nested within one of the major sources of nationally representative data in low- and middle-income countries, the Demographic and Health Surveys (DHS), range from a high of 97% for women in Rwanda in 2005, to a low of 63% for men in Malawi in 2004 and Zimbabwe in 2005.[3]

There are many potential reasons for low participation rates in HIV surveys (including concerns about the confidentiality of results, lack of incentive to participate, and survey fatigue[1,4], and non-participation can arise at different stages of HIV survey administration.[5] In this paper, we focus on refusal to be tested for HIV, which is typically the most important cause of missing data in HIV surveys.[6] In longitudinal studies, it has been shown that respondents who are HIV-positive are less likely to consent to be tested for HIV than HIV-negative individuals.[7–10] In Malawi, 46% of women and 39% of men who declined to be tested did so because of prior HIV testing, knowledge of HIV status, or fear of positive results.[11] Such reasons for declining to participate in an HIV survey have implications for the calculation of HIV prevalence. Neither complete case analysis (limiting the analysis only to people who consent to be tested for HIV) nor standard approaches to account for missing values generate unbiased estimates in the presence of selection on unobserved variables.[12,13] A potentially important situation leading to selection into survey participation based on unobserved variables occurs if HIV status itself predicts consent to be tested for HIV. This scenario is likely if people know that they are HIV-positive and fear that others will learn about their positive HIV status if they participate in a survey. It is also likely if people suspect that they are HIV positive (for instance, based on evaluation of past sexual behavior), if this suspicion predicts true HIV status, if they fear confirmation of this suspicion, or they fear that others might learn about their positive status. In these cases, standard approaches to correct HIV prevalence estimates for missing values (such as single imputation, multiple imputation, inverse probability weighting or propensity score reweighting) will be biased because they can only account for selection on observed factors, but HIV status is unobserved among those who refused to be tested. Another consequence of high refusal rates is that the uncertainty associated with estimating HIV prevalence can increase substantially, leading to wide confidence intervals.[3]

More generally, missing data is a common problem in epidemiologic studies, and the mechanisms through which this occurs can have an important impact on resulting estimates. Heckman-type selection models can provide asymptotically unbiased estimates of the parameters of interest, even when missing data are systematically related to unobserved characteristics of the individual.[14,15] These models will thus be useful whenever researchers cannot be certain that the assumption that is required for the standard approaches to generate unbiased results holds – i.e., that data are missing at random after selection on observed variables has been taken into account. However, in practice the use of Heckman-type selection models is limited by one requirement and one statistical assumption.

Heckman-type selection models require the existence of a selection variable that predicts participation in a survey but not the outcome of interest, other than through the effect on participation. Elements of survey design and implementation are often documented in datasets in epidemiology.[16] Characteristics of these elements are often likely to determine survey participation, and are thus potential candidates for selection variables if they are also plausibly uncorrelated with the characteristics of the individuals who are potential participants in a survey.[17] In HIV surveys, interviewer identity generally predicts consent to be tested, but it is unlikely that it also predicts HIV status. Previous research that has used interviewer identify as a selection variable in Heckman-type selection models has found evidence for selection on unobserved variables in several HIV surveys.[3,18–21]

The key statistical assumption that the standard Heckman-type selection models need to meet is that the relationship between consenting to be tested for HIV and HIV status follows a bivariate normal distribution after other covariates have been taken into account, i.e., that the error terms of the two equations in Heckman-type selection models are distributed as bivariate normal. While this assumption is convenient and tractable, it is a potentially serious limitation.[22–25] If this assumption is met, then the estimates obtained using the conventional bivariate probit Heckman-type selection model are consistent and asymptotically efficient. However, if the true distribution of the error terms is not bivariate normal, then the estimates are likely to be both inconsistent and inefficient.[26] Simulation studies have indicated that HIV prevalence estimates from selection models may indeed be sensitive to violations of this assumption.[27]

The robustness of results obtained from surveys involving missing data is particularly important.[23] The implementation of selection models can be viewed as a sensitivity analysis to adjust for potential bias using alternative sets of assumptions about the underlying mechanisms causing data to be missing. If it can be demonstrated that the results obtained in selection models are invariant to a variety of alternative assumptions regarding the mechanisms leading to missing data, our belief that the conclusions are not just a function of the model imposed by the researcher will be substantially strengthened. The lack of methods for evaluating the robustness of Heckman selection models is likely an impediment to wider use of this approach.

The aim of this paper is to develop and illustrate a means of determining the sensitivity of results from selection models to alternative ways of characterizing the functional form of the association between the participation equation (in this case, consent to be tested for HIV) and the outcome equation (in this case, HIV status). Copulae have been previously applied to recursive models involving a treatment that is affected by unobserved variables (such as health as function of medical care utilization[28–31]) and in censored models with continuous outcomes.[32] The main contribution of this paper is the application of copulae to binary outcomes with missing data. In addition, we use a variety of copulae (including the rotated Clayton, Joe, and Gumbel), allowing for large flexibility in modelling dependence. This flexibility is a key characteristic of our approach, because it allows us to capture a much wider set of possible dependence structures than those used in the previous literature.[33] With this method, and the number of alternative parametric specifications, we are therefore able to be more confident in assessing the robustness of results based on the standard Heckman-type selection models. For example, whereas previous implementations of the copula approach have generally focused on distributions that are similar to the bivariate normal (such as the Frank or Clayton[31]), we are able to consider asymmetric dependence. In addition to potential bias arising from misspecification of the error distribution, by potentially providing a more accurate representation of the underlying data structure, the copula approach may also provide more efficient estimates, allowing us to make better inferences. This approach with asymmetric copulae has not been previously implemented in the sample selection literature.

In what follows, we introduce and demonstrate our methodology for relaxing the assumption of bivariate normality in Heckman-type selection models that allow for non-linear association between participation and the outcome of interest. Although, in theory, semi-parametric or nonparametric approaches would not require any distributional assumptions, their application to estimating the intercept in sample selection models with binary data and a high degree of missing data is limited due to their inefficiency and computational feasibility. While the copula method does require parametric specification, our approach makes many distributional functional forms available, therefore making copulae a viable practical alternative to imposing bivariate normality. We illustrate the consequences of violating the normality assumption in a simulation study, and show that copulae can provide an effective and practical means of adjusting for this bias and inefficiency. Finally, we evaluate the robustness of estimates of HIV prevalence in Zambia. We provide the relevant code in order to make this approach easily accessible to researchers working with surveys containing missing data eAppendix.

