Joint models for mixed categorical outcomes: a study of HIV risk perception and disease status in Mozambique

LOQUIHA, Osvaldo; HENS, Niel; Martins-Fonteyn, Emilia; Meulemans, Herman; Wouters, Edwin; Temmerman, Marleen; Osman, Nafissa & AERTS, Marc (2017)
Joint models for mixed categorical outcomes: a study of HIV risk perception and disease status in Mozambique. In: Journal of Applied Statistics, 45 (10),p. 1781-1798.

DOI: 10.1080/02664763.2017.1391184
Handle: http://hdl.handle.net/1942/25106

Joint models for mixed categorical outcomes: A study of HIV risk perception & disease status in Mozambique

Osvaldo Loquiha[a,b], Niel Hens[b,c], Emilia Martins-Fonteyn[d], Herman Meulemans[d], Edwin Wouters[d], Marleen Temmerman[e,f], Nafissa Osman[g,h], Marc Aerts[b]

[a] Department of Mathematics and Informatics, Universidade Eduardo Mondlane, Av. Julius Nyerere 3453, Campus, Maputo, Mozambique;
[b] Interuniversity Institute for Biostatistics and statistical Bioinformatics (I-BioStat), Universiteit Hasselt, Agoralaan 1, B-3590 Diepenbeek, Belgium;
[c] Centre for Health Economics Research and Modeling Infectious Diseases and Centre for the Evaluation of Vaccination, Vaccine & Infectious Disease Institute, University of Antwerp, Universiteitsplein 1, B2610 Wilrijk, Belgium;
[d] Department of Sociology, University of Antwerp, City Campus, Prinsstraat 13, B-2000 Antwerp, Belgium;
[e] International Centre for Reproductive Health, Ghent University, De Pintelaan 185 P3, 9000 Ghent, Belgium;
[f] Centre of Excellence Women and Child Health, Aga Kan University, Nairobi, Kenya;
[g] Department of Obstetrics and Gynaecology, Maputo Central Hospital, Av. Agostinho Neto, Maputo, Mozambique;
[h] Faculty of Medicine, Eduardo Mondlane University, Av. Salvador Allende 702, Maputo, Mozambique.

ARTICLE HISTORY
Compiled June 14, 2017

ABSTRACT
Two types of bivariate models for categorical response variables are introduced to deal with special categories such as "unsure" or "unknown" in combination with other ordinal categories, while taking additional hierarchical data structures into account. The latter is achieved by the use of different covariance structures for a trivariate random effect. The models are applied to data from the INSIDA survey, where interest goes to the effect of covariates on the association between HIV risk perception (quadrinomial with an "unknown risk" category) and HIV infection status (binary). The final model combines continuation-ratio with cumulative link logits for the risk perception, together with partly correlated and partly shared trivariate random effects for the household level. The results indicate that only age has a significant effect on the association between HIV risk perception and infection status. The proposed models may be useful in various fields of application such as social and biomedical sciences, epidemiology and public health.

KEYWORDS
Bivariate categorical data; Continuation-ratio logits; HIV infection status; Mixed models; Perceived risk of HIV.

CONTACT Osvaldo Loquiha. Email: osvaldo.loquiha@uem.mz

1. Introduction

There are many settings where a combination of binary, ordinal and even continuous response variables occurs, and often a joint analysis of the responses offers more insights than separate analyses. Consider our motivational case where perception of risk for HIV infection (no,low,high, do not know) and HIV test status (positive, negative) are recorded for each individual in a household. These outcomes play an important role in the "health belief" model which emphasizes that an individual empowered with health information is more likely to comply with preventive or curative behaviours [5]. Self-perceived HIV risk is considered an integral component in motivating avoidance of HIV risk, and the congruency between self-perception of HIV risk and reported risk-taking behaviours might be especially important in the likelihood to engage in self-protective behaviours, such as condom use and uptake of HIV testing [30]. On the other hand, there is evidence that people who are aware of their HIV status can adopt practices to reduce HIV transmission and to access effective treatment [18].

While the association between perception of risk for HIV and disease status, and HIV related risk behaviours is well documented [1, 13, 29, 33, 34], it seems that this relation is complex and yet inconclusive. Koh and Yong [19] showed that for individuals with risky sexual behaviour there is a positive association between perception of risk and HIV status, i.e., the higher the perceived risk the higher the odds of a positive HIV status. However, another study reported a negative association such that risky groups tend to perceive no or low risk of HIV infection in contexts where the disease prevalence is high [22, 28].

The National Survey of Prevalence, Risk Behavioural and Information about HIV and AIDS of 2009 in Mozambique (INSIDA) allows us to study the joint distribution of the perception of risk for HIV infection and the HIV infection status as a function of several covariates on the individual level (such as age, gender, etc) as well as on the household level (such as household size, etc). Statistical models that jointly analyze perception of risk and HIV status provide more insights into the association between perception and HIV status and can be used, e.g. to test whether the association is homogeneous across the ordinal scale of perception, while accounting for the characteristics at the individual and at the household level. These models should reflect particular features of perception and disease status, such as that both are categorical and that perception of risk is typically a semi-ordinal variable, measured on a scale that usually contains a "do not know" or "unsure" risk category, next to an ordinal scale such as no, low, moderate and high risk, while disease status is a dichotomous variable.

Dale [9] introduced a bivariate logistic model that uses the bivariate Plackett distribution to specify the joint distribution with global cross-product ratios as marginal association measures. A global cross-product ratio measure results from a cumulative link function and is useful when variables are ordered. Molenbergs and Lesaffre [26] extended the Dale model to multivariate ordinal outcomes using a multivariate Plackett distribution. Glonek and McCullagh [15] provided a general definition of the class of regression models relating the joint distribution of categorical responses to predictors based on the multivariate logistic transform introduced by [23]. This multivariate logistic transform is a reparameterisation of cell probabilities

in terms of marginal logistic contrasts. An arbitrary set of logistic contrasts may however not correspond to a valid joint distribution. For that reason, [31] presented an efficient algorithm for detecting whether or not the inverse transform exists, and for computing it if it does.