Methods

Statistical Approach

We begin by modelling consent to be tested for HIV and HIV status simultaneously, an approach based on the adaptation of the original Heckman selection model estimator for binary outcomes.[14,34,35]

Consent to be tested is given by:

$$Consent_{ij}^{*}=X_{ij}^{T}\beta+Z_{j}^{T}\quad\alpha+u_{ij}, i=1,\ldots n,\quad j=1\ldots J \quad (1)$$

$$Consent_{ij}=1\quad if\quad Consent_{ij}^{*}>0,\quad Consent_{ij}=0\quad otherwise \quad (2)$$

The observed consent for person *i* with interviewer *j*, $Consent_{ij}$, is a dummy variable indicating acceptance of being tested, and is a function of a latent variable $Consent_{ij}^{*}$, which reflects the respondent's propensity to be tested. X_{ij} is a $p \times 1$ vector representing observed individual level characteristics with associated parameter vector β, Z_i is a $k \times 1$ vector of dummy variables representing interviewer identity with associated parameter vector α, and u_{ij} is a random error term. Although, in theory, identification can be achieved using the same set of regressors in both the participation equation and the outcome equation, in practice empirical identification in selection models requires at least one variable, the selection variable, to be present in the participation equation but not the outcome equation.[32,36] In this case, interviewer identity predicts consent to be tested but does not to enter into the HIV equation directly.

The equation for the HIV status HIV_{ij} of individual with i interviewer j is:

$$HIV^{*}_{ij}=X^{T}_{ij}\mathrm{Y}+\varepsilon_{ij} \quad (3)$$

$$HIV_{ij}=1 \quad if \quad HIV^{*}_{ij}>0, \quad HIV_{ij}=0 \quad otherwise \quad (4)$$

$$HIV_{ij} \quad observed \quad only \quad if \quad Consent_{ij}=1, \quad missing \quad otherwise \quad (5)$$

where γ is a parameter vector and ε_{ij} is a random error term. The structural assumption used in previous studies to estimate HIV prevalence is that the error terms in both equations (u_{ij}, ε_{ij}) are independent and identically distributed (i.i.d.) as bivariate normal, with means equal to zero, constant variances equal to one, and covariance (correlation coefficient)ρ. That is, the joint distribution of u_{ij}, ε_{ij} is given by $F(u_{ij}, \varepsilon_{ij}) = \Phi_2 u_{ij}, \varepsilon_{ij};\rho)$, where Φ_2 is the standardized bivariate normal cumulative distribution function (cdf). This model can be fitted using classic maximum likelihood. The standard selection model that relies on joint normality is equivalent to specifying the Gaussian copula in our framework; therefore we use this model as the baseline for our comparisons.

In order to allow for non-linear association between the consent and HIV status equations, we model the dependency of the error terms in the two equations using copulae. Broadly speaking, these are functions that connect multivariate distributions to their one dimensional margins, such that if F is a two-dimensional cdf with one-dimensional margins ($F_1(y_1)$, $F_2(y_2)$), then there exists a two-dimensional copula C such that $F(y_1, y_2) = C\,(F_1(y_1), F_2(y_2); \theta)$, where y_1 and y_2 are two random variables, and θ is an association parameter measuring the dependence between the two marginals.[37] If HIV-positive persons are refusing to be tested on the basis of knowledge of their HIV status,[7–10] we would expect a value of $\rho < 0$. When we estimate the model for Zambia using symmetric copulae (Gaussian, Frank, Student-t) that do not impose a sign on the relationship between consent and HIV status, the dependence is estimated to be negative in the data, and when we implemented asymmetric copulae that specify positive associations (Clayton 0 and 180, Joe 0 and 180, Gumbel 0 and 180), we found that the models did not converge. Therefore, we focus on those copulae that allow for negative association. However, in other contexts there could just as easily be a positive relationship, when this method is equally applicable. The models we consider are therefore: Gaussian (C_g), equivalent to the standard bivariate normal probit model; Frank (C_f); 90 and 270 degrees rotated Clayton (C_{c90}, C_{c270}); 90 and 270 degrees rotated Joe (C_{J90}, C_{J270}); 90 and 270 degrees rotated Gumbel (C_{G90}, C_{G270}); and Student-t (C_t). These copulae are reported in Table 2 and illustrated in Figure 1. While the Gaussian, Frank and Student-t copulae are symmetric, the rotated Clayton, Joe and Gumbel copulae allow for stronger negative dependence in the tails of the distribution. The 90- and 270-degrees rotated versions can be obtained using the following equations[38]:

$$\begin{aligned} C_{90} = & \quad F_2\left(y_2\right) - C\left(1 - F_1\left(y_1\right), F_2\left(y_2\right); \theta\right) \\ C_{270} = & \quad F_1\left(y_1\right) - C\left(F_1\left(y_1\right), 1 - F_2\left(y_2\right); \theta\right) \end{aligned}$$

These forms of dependence are particularly applicable in the context of HIV prevalence estimation, as we might expect respondents with a strong negative score on the latent test variable to be of particularly high risk of being HIV-positive.

In the sample selection context, the data identify the three possible events ($Consent_{ij}$ = 1, HIV_{ij} = 1), ($Consent_{ij}$ = 1, HIV_{ij} = 0) and ($Consent_{ij}$ = 0), with probabilities:

$$\begin{gathered} P\left(Consent_{ij}{=}1, \quad HIV_{ij}{=}1\right) = p_{11ij} = C\left(\Phi\left(X_{ij}^T\beta - Z_j^T\alpha\right), \Phi\left(X_{ij}^T \mathrm{Y}\right); \theta\right) \\ P\left(Consent_{ij}{=}1, \quad HIV_{ij}{=}0\right) = p_{01ij} = \Phi\left(X_{ij}^T\beta + Z_j^T\alpha\right) - p_{11ij} \\ P\left(Consent_{ij}{=}0\right) = p_{0ij} = 1 - \Phi\left(X_{ij}^T\beta + Z_j^T\alpha\right) \end{gathered}$$

The log-likelihood function is therefore:

$$\ell\left(\boldsymbol{\delta}\right) = \sum_{i=1}^{n} Consent_{ij} \times HIV_{ij}\, log\left(p_{11ij}\right) + Consent_{ij} \times \left(1 - HIV_{ij}\right) log\left(p_{01ij}\right) + \left(1 - Consent_{ij}\right) log\left(p_{0ij}\right) \quad (11)$$

where δ^{T} (β^{T}, a^{T}, γ^{T}, θ).