McMillan and Hanson [25] presented a `SAS macro` that fits a bivariate Dale model when the second variable is meaningless if the first variable equals a special value (e.g. one cannot specify the number of drinks per drinking occasion if one never has any drinking occasions). The approach used by [25] consists of two parts: one with a dichotomous outcome (e.g.: drinking yes/no) and a second part, conditional on the dichotomous outcome to be positive, where both outcomes are ordinal (frequency and quantity of drinking). In our setting, as explained further, the first variable can also take a special value "unsure" or "unknown" and if "known", it is of an ordinal nature, turning it into a semi-ordinal type. But the second variable is meaningful, regardless of whether the first one is known or not.

Here we consider two particular multivariate logit transforms defining regression models for the joint distribution of the perception of HIV risk (semi-ordinal) and the HIV infection status (binary). They differ in their construction of the ordinal part of the risk perception variable following a continuation-ratio logits or a cumulative logits model, and the corresponding odds ratios when cross-tabulating risk perception with infection status. The continuation-ratio logit model is best suited when each category of the response variable is of intrinsic interest and when they are seen as levels of achievement that can only be entered if the previous was achieved [2, 8]. Probabilities such as "unknown" vs "known" risk, and conditional probabilities "no risk" vs "some risk" (if "known" risk), and "low risk" vs "high risk" (if "some risk"), can be modelled by a continuation-ratio logit model without any need for a two-part analysis to deal with the special "unknown" category. In a similar way a cumulative logits approach for the ordinal categories no, low, high risk following the first contrast "unknown" vs "known" risk is a very natural alternative modelling strategy, reflecting the discretization of a latent continuous risk perception variable.

Another intrinsic feature of the INSIDA data is its hierarchical nature with households as natural clusters resulting in correlated data for respondents from the same household. A typical strategy to account for this hierarchical structure in a multivariate setting is the use of multivariate random effects [12, 16, 32] We will extend our bivariate marginal model for HIV risk perception and infection status with different multivariate structures for the random effects distribution, including independent, shared and correlated random effects.

Relevant information on the INSIDA survey is summarized in the following section. The marginal model for the joint distribution of HIV risk perception and infection status is formulated in Section 3 while its extension with household specific random effects is described in Section 4. Section 5 summarizes and discusses the results of the application of the different models to the INSIDA data. The paper ends with a final discussion in Section 6.

2. The INSIDA survey

The 2009 INSIDA survey is a cross-sectional two-stage survey on households, carried out by the National Institute of Health in collaboration with the National Bureau of Statistics of Mozambique. It was the first survey designed to collect comprehensive data on the prevalence of HIV infection, knowledge, attitude, behaviour risk factors and access to information on HIV and AIDS in the Mozambican population [17].

It was designed to interview 6232 households, 10800 men and women aged 15 to 64 years, 1770 teenagers aged 12 to 14 and 4300 children from 0 to 11 years old. It was also expected that 13600 individuals would be tested for HIV. The actual number of participating households was 6097 with a total of 14964 individuals. For our objectives we consider the subset of individuals (men and women) aged 15-64 years, being sexual active in the last 12 months prior to the survey (according to the sexual activity variable in the survey). Not all households had eligible individuals as to this definition, resulting in a total of 10548 individuals from 5573 households. Data on HIV risk perception and infection status from 730 households were complete, in the sense that data on both response variables were available for all members of the households; for 83 households all members had data on one response variable, and 4760 households had at least one member lacking the value for both response variables. Table 1 lists the variables of interest with some basic descriptive information.

TABLE 1 ABOUT HERE

Perception of risk of contracting HIV was a general HIV risk assessment based on the question: "Do you think your chances of getting AIDS are small, moderate, great, or no risk at all?". Possible answers were: "none", "small", "moderate", "great", "HIV+" and "do not know". Participants who responded "HIV+" (1.1%) where excluded, since they were no longer at risk of HIV infection. The categories "do not know", "none" and "small" were relabeled as "unknown", "no" and "low" respectively. When missing data and covariates are taken into account, the resulting multidimensional contingency table with all 5 categories for perception of risk had about 20% cells (out of 960) with 0 frequency for moderate risk, while the other risk categories had less than 8% cells with 0 frequency. This led to separation and thus convergence problems for the joint modelling of the outcomes with all 5 categories for perception of risk and therefore the two categories "moderate" and "great" risk were combined into one "high"-category.

The HIV infection status was a binary variable derived from an HIV test: 1 if positive and 0 otherwise. The HIV test was done posterior to the interview using Dried Blood Samples (DBS) and two sequential ELISA tests. After counseling and consent, household members would then receive their test results.

Risky behaviour was assessed based on a combination of three risky behavioural patterns: multi-sexual partners during the last 12 months, inconsistent condom use during the last three sexual intercourses with any partners, and having had an STD during the last 12 months [11, 22]. From the answers to the corresponding yes/no questions, a "global" ordinal measure of risky sexual behaviour RSB, ranging from 0 (no risk if all negative) to 3 (high risk if all positive) was defined.

Socio-demographic characteristics such as place of residence (urban, rural), sex

of respondent (male, female), age (grouped) and knowledge of at least one way to reduce HIV infection were also included as covariates [13, 29]. Place of residence accounts for socio-economic inequalities between rural and urban areas on perception of risk of HIV and HIV prevalence. Knowledge of at least one way to reduce HIV infection was considered as a proxy for awareness of risk behaviour that leads to HIV contamination.

Table 1 also shows that the percentage of missing observations in the response variables is 8.1% for the infection status and 1.4% for the risk perception. It is quite low for the different covariates (from 0% to 3.7%) except for condom use (13.6%). This relatively higher percentage for condom use, as compared for instance to the very low percentage (0.04%) for the number of sexual partners (last 12 months) is rather ascribed to the ambiguity of what is perceived as consistent or inconsistent use than to particular non-ignorable mechanisms. Although a complete case analysis implies a sample size reduction of about 25%, we decided to not focus on any more advanced missing data methods but rather on the development of the joint models, as the final dataset still contains observations of 7944 individuals within 4912 households.

3. The marginal joint model for HIV risk perception & infection status

Assume that there are $i = 1, \ldots, N$ households and $j = 1, \ldots, n_i$ individuals in each household. So n_i is the number of eligible household members (ranging from 1 to 23, see Table 1). Let $\boldsymbol{Y}_{ij} = (Y_{ij1}, Y_{ij2})^T$ be a random vector of categorical outcomes: Y_{ij1} the perception of risk for HIV infection with 4 categories, coded as 0 to 3, corresponding to "unknown", "no","low" and "high" risk respectively, and Y_{ij2} the binary infection status (0=HIV negative, 1=HIV positive). Let $\boldsymbol{X}_{ij} = (X_{ij0}, X_{ij1}, \ldots, X_{ijq})^T$ denote the vector containing q covariates of interest with $X_{ij0} = 1$ for all (i, j). The majority of the covariates are at the individual level, some of them however or on the household level and do not need a j subscript.

The multinomial probabilities (suppressing the ij subscript for simplicity) $\mathrm{P}(Y_1 = k, Y_2 = \ell)$ with $k = 0, 1, 2, 3$ and $\ell = 0, 1$ corresponding to the 4×2 contingency table define the joint probability distribution of the response vector $\boldsymbol{Y} = (Y_1, Y_2)^T$. Based on the multinomial log-likelihood, the maximum likelihood (ML) estimates can be computed. Instead of modelling the joint probabilities it is often of more interest to reparametrize the (log-)likelihood in terms of 7 out of 8 non-redundant parameters related to the marginal distribution of Y_1 (3 parameters), the "success" probability $\pi = \mathrm{P}(Y_2 = 1)$, and 3 association parameters. The choices made for these 7 parameters should reflect the nature of variable Y_1 with the special category "unknown" (coded as 0) and the ordinal categories coded as 1,2,3.

Here we opted for two natural but different parametrization: the continuation-ratio logits (CR) approach and a combination with a cumulative logits/proportional odds (CR-PO) model. As depicted in Figure 1 (left upper graph), the continuation-ratio approach follows a particular but in this case natural sequential pattern and is based on the odds $\mathrm{P}(Y_1 = 0)/\mathrm{P}(Y_1 > 0)$ corresponding to not knowing about one's risk, $\mathrm{P}(Y_1 = 1|Y_1 \geq 1)/\mathrm{P}(Y_1 > 1|Y_1 \geq 1)$ corresponding to perceiving "no risk" given that one knows about one's risk, $\mathrm{P}(Y_1 = 2|Y_1 \geq 2)/\mathrm{P}(Y_1 > 2|Y_1 \geq 2)$ corresponding to perceiving low rather than high risk, given that one perceives some risk. So, also suppressing the conditioning in the notation, the CR marginal distribution is

determined by:

$$\text{Odds}_k^{\text{CR}} = \text{P}(Y_1 = k)/\text{P}(Y_1 > k), \quad k = 0, 1, 2.$$

Implied by the CR structure, the three association parameters are defined by the OR's of the corresponding 2×2 tables formed by cross classifying the CR structure of Y_2 with Y_1 (see Figure 1, left lower graph):

$$\text{OR}_k^{\text{CR}} = \frac{\text{P}(Y_1 = k, Y_2 = 0)\text{P}(Y_1 > k, Y_2 = 1)}{\text{P}(Y_1 > k, Y_2 = 0)\text{P}(Y_1 = k, Y_2 = 1)}, \quad k = 0, 1, 2,$$

reflecting for $k = 0$ how the odds to be HIV infected depends on knowing or not knowing about one's risk, for $k = 1$ how it depends on perceiving risk yes or no, and for $k = 2$ how it depends on perceiving high or low risk.

The combined CR-PO parameterisation (see Figure 1, right upper graph) for

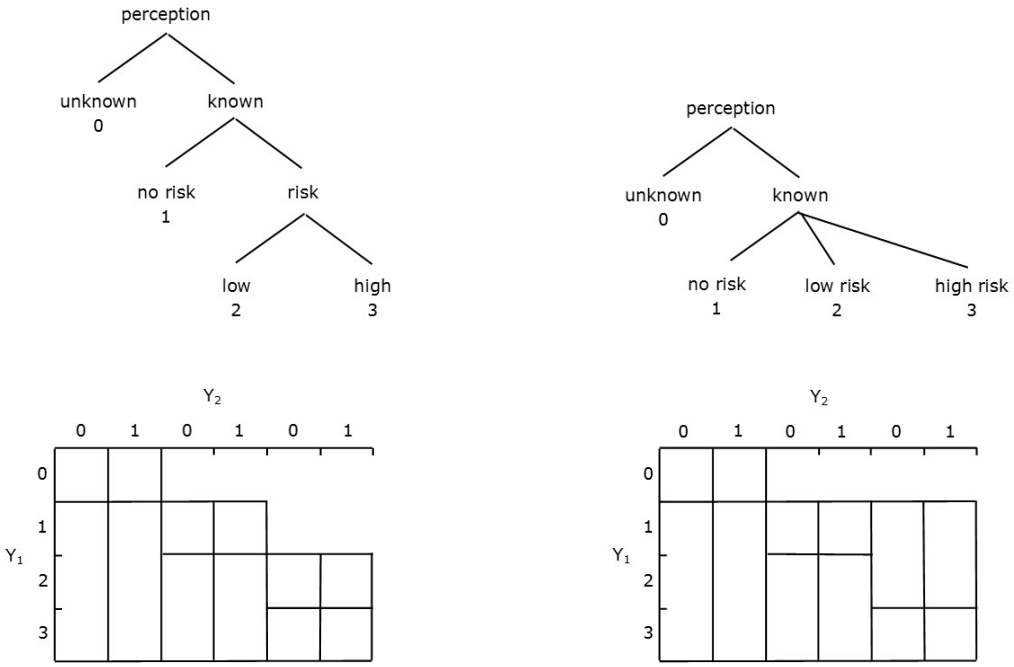

Figure 1. Graphical presentation of the continuation ratio (CR) and cumulative-proportional odds model (CR-PO). Upper panels show the structure of the three odds representing the partly ordinal structure of the HIV risk perception Y_1: on the left the CR model, and on the right the first logit as in the CR model, and the remaining ordinal part as a PO model (CR-PO model). Lower panels show the three 2×2 tables combining the CR (left) and CR-PO (right) structure with the HIV infection status Y_2.