Maximization is based on a trust region Newton algorithm and not the usual Newton-Raphson algorithm, resulting in more stable computation and better convergence properties, which is valuable because another common criticism of these models is that they can often fail to converge.

We assess the degree of association between the consent and HIV status equations using a nonparametric measure of rank (Kendall's Tau, τ), which is more appropriate than the correlation coefficient (ρ) as the dependence modelled by copulae is typically non-linear. τ can be interpreted in the same manner as ρ in the sense that it ranges between -1 and +1; therefore if persons who refuse to be tested are more likely to be HIV positive, we would expect to see a value of $\tau < 0$. The approximate posterior cumulative distribution function $F_{\tau}\hat{}$ is obtained by simulating a set of random values, $\{\theta_r$: r = 1, ... ,$R\}$, from the multivariate normal posterior of δ such that:

$$\hat{F}\left(\tau\right) = \frac{1}{R}\sum_{r=1}^{R} H\left(\tau - \tau\left(\theta_r\right)\right)$$

where H is the Heaviside function (jumping from 0 to 1 at τ). Confidence intervals are obtained from quantiles of this distribution. Intervals for $\tau(\theta)$ may also be obtained by bootstrapping. The HIV prevalence estimate is computed as a weighted average of individual predicted values with survey weights, w_{ij} :

$$\hat{P}\left(HIV=1\right)=\left(\sum_{i=1}^{n} w_{ij}\hat{P}\left(HIV_{ij}=1|X_{ij}\right)\right)/\sum_{i=1}^{n} w_{ij}$$

We use a Taylor-series expansion to derive the large-sample variance estimator for the point estimate of HIV prevalence, which simultaneously acknowledges uncertainty due to cluster effects and the presence of sampling weights.[39]

There are no disadvantages to not specifying the standard normality assumption as we are using a likelihood based model; hence, asymptotic theory will still hold under the usual regularity conditions, and we can evaluate model fit using information criteria (for example, the Akaike Information Criterion [AIC]). However, it is important to understand the relative performance of the standard Heckman-type model in comparison with the copula approach. Therefore, we undertake a simulation study in order to determine the conditions under which the normality assumption performs well, and to assess the extent of bias which arises from misspecification of the error terms' distribution.

Simulation Study

We follow the approach implemented in Clark and Houle by generating a dataset based on a real HIV survey (the 2007 Zambian DHS).[27] Therefore, our simulations closely match the observed consent rates and HIV prevalence in the data used in the empirical part of this paper. We construct latent variables for consent and HIV status, and allow for interviewer identity to influence the probability of consent. Then we draw error terms for the latent variable equations in order to induce a correlation between consent and HIV status (which we censor for individuals with $Consent_{ij}$ = 0). As we know the true HIV prevalence, we can evaluate the relative performance of imputation, the standard selection model, and our copula selection model. By varying the structure of the error terms, we assess the extent to which the standard selection model is sensitive to the assumption of bivariate normality, and whether the copula approach can be used to correct for potential bias and inefficiency.

We confirm that the imputation model performs poorly when there is correlation between consent and HIV status (bias of between 40% and 50%), and that selection models are appropriate for correcting for this correlation. We find that the performance of the bivariate normal selection model is related to the strength of the relationship between the selection variable and consent. This closely parallels the case of instrumental variables, and is consistent with previous results.[40] When the relationship between interviewer identity and consent is less strong, bias and inefficiency can arise when the model is misspecified. For example, when normal errors are cubed we find the mean bias of the standard Heckman-type model is -14%, while the bias in the copula model is less than half this amount, as well as being more efficient. The distribution of the normal and copula estimators, along with that for the imputation model, is shown in Figure 2. Further details are presented in the eAppendix.

Data

We use data from the 2007 Zambian DHS (publically accessible from www.dhsprogram.com). We adopt the same explanatory variables and specification as used in previous research,[3] the code for which is freely available online from http://hdl.handle.net/1902.1/17657. As outlined in model (1), interviewer identity enters into the consent equation as a series of dummy variables, one for each interviewer. As some interviewer fixed effects are collinear with other variables in the model, interviewers with fewer than 50 interviewees, or those with interviewer effects which are collinear, are combined into a single category. After combining, there are 29 interviewers for men, and 45 for women. We focus on estimating selection models for persons who refused to consent to be tested, as opposed to respondents who have missing HIV data due to non-contact, as there are relatively few of these persons compared with those who refuse. However, the methodology we propose could be easily applied to respondents who were not contacted. Table 3 illustrates the composition of the analysis sample for men and women separately. Excluding non-contacts, of the eligible 6,416 men, 1,318 (21%) declined to take a HIV test; of the eligible 7,025 women in the survey, 1,400 (20%) declined to take a HIV test. Table 3 also illustrates the HIV prevalence estimate based on the complete case analysis (respondents with a valid HIV test), which is estimated to be 12% for men and 16% for women.

All our estimates of HIV prevalence are weighted and take account of the complex survey design of the DHS.[41] Statistical analyses were performed in R version 3.1.1 , using the SemiParBIVProbit package.[42]

Results

Table 4 presents estimates for the rank association between consenting to be tested and HIV status (Kendall's Tau, τ) for each of the nine copula models employed, along with the corresponding 95% confidence intervals (CIs), which account for clustering at the primary sampling unit level. A measure of model fit (the AIC) is also presented in the final column of Table 4. While the AIC is not adjusted for clustering, this limitation is unlikely to affect the preferred ordering of the models.[43] For men, there is support for the hypothesis of selection bias, with a negative association for each of the copula models, and the 95% CI for τ excludes zero in each case. The τ of -0.53 for the normal model corresponds to a ρ (correlation coefficient) of -0.73. On the basis of the AIC, the model with the best fit is the C_{ij}.