the marginal distribution of Y_2 starts again with $\text{P}(Y_1 = 0)/\text{P}(Y_1 > 0)$, but then, given that $Y_1 > 0$, proceeds with the "cumulative" logits, again suppressing the conditioning on $Y_1 > 0$:

$$\text{Odds}_k^{\text{PO}} = \text{P}(Y_1 \leq k)/\text{P}(Y_1 > k), \quad k = 1, 2.$$

and implied association parameters $\mathrm{OR}_0^{\mathrm{CR}}$ and

$$\mathrm{OR}_k^{\mathrm{PO}} = \frac{\mathrm{P}(Y_1 \leq k, Y_2 = 0)\mathrm{P}(Y_1 > k, Y_2 = 1)}{\mathrm{P}(Y_1 > k, Y_2 = 0)\mathrm{P}(Y_1 \leq k, Y_2 = 1)}, \quad k = 1, 2.$$

Note that, because of the conditioning on $Y_1 > 0$, $\mathrm{Odds}_1^{\mathrm{CR}} = \mathrm{Odds}_1^{\mathrm{PO}}$ and that $\mathrm{OR}_1^{\mathrm{CR}} = \mathrm{OR}_1^{\mathrm{PO}}$ and consequently only the 2 parameters related to $k = 2$ make both parameterisations different. $\mathrm{OR}_2^{\mathrm{PO}}$ quantifies the dependency between being HIV infected and on perceiving high or no high (low or no) risk.

The above defined odds for Y_1, the probability π for Y_2 and the odds ratio parameters relating Y_1 and Y_2 can modelled as function of the covariates $\boldsymbol{X}_{ij}$, using logit and log link functions as follows (suppressing any CR or PO superscript, but reintroducing ij subscripts). For the marginal distribution of Y_{ij1}, the generalized logit model is given by

$$\log(\mathrm{Odds}_{k,ij}) = \boldsymbol{X}_{ij}\boldsymbol{\beta}_{1k}, \quad k = 0, 1, 2, \tag{1}$$

for the probability to be HIV positive of Y_{ij2}, the logit model is given by

$$\mathrm{logit}(\pi_{ij}) = \boldsymbol{X}_{ij}\boldsymbol{\beta}_2, \tag{2}$$

and for the dependency parameters between Y_1 and Y_2, the model is defined as

$$\log(\mathrm{OR}_{k,ij}) = \boldsymbol{X}_{ij}\boldsymbol{\beta}_{3k}, \quad k = 0, 1, 2, \tag{3}$$

where $\boldsymbol{\beta}_{1k}^T = (\beta_{10,k}, \beta_{11,k}, \ldots, \beta_{1q,k})$, $\boldsymbol{\beta}_2^T = (\beta_{20}, \beta_{21}, \ldots, \beta_{2q})$ and $\boldsymbol{\beta}_{3k}^T = (\beta_{30,k}, \beta_{31,k}, \ldots, \beta_{3q,k})$ are vectors of regression parameters. Of course, not necessarily the same covariates play a role for different parameters.

In the mixed CR-PO approach, the "known" perception categories 1 to 3 of Y_1 can be seen as a categorisation of a latent continuous degree of perception having a logistic distribution, which is an appealing concept and which motivates the use of common slopes across the CR-logits (resulting in the "proportional odds" PO model). Common slopes can also be considered for the CR model, at least again for logits related to the "known" perception categories. The intercepts $\beta_{10,k}$ are taken differently for $k = 0, 1, 2$ for both type of models and for the PO part of the model with the additional constraint $\beta_{10,1} \leq \beta_{10,2}$ (following from the cumulative nature). But it can be examined whether the slopes (or part of them) $\beta_{11,k}, \ldots, \beta_{1q,k}$ are not depending on k, thus reducing the number of parameters considerably.

The model defined by (1)-(3) will be referred to as the "marginal bivariate" model. This model takes into account the semi-ordinal nature of Y_1 and allows to study the dependency between perception of risk (Y_1) and HIV status (Y_2). It ignores however the hierarchical data structure and corresponding intra-household correlation. Indeed, one might expect the observations of $\boldsymbol{Y}_{ij}$ for two individuals within a household to be more alike than for two individuals from different households, even after adjusting for covariates at the household level. Approaches using different random effects structures are proposed in the next section.

4. The mixed effects joint model for HIV risk perception & infection status

To model the heterogeneity across households, or equivalently the intra-household association, models (1), (2) and (3) can be extended with random household effects as follows:

$$\log(\text{Odds}_{k,ij}) = \boldsymbol{X}_{ij}\boldsymbol{\beta}_{1k} + b_{1ki}, \quad k = 0, 1, 2, \tag{4}$$

$$\text{logit}(\pi_{ij}) = \boldsymbol{X}_{ij}\boldsymbol{\beta}_{2} + b_{2i}, \tag{5}$$

and

$$\log(\text{OR}_{k,ij}) = \boldsymbol{X}_{ij}\boldsymbol{\beta}_{3k} + b_{3ki}, \quad k = 0, 1, 2, \tag{6}$$

with a total of 7 random effects

$$\boldsymbol{b}_i = ((b_{10i}, b_{11i}, b_{12i}), b_{2i}, (b_{30i}, b_{31i}, b_{32i})), \tag{7}$$

having a joint zero-mean normal distribution. A fully unstructured 7-dimensional multivariate random effects parameterisation however was computationally not feasible, and therefore simplified structures were considered, with shared random effects for each type of parameter $\log(\text{Odds}_{k,ij}), \text{logit}(\pi_{ij})$ and $\log(\text{OR}_{k,ij})$

$$\boldsymbol{b}_i = ((b_{1i}, \gamma_{11}b_{1i}, \gamma_{12}b_{1i}), b_{2i}, (b_{3i}, \gamma_{31}b_{3i}, \gamma_{32}b_{3i})), \tag{8}$$

defined by scalar scale parameters $\gamma_{11}, \gamma_{12}, \gamma_{31}, \gamma_{32}$ and random effects

$$(b_{1i}, b_{2i}, b_{3i}) \sim N_3(\mathbf{0}, \boldsymbol{\Sigma}). \tag{9}$$

The use of shared random effects is a well-established method for reducing the dimension of the random effects structure and obtaining more parsimonious models [see 24]. The CR-PO model needs the additional constraint $\beta_{10,1} + \gamma_{11}b_{1i} \leq \beta_{10,2} + \gamma_{12}b_{1i}$ to make sure that the cumulative probabilities are ordered appropriately. The condition that $\gamma_{11} \leq \gamma_{12}$ is sufficient but not necessary. For any practical purposes it is sufficient to constrain the parameter estimates such that

$$\beta_{10,1} + 4\gamma_{11}\sigma_1 \leq \beta_{10,2} + 4\gamma_{12}\sigma_1, \tag{10}$$

as $P(|b_{1i}| < 4\sigma_1) > 0.9999$.This constraint would actually become more problematic for the most complex random effects structure (7). However, if (10) is not satisfied, negative probabilities can be obtained which may require redefinitions on the estimation algorithms (e.g. step-halving). On the other hand violations of (10) could indicate that the CR-PO model with random effects structure (8) does not provide a good fit to the data and should be dropped. Fortunately, we did not face such problems with the models presented in this paper.