For women, the measure of association between testing and HIV status is also negative, although the association is less strong than for men, with the 95% CIs in most models including zero. The τ of -0.19 in the normal model corresponds to a ρ of -0.30. On the basis of the AIC, the preferred copula specification for women is C_g or C_{c270} .

Table 5 gives the corresponding HIV prevalence estimates. Point estimates for all copula models for men are similar, ranging from 19% to 21%, with the preferred model (C_{J90} copula) indicating a population HIV prevalence of 21% (with a corresponding 95% CI of 16% to 25%). As with men, HIV prevalence estimates for women are not sensitive to the choice of the copula function, ranging between 18% and 19%. The results for the preferred copula model (C_g) is 19% (with a 95% CI of 13% to 24%).

Discussion

Longitudinal evidence has demonstrated that people who do not consent to be tested in HIV surveys are more likely to be HIV positive than people who do consent to be tested.[7–10] Heckman-type selection models can be used to correct for the bias in data that are missing due to unobserved variables. However, the practical use of these selection models has been criticized for the strong assumptions required for their implementation.[22–25] Our method provides estimates of HIV prevalence that are corrected for missing data on unobserved variables, without relying on the assumption of bivariate normality for identification. This study shows how the credibility of conclusions from selection models can be enhanced by demonstrating that identification does not rely on a specific functional form for estimation – here, for the example of estimating HIV prevalence in Zambia. The wider variety of error distributions we consider provide a more meaningful assessment of the importance of the bivariate normality assumption than was previously possible using existing methods.

Our results indicate population HIV prevalence for men in the preferred selection model that is statistically larger than that based on the assumption of missing at random for the data on respondents who refuse to consent to be tested. The preferred copula model for men, the Joe 90 (C_{J90}), indicates the presence of asymmetric dependence. This finding highlights the importance of our contribution of allowing for a large number of parametric structures. The previous literature relied on a more narrow set of models, which did not include the rotated Joe, Gumbel or Clayton copulas.[33] In addition, we find that the corresponding 95% CI for the Joe 90 copula estimate is substantially narrower than that obtained from the bivariate normal model, indicating an efficiency gain from implementing a dependence structure that may more accurately reflect the true underlying distribution of the data.

In this analysis, imputation models, which require that the strong assumption of data being missing at random is met, produced results that are almost identical to the complete case analysis of respondents who have a valid HIV test, which is similar to previous findings.[3,6,44] Given the increasing focus on treatment-as-prevention in HIV research and policy, it is likely that the HIV surveys will increase in both frequency and coverage in many settings. Therefore, the issue of non-response bias in such surveys will likely increase in importance. Moreover, knowledge of HIV status, and therefore the potential for selection bias that depends on the unobserved variable HIV status is also likely to increase as a result. The development of approaches to correct for selection on unobserved variables while relying on as few assumptions as possible, as well as approaches to test the robustness of the results from such selection models to variation in assumptions, are an important aim. The use of copula functions in Heckman-type selection models is an important advance toward this aim. We believe that our approach using several parametric assumptions in the implementation of Heckman-type selection models makes the use of these models an even more viable alternative to the other approaches to correct for selection bias, which require that the strong and untestable assumption that data are missing at random.

Our simulation results indicate that estimates obtained from the standard selection model that assumes bivariate normality can be biased and inefficient when the structure of the error term is misspecified. The copula models we propose perform well under a variety of different correlational structures, including scenarios with asymmetry. While these conclusions are valid for the simulation settings considered here, it cannot be determined *a priori* whether relaxing the assumption of normality will lead to dramatically different estimated prevalence, as the error terms are not observed and the true structure is unknown. It is difficult to simulate the highly complex processes that likely underlie the relationship between consent to HIV testing and HIV status. However, these results do suggest that there are a variety of scenarios where an incorrect normality assumption leads to biased results, and where the copula approach can correct for this bias.

The methodology we outline is easily implemented in standard statistical software (http://cran.r-project.org/web/packages/SemiParBIVProbit), and we provide the code for all the analyses discussed in this paper (eAppdendix). Assessing the sensitivity of selection model results to relaxing the bivariate normality assumption is easily achieved with this approach, not only in the specific context of HIV prevalence estimation but also in other empirical applications.

There are a number of avenues for future research. First, the literature on copula model selection for censored data is underdeveloped. Implementing goodness-of-fit tests is difficult due to the combination of censoring, the fact that the error terms are unobserved, and the fact that the outcomes are binary. We have focused here on conventional information criteria, but goodness-of-fit tests in this context are an important area for development, which could substantially improve the performance of copula models. Secondly, there are advantages and disadvantages associated with the copula approach compared with semi-parametric and nonparametric models. The latter have the advantage of not requiring the true parametric model to be specified by the researcher. However, while theoretically possible,[26] the intercept is typically not identified in these models, and so this approach is not suitable for estimating population means based on binary outcomes, such as HIV prevalence. Semi-parametric approaches that allow for the estimation of the intercept require additional assumptions and have only been developed for the case of continuous outcomes.[45,46] Additionally, semi-parametric approaches typically generate estimates that are inefficient relative to fully parameterized models, may not allow diagnostics, are limited with regards to the inclusion of a large set of covariates, and may be computationally demanding.[47] In contrast, the computational simplicity of the copula approach allows the practitioner to exploit familiar tools such as maximum likelihood without requiring simulation methods or numerical integration. Maximum likelihood, in turn, allows for the simultaneous estimation of all the parameters of the model and, if the usual regularity conditions are met, ensures consistent, efficient and asymptotically normal estimators.[32] Finally, copula modelling allows for direct estimation of the dependence structure in the sample selection model, while semi-parametric methods do not.[48]

Further analysis should focus on establishing the validity of the other main requirement in sample selection models underlying the estimation of HIV prevalence in the presence of non-response, namely the existence of a selection variable that does not independently affect the outcome of interest. While interviewer identity is plausibly a function only of survey design, and not related to individual-level characteristics, this claim is difficult to prove conclusively. As a robustness check we included a cluster random effect in our model using a two-stage procedure in order to account for potential correlation between interviewer allocation and the characteristics of the individual's primary sampling unit.[34] HIV prevalence estimates in this analysis were similar, but this approach is inefficient and resulted in an attenuated relationship between consent and interviewer identity. Therefore, incorporating random effects directly into these types of selection models is another important direction for future research. In general, as we never observe the HIV status of respondents who refuse to be tested, establishing whether estimates based on selection models can be supported with objective external data, such as alternative selection variables or mortality records, would help validate this approach.