Different choices considered for the covariance matrix $\boldsymbol{\Sigma} = \boldsymbol{V}\boldsymbol{R}\boldsymbol{V}$ with

$$\boldsymbol{V} = \begin{pmatrix} \sigma_1 & 0 & 0 \\ 0 & \sigma_2 & 0 \\ 0 & 0 & \sigma_3 \end{pmatrix}$$

are based on different choices for the correlation matrix $\boldsymbol{R}$. Matrix $\boldsymbol{V}$ implies that the random effects vector $\boldsymbol{b}_i$, as given by (7) and (8), has variance components

$$\text{varcomps}(\boldsymbol{b}_i) = ((\sigma_1^2, \gamma_{11}^2\sigma_1^2, \gamma_{12}^2\sigma_1^2), \sigma_2^2, (\sigma_3^2, \gamma_{31}^2\sigma_3^2, \gamma_{32}^2\sigma_3^2)), \tag{11}$$

allowing flexibility of different scales of variation over households for different parameters, in the same direction (positive γ's) or in the opposite direction (negative γ's).

A fully parameterized $\boldsymbol{R}$ matrix with three correlations $\rho_{k\ell} = \text{cor}(b_{ki}, b_{\ell i})$, $1 \leq k < \ell \leq 3$ appeared to be computationally not feasible. The following simplified structures are considered:

Independent RE-model

This model with independent random effects (b_{1i}, b_{2i}, b_{3i}) corresponds to the choice $\boldsymbol{R}_{\text{IND}} = I_3$, the identity matrix.

Shared RE-model

This model with all random effects (b_{1i}, b_{2i}, b_{3i}) shared corresponds to the choice $\boldsymbol{R}_{\text{SHARED}} = J_3$, the all-ones matrix, and implying one random effect $b_{1i} \sim N(0, \sigma_1^2)$ and $b_{2i} = (\sigma_2/\sigma_1)b_{1i}$ and $b_{3i} = (\sigma_3/\sigma_1)b_{1i}$. The vector of variance components remains the same as in (11).

Correlated RE-models

A first choice

$$\boldsymbol{R}_{\text{COR}_1} = \begin{pmatrix} 1 & \rho & 1 \\ \rho & 1 & \rho \\ 1 & \rho & 1 \end{pmatrix} \tag{12}$$

corresponds to two correlated random effects (b_{1i}, b_{2i}) and shared effects $b_{3i} = (\sigma_3/\sigma_1)b_{1i}$. Consequently (b_{2i}, b_{3i}) are correlated too, with the same correlation coefficient ρ.

A second choice

$$\boldsymbol{R}_{\text{COR}_2} = \begin{pmatrix} 1 & 1 & \rho \\ 1 & 1 & \rho \\ \rho & \rho & 1 \end{pmatrix} \tag{13}$$

corresponds to two correlated random effects (b_{1i}, b_{3i}) and shared effects $b_{2i} = (\sigma_2/\sigma_1)b_{1i}$; and so (b_{2i}, b_{3i}) are correlated too.

And a similar third choice

$$\boldsymbol{R}_{\text{COR}_3} = \begin{pmatrix} 1 & \rho & \rho \\ \rho & 1 & 1 \\ \rho & 1 & 1 \end{pmatrix} \tag{14}$$

corresponds to two correlated random effects (b_{1i}, b_{2i}) and shared effects $b_{3i} = (\sigma_3/\sigma_2)b_{2i}$, implying that also (b_{1i}, b_{3i}) are correlated with correlation ρ.

Maximum likelihood estimation

Let $a_{ij,k\ell}$ represent the cell count corresponding to cell $(Y_{ij1} = k, Y_{ij2} = \ell)$ with $k = 0, 1, 2, 3$ and $\ell = 0, 1$, and $\mu_{ij,k\ell} = \text{P}(Y_{ij1} = k, Y_{ij2} = \ell | \boldsymbol{X}_{ij}, \boldsymbol{\beta}_{1k}, \boldsymbol{\beta}_2, \boldsymbol{\beta}_{3k}, \boldsymbol{b}_i)$ the respective cell probability. For the multinomial distribution, the kernel of the likelihood function in terms of $a_{ij,k\ell}$ and $\mu_{ij,k\ell}$ is given as

$$L = \prod_{i=1}^{N} \int_{\boldsymbol{b}_i} \left\{ \prod_{j=1}^{n_i} \mu_{ij,k\ell}^{a_{ij,k\ell}} \right\} f(\boldsymbol{b}_i) d\boldsymbol{b}_i \tag{15}$$

with $f(\boldsymbol{b}_i)$ the Gaussian density function for $\boldsymbol{b}_i$ with mean $\boldsymbol{0}$ and variance $\boldsymbol{\Sigma}$. Maximization of (15) is subject to the constraints that $\sum_{k,\ell} \sum_j^{n_i} a_{ij,k\ell} = n_i$ and $\sum_{k,\ell} \mu_{ij,k\ell} = 1$. We approximate the integral in (15) using numerical integration carried out with adaptive quadratures, implementing 7 or more quadrature points. The maximum likelihood estimators for $\boldsymbol{\Theta} = \{\boldsymbol{\beta}_{1k}, \boldsymbol{\beta}_2, \boldsymbol{\beta}_{3k}, \boldsymbol{\Sigma}\}$ were then obtained via the Newton–Raphson iterative algorithm

$$\boldsymbol{\Theta}^{(t+1)} = \boldsymbol{\Theta}^{(t)} + \boldsymbol{H}(\boldsymbol{\Theta}^{(t)})^{-1} \boldsymbol{U}(\boldsymbol{\Theta}^{(t)})$$

with the score function $\boldsymbol{U}(\boldsymbol{\Theta}) = \partial \log L / \partial \boldsymbol{\Theta}$ and the Hessian matrix $\boldsymbol{H}(\boldsymbol{\Theta}) = \partial^2 \log L / \partial \boldsymbol{\Theta} \partial \boldsymbol{\Theta}'$ Parameter estimation was implemented in `PROC NLMIXED` of `SAS/STAT` software, version 9.2 (illustrative code available as supplementary material).