In sum, we introduce and demonstrate a new approach for relaxing the assumption of bivariate normality in Heckman-type selection models with binary outcomes using copulas. Our simulation study illustrates that this methodology can be used to correct for the bias and inefficiency associated with misspecification of the dependence structure between selection into the data and the outcome of interest. In empirical work, establishing that selection model estimates are robust to alternative functional form specifications for the relationship between selection and the outcome increases the credibility of these estimates.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgements

We are grateful to Slawa Rokicki for invaluable suggestions and advice. We also thank the editor, two anonymous referees, seminar participants at Harvard University, University College London, and the UNAIDS Reference Group on Estimates, Modelling and Projections for comments.

The Program on the Global Demography of Aging receives funding from the National Institute on Aging, Grant No. 1 P30 AG024409-09.

References

1. Gersovitz M. HIV testing: principles and practice. World Bank Res. Obs. 2011; 26:1–41.
2. Boerma JT, Ghys PD, Walker N. Estimates of HIV-1 prevalence from national population-based surveys as a new gold standard. Lancet. 2003; 362:1929–1931. [PubMed: 14667753]
3. Hogan DR, et al. National HIV prevalence estimates for sub-Saharan Africa: controlling selection bias with Heckman-type selection models. Sex. Transm. Infect. 2012; 88:i17–i23. [PubMed: 23172342]
4. Sterck O. Why Are Testing Rates So Low in Sub-Saharan Africa? Misconceptions and Strategic Behaviors. Forum for Health Economics and Policy. 2013; 16
5. Marston M, Harriss K, Slaymaker E. Non-response bias in estimates of HIV prevalence due to the mobility of absentees in national population-based surveys: a study of nine national surveys. Sex. Transm. Infect. 2008; 84:i71–i77. [PubMed: 18647870]
6. Mishra V, Barrere B, Hong R, Khan S. Evaluation of bias in HIV seroprevalence estimates from national household surveys. Sex. Transm. Infect. 2008; 84:i63–i70. [PubMed: 18647869]
7. Bärnighausen T, Tanser F, Malaza A, Herbst K, Newell M. HIV status and participation in HIV surveillance in the era of antiretroviral treatment: a study of linked population-based and clinical data in rural South Africa. Trop. Med. Int. Health. 2012; 17:e103–e110. [PubMed: 22943374]
8. Floyd S, et al. Underestimation of HIV prevalence in surveys when some people already know their status, and ways to reduce the bias. AIDS. 2013; 27:233–242. [PubMed: 22842993]

9. Obare F. Nonresponse in repeat population-based voluntary counseling and testing for HIV in rural Malawi. Demography. 2010; 47:651–665. [PubMed: 20879682]

10. Reniers G, Eaton J. Refusal bias in HIV prevalence estimates from nationally representative seroprevalence surveys. AIDS Lond. Engl. 2009; 23:621.

11. Kranzer K, et al. Individual, household and community factors associated with HIV test refusal in rural Malawi. Trop. Med. Int. Health. 2008; 13:1341–1350. [PubMed: 18983282]

12. Conniffe D, O'Neill D. Efficient Probit Estimation with Partially Missing Covariates. Adv. Econom. 2011; 27:209–245.

13. Donders ART, van der Heijden GJ, Stijnen T, Moons KG. Review: a gentle introduction to imputation of missing values. J. Clin. Epidemiol. 2006; 59:1087–1091. [PubMed: 16980149]

14. Heckman JJ. Sample selection bias as a specification error. Econom. J. Econom. Soc. 1979:153–161.

15. Vella F. Estimating models with sample selection bias: a survey. J. Hum. Resour. 1998:127–169.

16. O'Muircheartaigh C, Campanelli P. The relative impact of interviewer effects and sample design effects on survey precision. J. R. Stat. Soc. Ser. A Stat. Soc. 1998; 161:63–77.

17. Bärnighausen T, Bor J, Wandira-Kazibwe S, Canning D. Interviewer identity as exclusion restriction in epidemiology. Epidemiology. 2011; 22:446. [PubMed: 21464660]

18. Bärnighausen T, Bor J, Wandira-Kazibwe S, Canning D. Correcting HIV prevalence estimates for survey nonparticipation using Heckman-type selection models. Epidemiology. 2011; 22:27–35. [PubMed: 21150352]

19. McGovern ME, Bärnighausen T, Salomon JA, Canning D. Using Interviewer Random Effects to Calculate Unbiased HIV Prevalence Estimates in the Presence of Non-Response: a Bayesian Approach. PGDA Work. Pap. 2013

20. Janssens W, van der Gaag J, de Wit TFR, Tanovia Z. Refusal bias in the estimation of HIV prevalence. Demography. 2014:1–27. [PubMed: 24357021]

21. Reniers G, Araya T, Berhane Y, Davey G, Sanders EJ. Implications of the HIV testing protocol for refusal bias in seroprevalence surveys. BMC Public Health. 2009; 9:163. [PubMed: 19476618]

22. Arpino B, Cao ED, Peracchi F. Using panel data for partial identification of human immunodeficiency virus prevalence when infection status is missing not at random. J. R. Stat. Soc. Ser. A Stat. Soc. 2013 n/a–n/a doi:10.1111/rssa.12027.

23. Geneletti S, Mason A, Best N. Adjusting for selection effects in epidemiologic studies: why sensitivity analysis is the only ‘solution’. Epidemiology. 2011; 22:36–39. [PubMed: 21150353]

24. Puhani P. The Heckman correction for sample selection and its critique. J. Econ. Surv. 2000; 14:53–68.

25. Vytlacil E. Independence, monotonicity, and latent index models: An equivalence result. Econometrica. 2002; 70:331–341.

26. De Luca G. SNP and SML estimation of univariate and bivariate binary-choice models. Stata J. 2008; 8:190.

27. Clark SJ, Houle B. Evaluation of Heckman Selection Model Method for Correcting Estimates of HIV Prevalence from Sample Surveys via Realistic Simulation. Cent. Stat. Soc. Sci. Work. Pap. No 120 Univ. Wash. 2012

28. Dancer D, Rammohan A, Smith MD. Infant mortality and child nutrition in Bangladesh. Health Econ. 2008; 17:1015–1035. [PubMed: 18636430]

29. Murteira JM, Lourenço ÓD. Health care utilization and self-assessed health: specification of bivariate models using copulas. Empir. Econ. 2011; 41:447–472.

30. Prieger JE. A flexible parametric selection model for non-normal data with application to health care usage. J. Appl. Econom. 2002; 17:367–392.

31. Winkelmann R. Copula Bivariate Probit Models: With an Application to Medical Expenditures. Health Econ. 2012; 21:1444–1455. [PubMed: 22025413]

32. Smith MD. Modelling sample selection using Archimedean copulas. Econom. J. 2003; 6:99–123.

33. Radice R, Marra G, Wojtys M. Copula Regression Spline Models for Binary Outcomes With Application in Health Care Utilization. Univ. Coll. Lond. Res. Rep. 2013; 321

34. Dubin JA, Rivers D. Selection bias in linear regression, logit and probit models. Sociol. Methods Res. 1989; 18:360–390.

35. Van de Ven WP, Van Praag B. The demand for deductibles in private health insurance: A probit model with sample selection. J. Econom. 1981; 17:229–252.

36. Madden D. Sample selection versus two-part models revisited: the case of female smoking and drinking. J. Health Econ. 2008; 27:300–307. [PubMed: 18180064]

37. Trivedi PK, Zimmer DM. Copula Modeling: An Introduction for Practitioners. Found. Trends R Econom. 2007; 1:1–111.

38. Brechmann EC, Schepsmeier U. Modeling dependence with C-and D-vine copulas: The R-package CDVine. J. Stat. Softw. 2012; 52:1–27.

39. Lumley T. Analysis of complex survey samples. J. Stat. Softw. 2004; 9:1–19.

40. Leung SF, Yu S. On the choice between sample selection and two-part models. J. Econom. 1996; 72:197–229.

41. Corsi DJ, Neuman M, Finlay JE, Subramanian S. Demographic and health surveys: a profile. Int. J. Epidemiol. dys. 2012; 184

42. Marra G, Radice R. SemiParBIVProbit: Semiparametric Bivariate Probit Modelling. R Package Version. 2014:32–11.

43. Dziak J, Li R. Variable Selection with Penalized Generalized Estimating Equations. Methodol. Cent. Pa. State Univ. 2006

44. Zaidi J, Grapsa E, Tanser F, Newell M-L, BÃ¤rnighausen T. Dramatic increase in HIV prevalence after scale-up of antiretroviral treatment. AIDS. 2013; 27:2301–2305. [PubMed: 23669155]

45. Andrews DW, Schafgans MM. Semiparametric estimation of the intercept of a sample selection model. Rev. Econ. Stud. 1998; 65:497–517.

46. Schafgans M, Zinde-Walsh V. On intercept estimation in the sample selection model. Econom. Theory. 2002; 18:40–50.

47. Bhat CR, Eluru N. A copula-based approach to accommodate residential self-selection effects in travel behavior modeling. Transp. Res. Part B Methodol. 2009; 43:749–765.

48. Genius M, Strazzera E. Applying the copula approach to sample selection modelling. Appl. Econ. 2008; 40:1443–1455.

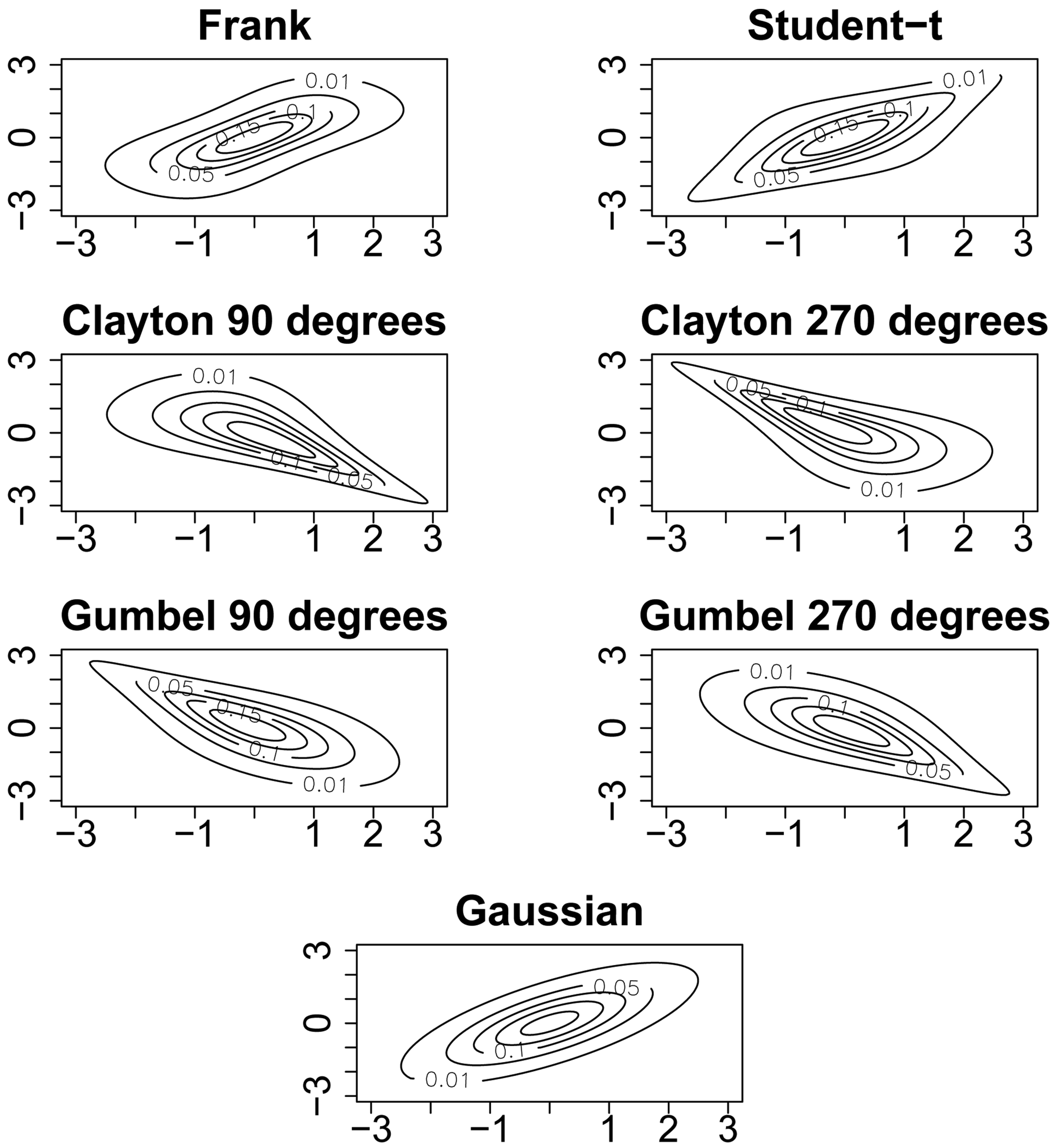

Figure 1.