5. Application to the INSIDA survey

In this section we briefly report on the application of both type of models (CR and CR-PO) on the INSIDA data, the selection of covariates and the covariance structure of the random effects, and discuss and interpret the estimates as obtained from the best fitting model.

Model building and model selection

Due to the computational complexity, extensive model building was not feasible for the mixed effects joint models, contrary to the marginal models which ran without any difficulties with both model parametrization. As a pragmatic approach, the optimal set of covariates was first selected for the three model components (1)-(3) for each type of marginal joint model (continuation ratio & cumulative logit models) using a

backward selection procedure and then this set was investigated for the possibility of common slopes and random effects.

The final set of covariates is listed in Tables 3 and 4. A likelihood ratio (LR) test was applied to check the possibility to have common slopes for the ordinal logits, resulting in LR=13, df=10 (p-value=0.2237) and LR=15, df=10 (p-value=0.1321) for the CR and CR-PO models respectively. Table 2 below compares the fit of the different models in terms of their values for -2×loglikelihood, Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC). The last two columns show the ranking of all models according to AIC and BIC. Both rankings are very close and show that the marginal models on position 8 and 9 are inferior to most of the mixed effects models, confirming the need to include household random effects. So we focus discussion on the latter type of models.

The model outperforming all others is the CR-PO model 6 with correlation matrix (14) for the random intercepts. Within the CR-PO family the second best model (and overall second best model according to BIC) is model 3 with independent random intercepts, closely followed by the CR and CR-PO model 4 with correlation matrix (12) for the random intercepts.

TABLE 2 ABOUT HERE

A note of cautiousness however is in place here. The fits of the CR models 5 and 6 and CR-PO model 5 (indicated with an asterisk in Table 2) gave the `SAS/STAT` warning "hessian matrix not positive definite" implying that estimates for the standard errors of some parameter estimates are not available. Applying different starting values and exploring other options to improve the numerical computations were not successful. The most common reason for this `SAS/STAT` warning relates to scaling issues or misspecified or overspecified models. In our view, we are hitting the limits of identifiability and overspecification with some of our models. Results of these models were excluded of further analysis in this paper.

Parameter estimates and interpretation

Table 3 and 4 show the estimates of the best CR model 4 and the best CR-PO model 6 (being the overall best one). The estimates for the parameters and standard errors for the marginal distribution of the HIV risk perception, the infection status and the assocation between both are quite similar between both models. Obviously, as the covariance structures of the random effects for both models are different, so are the estimates for the variance components.

TABLE 3 ABOUT HERE

Further discussion is focused on CR-PO model 6. First note that the model constraint (10) is satisfied as $\hat{\beta}_{10,1} + 4\hat{\gamma}_{11}\hat{\sigma}_1 = 5.73 \leq \hat{\beta}_{10,2} + 4\hat{\gamma}_{12}\hat{\sigma}_1 = 6.36$. The odds to not know the risk is about 48% higher for households in the rural areas, 112% higher for women, 17% higher for the older individuals (age ≥ 45) as compared to the youngest ones (age 15-24), and 25% higher for a unit increase on the risky sexual behaviour scale. Having knowledge on how to reduce risk of HIV infection reduces the odds to not know the risk with about 50%. Given "known" risk perception, the odds to perceive no risk (versus low or high risk), as well as the odds to perceive low risk versus high risk, is about 97% higher for households in the rural areas and 43% higher for the older

individuals (age$\geq$ 45) as compared to the youngest ones (age 15-24). The odds of no risk is about 29% lower for women, 43% lower when having knowledge how to reduce HIV risk, and 41% lower for any unit increase on the risky sexual behaviour scale.

TABLE 4 ABOUT HERE

For the HIV infection status, the odds to be positive is about 61% lower for households in the rural areas and 68% higher for women. The odds to be positive is highest for the age group 25-34, being 166% higher as compared to the youngest ones (age 15-24); and 116% higher for the age group 35-44 and 60% higher for the age group ≥ 45, both as compared to the youngest group. Finally, the odds increases 26% with any unit increase on the risky sexual behaviour scale. The effect of having knowledge on how to reduce HIV risk was not significant.

Of the covariates, age was the only factor with a significant effect on the odds ratio models (6), more precisely for the odds ratio in the 2×2 table cross classifying "unknown"/"known" risk perception against positive/negative infection status. For the youngest age-group (15-24) there is a significant "positive" association with an odds ratio estimate of 1.97 (95% C.I.:1.16 –3.35), implying that the odds to be HIV positive were 97% higher when one does not know about his/her risk, as compared to "known" risk. This association is only significantly different for the age-group 25-34, for which the odds ratio estimate reduces to 0.98, implying a lower odds when risk is unknown than when it is known. As there was no evidence of an effect of risky sexual behavior on the association between HIV risk perception and HIV status, our study does not confirm the earlier (contradictory) findings of [19] or [28]. However, since $\beta_{30,0} > \beta_{30,1}$ the odds of HIV infection is significantly higher for individuals with "unknown" risk than with "known" risk, and it increases with an increase on the risk perception scale (since $\beta_{30,2} > \beta_{30,1}$), the latter suggesting a positive association between risk perception and HIV status though not statistically significant.

The above interpretations of the fixed effects are conditional on household. The heterogeneity across households is characterized by the variance components. The effect of household seems to mainly play at the level of the probability to have a HIV positive status, with estimated variance $\hat{\sigma}_2 = \sqrt{2.421} = 1.556$ (p-value < 0.001 using a $\chi^2_{0,1}$mixture). Using the approximate formula $\beta^M = \beta^{RE}/\sqrt{c^2\sigma^2_{RE} + 1}$ with $c = 16\sqrt{3}/15\pi$ between marginal effect β^M and conditional effect β^{RE} in a random intercept logistic model with random effect variance σ^2_{RE} [10, 27], we get the following unconditional, population averaged effects: the odds to be HIV positive is about 50% lower for households in the rural areas and 47% higher for women; is highest for the age group 25-34, being 106% higher as compared to the youngest ones (age 15-24), 76% higher for the age group 35-44 and 41% higher for the age group ≥ 45, both as compared to the youngest group; increases by 18% with any unit increase on the risky sexual behaviour scale. An estimate of an approximate intra-household correlation (correlation between HIV infection status between members of the same household) can be calculated as $\hat{\sigma}^2_2/(\hat{\sigma}^2_2 + \pi^2/3) = 0.42$. However, the basis for this calculation is not very strong and caution is needed with its use [20], but it gives an idea of the order of magnitude.

The estimated correlation between the random effects b_{1i} and b_{2i} was found to be significantly negative, implying that (since $\hat{\gamma}_{11} \geq 0$ and $\hat{\gamma}_{12} \geq 0$) for a household

whose members have a higher probability to be HIV positive (as compared to e.g. a household for which $b_{1i} = 0$), the probability of these members perception of risk tend to be more on the left end of the scale of HIV risk perception (as compared to e.g. a household for which $b_{2i} = 0$), where that scale is defined as 0="unknown risk",1="no risk", 2="low risk", and 3="high risk".

As the random effect b_{2i} was correlated significantly with b_{1i} and shared its effect with b_{3i}, we decided to keep all other variance related parameters in the model, despite the fact that they turned out to be not significantly different from 0 (at level 0.05).

6. Discussion

This article presents two types of bivariate models for categorical response variables, being HIV risk perception Y_1 having one category "unknown risk" and three other ordinal categories ("no risk", "low risk","high risk"), and HIV infection status Y_2 (positive or negative). As interest goes to modelling the association between Y_1 and Y_2 as a function of covariates, while taking into account the hierarchical data structure, joint models with random household effects were considered. One model (CR model) starts with a continuation ratio-logits model for Y_1, being a natural choice in this particular setting, and combines it with the binomial distribution of Y_2 using global cross-product ratios as marginal association measures (following Dale's approach [9], and further extended with (partly) correlated and/or shared random effects for each of the three model components. A second model (CR-PO model) follows the same construct but starts with a model for Y_1 that combines continuation-ratio logits with cumulative logits (proportional odds). In the particular setting of HIV risk perception and infection status, both joint models only differ in two parameters, one parameter related to the risk perception distribution and one association parameter. In case Y_1 would have five or more parameters (e.g. including a "moderate risk" category), both type of joint models would differ in four or more parameters.

By specifying a model for the cross product ratio we can investigate different notions of dependence between the outcomes, e.g., whether outcomes are positively dependent, or if this dependence increases with respect to certain covariates. Models based on Figure 1 and the formulas for the odds ratios OR_k^{CR} and OR_k^{PO} $(k = 0, 1, 2)$ are of interest given the logits as defined for Y_1. In terms of our application, OR_0^{CR} or OR_0^{PO} quantifies the factor of change in the odds of not knowing his/her risk (as compared to knowing one's risk), while OR_1^{CR} (or OR_1^{PO}) quantifies the factor of change in the odds of perceiving no risk (as compared to perceiving low or high risk), given the changes in the relevant covariates. According to Table 4, the odds of HIV positive when one does not know his/her risk is estimated as 1.97 for the age group (15-24) and hence the probability of not knowing ones risk is about twice as large as of knowing (leading to probabilities about 2/3 and 1/3 respectively) when HIV positive. This corroborates with the findings of [29] which states that in Mozambique personal risk of HIV infection is greatly underestimated by individuals' risk perception, despite the high HIV prevalence [3]. This odds ratio decreases by about 50% (95% C.I.:0.30 – 0.81) when switching from the age group (15-24) to the age group (25-34), leading to an odds of about 1 or consequently more or less equal probability of knowing or not knowing individual's risk.

This modelling approach can be applied to similar settings as in [22] or slightly modified to deal with related but somewhat different settings as in quantity-frequency surveys [see 25]. Quantity-frequency response scales typically have two dimensions being the frequency (never, monthly or less, 2-4 times per month,...) and the quantity (1-2 drinks, 3-4 drinks,...). The second dimension only applies conditional on the first one being different from "never". This can be accommodated in our model by excluding the association parameters corresponding $k = 0$, being OR_0^{CR} and OR_0^{PO}. Our model then boils down to the one described in [25] but would allow to fit the data as one joint model with inclusion of common effects across the model components as well as different multivariate random effects structures.

The proposed full likelihood approach with different multivariate random effects structures and facing computational limits might become a real challenge to larger tables. Such an exercise however is considered worthwhile, as it forces one to think about all model components in detail. A GEE approach is a natural alternative approach [6, 14, 21]. However, implementing this approach and fitting such a model to our data with different types of working correlations (again facing us with many different options) would be far from straightforward. It is a very interesting topic of research but considered beyond the scope of this contribution.

For the application to the INSIDA data, different models with varying sets of covariates for the three components of the model and different covariance structures for the household random effects were fit, and the final model was selected using AIC and BIC, being the CR-PO model with partly correlated and partly shared random effects. Not unexpected, the estimates for the (conditional on household) parameters for the joint distribution of (Y_1, Y_2) of the (overall) best CR-PO model and the (second) best CR model were quite close. Computational difficulties were encountered for some of the models, indicating that information in the data for some of the parameters is quite scarce. The proposed models may be useful in various fields of application such as social and biomedical sciences, epidemiology and public health.