Illustration of Modelling Dependence Using Copulae. Observations are drawn from the corresponding bivariate distributions with n=1,000 and $\tau = -0.50$. See the eAppendix for the code for drawing from these distributions.

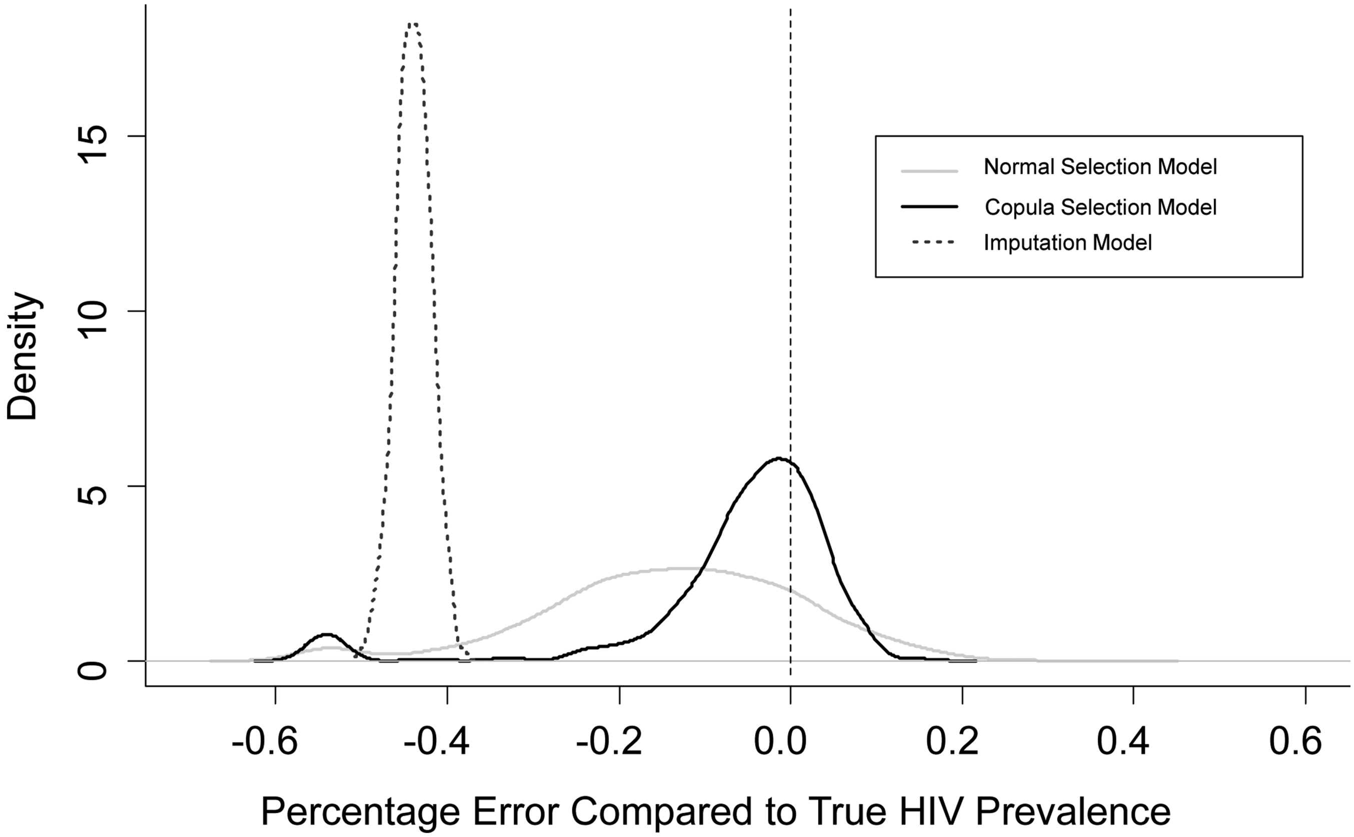

Figure 2.

Simulation Results for HIV Prevalence Estimates with Non-Normal Errors. This scenario illustrates the case with cubed normal errors. The distribution of the proportional error of estimates of HIV prevalence obtained from the normal selection model (Gaussian Copula), a Copula selection model and an imputation model are shown. The simulation is based on the 2007 Zambia Demographic and Health Survey for men, with n=6,500 and 1,000 replications. For each replication, the proportional error for each estimator is calculated as $mean\ (HIV_{Model} - HIV_{True})/HIV_{True}$. The copula model is defined as the copula with the best fit in each replication according to the Akaike Information Criterion (AIC). Errors for the latent variables for consent and HIV status were drawn from a bivariate normal distribution with $mean = 0$ and $\tau = -0.50$, cubed, and then scaled to have mean 0. The mean true HIV prevalence was 21%, observed HIV prevalence (for those with consent=1) was 12%. Consent to be tested was 81%, and the F statistic for interviewer identity was 3.5. The F statistic is calculated as a joint test of significance for interviewer identity in a regression of consent on interviewer identity with the inclusion of the model control variables. See the eAppendix for further details, including the R code for replicating the simulations.

Table 1

Participation Rates for HIV Testing in Demographic and Health Surveys[a]

Demographic and Health Surveys	Participation Rates for Men (%)	Participation Rates for Women (%)
Cote d'lvoire 2005	76	79
Malawi 2004	63	70
Tanzania 2003	77	84
Tanzania 2007	80	90
Zimbabwe 2005	63	76
Lesotho 2004	68	81
Liberia 2007	81	88
Sierra Leone 2008	87	90
Zambia 2007	72	77
Cameroon 2004	90	92
Ethiopia 2005	76	83
Mali 2006	85	93
Niger 2006	84	91
Senegal 2005	75	84
Swaziland 2006	78	87
Rwanda 2005	96	97
Burkina Faso 2003	86	92
Congo 2007	86	90
Ghana 2003	80	89
Guinea 2005	88	92
Kenya 2003	70	76
Kenya 2008	79	86
Mali 2001	76	85
Zambia 2001	73	79

[a] Source: Hogan et al.[3] Data are publically avail able from www.dhsprogram.com.

Table 2

Definition of Copula Functions[a]

Copula	$C(F_1(y_1),F_2(y_2);\theta)$
Normal: C_n	$\Phi_2(\Phi^{-1}(F_1), \Phi^{-1}(F_2); \theta)$
Frank: C_f	$-\theta^{-1}\ln\left(1+\frac{\left(e^{-\theta F_i}-1\right)\left(e^{-\theta F_2}-1\right)}{\left(e^{-\theta}-1\right)}\right)$
Clayton: C_c	$(F_1^{-\theta}+F_2^{-\theta}-1)^{-1/\theta}$
Student: C_t	$t_{2v}(t_v^{-1}(F_1), (t_v^{-1}(F_2); \theta)$
Joe: C_j	$1-((1-F_1)^{\theta}+(1-F_2)^{\theta}-(1-F_1)^{\theta]}(1-F_2)^{\theta})^{1/\theta}$
Gumbel: C_g	$\exp(-((-\log(F_1))^{\theta}+(-\log(F_2))^{\theta})^{1/\theta})$

[a] $t_{2v}(.,.;\theta)$ denotes the cumulative distribution function of a standard bivariate Student-t distribution with correlation coefficient θ and v degrees of freedom. t_v^{-1} denotes the inverse univariate Student-t distribution function with v degrees of freedom.

Table 3

Summary Statistics for Men and Women (Zambia Demographic and Health Survey 2007)[a]

	HIV Prevalence		HIV Test	
	%	**(95% CI)**	**Consented No. (%)**	**Refused No. (%)**
Men	12	(11 to 13)	5098 (79)	1318 (21)
Women	16	(15 to 17)	5625 (80)	1400 (20)

[a] HIV prevalence estimates are based on analysis of respondents who have a valid HIV test and are adjusted for survey design. Non-contacts are excluded. CI = confidence interval.

Table 4

Measures of Association between Consent to be HIV Tested and HIV Status for Men and Women (Zambia Demographic and Health Survey 2007)[a]

	Men			Women		
Copula Model	**Kendall's Tau**	**(95% CI)**	**AIC**	**Kendall's Tau**	**(95% CI)**	**AIC**
Normal	−0.53	(−0.76 to −0.13)	9672.52	−0.19	(−0.47 to 0.12)	11237.27
Frank	−0.58	(−0.72 to −0.24)	9667.97	−0.17	(−0.44 to 0.17)	11237.54
Student T	−0.53	(−0.79 to −0.07)	9675.19	−0.19	(−0.49 to 0.15)	11238.36
Clayton 90	−0.31	(−0.80 to −0.05)	9677.25	−0.13	(−0.60 to −0.02)	11237.64
Clayton 270	−0.71	(−0.84 to −0.53)	9666.31	−0.27	(−0.74 to −0.05)	11237.27
Joe 90	−0.72	(−0.84 to −0.55)	9666.21	−0.28	(−0.74 to −0.05)	11237.28
Joe 270	−0.32	(−0.80 to −0.05)	9678.22	−0.13	(−0.60 to −0.01)	11237.89
Gumbel 90	−0.61	(−0.82 to −0.35)	9670.97	−0.23	(−0.68 to −0.03)	11237.37
Gumbel 270	−0.43	(−0.82 to −0.11)	9676.32	−0.16	(−0.64 to −0.02)	11237.69

[a] Estimates are presented for selection models based on the maximization of model (11), and the copula functions defined in Table 2. The Akaike Information Criterion (AIC) is shown in columns 4 and 8. The selection variable is a series of fixed effects for interviewer identity, of which there are 29 for men and 45 for women. Additional control variables include urban setting, region, interview language, ethnicity, religion, marital status, high−risk sexual behavior in the past year, condom use at last sex, sexually transmitted disease in the past year, tobacco and alcohol use, knowing someone with HIV/AIDS, willingness to care for a family member with HIV/AIDS, and having had a previous HIV test.[3,18] Non−contacts are excluded. Confidence intervals (CI) are adjusted for clustering at the primary sampling unit (PSU) level.

Table 5

HIV Prevalence Estimates for Men and Women (Zambia Demographic and Health Survey 2007)[a]

	Men		Women	
Copula Model	**HIV Prevalence**	**(95% CI)**	**HIV Prevalence**	**(95% CI)**
Normal	20	(13 to 28)	19	(13 to 24)
Frank	21	(15 to 26)	18	(14 to 23)
Student T	21	(13 to 29)	19	(14 to 25)
Clayton 90	19	(7 to 30)	19	(13 to 25)
Clayton 270	21	(16 to 25)	18	(14 to 22)
Joe 90	21	(16 to 25)	18	(14 to 22)
Joe 270	19	(8 to 31)	19	(12 to 26)
Gumbel 90	21	(14 to 27)	18	(14 to 23)
Gumbel 270	20	(10 to 30)	19	(13 to 25)

[a] HIV prevalence is based on individuals who have a valid HIV test and predicted HIV status from selection models based on the maximization of model (11), and the copula functions defined in table 2. The selection variable is a series of fixed effects for interviewer identity, of which there are 29 for men and 45 for women. Additional control variables include urban setting, region, interview language, ethnicity, religion, marital status, high-risk sexual behavior in the past year, condom use at last sex, sexually transmitted disease in the past year, tobacco and alcohol use, knowing someone with HIV/AIDS, willingness to care for a family member with HIV/AIDS, and having had a previous HIV test.[3,18] Non-contacts are excluded. Confidence intervals (CI) are adjusted for clustering at the primary sampling unit level and prevalence estimates are weighted. The preferred model according to the Akaike Information Criterion (AIC) is Joe 90 for men, and the Normal and Clayton 270 models for women.