The results from the final CR-PO model as fitted to the INSIDA data indicate that only age has a significant effect on the association between HIV risk perception and infection status. The variable risky sexual behaviour, of special interest as it has been mentioned before in literature in relation to the HIV perception & status association, was not significant for the association model. The random household effect was mainly relevant for the HIV status. The negative association between perception of risk and HIV infection status within a household implies that assessing household members health status through his/her perception of risk is a very biased procedure, especially for younger individuals. Those that do not correctly perceive their HIV risk may unknowingly transmit the disease, thus increasing the HIV incidence and prevalence rates. We believe that a universal or mandatory HIV screening for special population groups should therefore be considered as a complementary tool in the fight against HIV/AIDS in the country. Mandatory testing has gained lot of political support in addressing the problem of HIV infection in certain regions of Gulf countries (Saudi Arabia, UAE), China, India, Ethiopia, Cambodia, Senegal and others [7].

One limitation in the INSIDA application is related to the considerable amount of missing data for both, the response variables and the covariates. Further investigation is required in order to assess the impact that missing mechanisms other than missing

completely at random may have on the applicability of our models in general and the results and conclusions about the application to the INSIDA data in particular.

Acknowledgement

The authors would like to acknowledge the sponsors of this study: the Flemish Interuniversity Council (VLIR-UOS) and Eduardo Mondlane University (UEM) through the DESAFIO Program. The authors would also like to acknowledge the support given by the Demographic and Health Surveys (DHS) Program from USAID for providing the data. The authors gratefully acknowledge the contribution of the reviewers for their valuable comments that have led to an improved version of the manuscript.

Disclosure statement

The authors declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The authors disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: The present study received financial support from Universidade Eduardo Mondlane (UEM)/Vlaamse Interuniversitaire Raad (VLIR-UOS) DESAFIO Program.

References

[1] P. Akwara, N. Madise, and A. Hinde, *Perception of risk of HIV/AIDS and sexual behaviour in Kenya*, Journal of Biosocial Science. 35 (2003), pp. 385–411.

[2] C.V. Ananth and D.G. Kleinbaum, *Regression models for ordinal responses: A review of methods and applications*, International Journal of Epidemiology. 26 (1997), pp. 1323–1333.

[3] C.M. Audet, J. Burlison, T. D. Moon, M. Sidat, A.E. Vergara, and S. H. Vermund, *Sociocultural and epidemiological aspects of HIV/AIDS in Mozambique*, BMC International Health and Human Rights. 10 (2010).

[4] F. Bartolucci, R. Colombi, and A. Forcina, *An extended class of marginal link functions for modelling contingency tables by equality and inequality constrains*, Statistica Sinica. 17 (2007), pp. 691 – 711.

[5] S.E. Boslaugh,*Health belief model*. Salem Press Encyclopedia. (2014), pp. 533–546.

[6] P.J. Catalano, *Bivariate modelling of clustered continuous and ordered categorical outcomes*, Statistics in Medicine. 16 (1997), pp. 883–900.

[7] V.K. Chattu, *Global Health & HIV/AIDS - A Critical Debate on Mandatory HIV Testing Policy*, Journal of Human Virology & Retrovirology. 1 (2014).

[8] R. Colombi and A. Forcina, *Marginal regression models for the analysis of positive association of ordinal response variables*, Biometrika, 88 (2001), pp. 1007–1019.

[9] J.R. Dale,*Global cross-ratio models for bivariate, discrete, ordered responses*, Biometrics. 42(1986), pp. 909–917.

[10] P. Diggle, P. Heagerty, K.-Y. Liang, and S.L. Zeger, *Analysis of Longitudinal Data*, 2nd edition, Oxford Science Publications, UK, 2002

[11] L.Eaton, A. J. Flisher, and L.E. Aaro, *Unsafe sexual behaviour in South African youth*, Social Science and Medicine. 56 (2003), pp. 149–165.

[12] S. Fieuws, G. Verbeke, and G. Molenberghs, *Random-effects models for multivariate repeated measures*, Statistical Methods in Medical Research. **1**6 (2007), pp. 387–397.

[13] J.D. Fishel, S.E.K. Bradley, P.W. Young, F. Mbofana, and C. Botao, *HIV among couples in Mozambique: HIV status, knowledge of status, and factors associated with HIV serodiscordance*, DHS Further Analysis Report, 2011.

[14] G.M. Fitzmaurice and N.M. Laird, *Regression models for a bivariate discrete and continuous outcome with clustering*, Journal of the American Statistical Association. 90 (1995), pp. 845–852.

[15] G. F. V. Glonek and P. McCullagh, *Multivariate logistic models*, Journal of the Royal Statistical Society, 57 (1995), pp. 533–546.

[16] H. Goldstein, J. Carpenter, M.G. Kenward, and K.A. Levin, *Multilevel models with multivariate mixed response types*, Statistical Modelling. 9 (2009), pp. 173–197.

[17] Instituto Nacional de Saúde (INS) and Instituto Nacional de Estatística (INE), *Inquérito Nacional de Prevalência, Riscos Comportamentais e Informação sobre o HIV e SIDA em Moçambique 2009*, Technical report, 2010.

[18] F. Kirakoya-Samadoulougou, S. Yaro, A. Deccache, P. Fao, M. -C. Defer, N. Meda, A. Robert, and N. Nagot, *Voluntary HIV testing and risky sexual behaviours among health care workers: a survey in rural and urban Burkina Faso*, BMC public health. 13 (2013).

[19] K.C. Koh and L.S. Yong, *HIV risk perception, sexual behavior, and HIV prevalence among men-who-have-sex-with-men at a community-based voluntary counseling and testing center in Kuala Lumpur, Malaysia*, Interdisciplinary Perspectives on Infectious Diseases. (2014), doi: 10.1155/2014/236240,

[20] A. Laenen, T. Vangeneugden, H. Geys, and G. Molenberghs, *Generalized reliability estimation using repeated measurements*, British Journal of Mathematical & Statistical Psychology. 59 (2006), pp. 113–131.

[21] K.Y. Liang, and S.L. Zeger, *Longitudinal Data Analysis Using Generalized Linear Models*, Biometrika. 73 (1986), pp. 13–22.

[22] E. Martins-Fonteyn, O. Loquiha, E. Wouters, I. Raimundo, N. Hens, M. Aerts,

and H. Meulemans, *HIV Susceptibility Among Migrant Miners in Chokwe: A Case Study*, International Journal of Health Services. preprint (2015). Available at sagepub.com/journalsPermissions.nav, doi: 10.1177/0020731415585988.

[23] P. McCullagh and J.A. Nelder,*Generalized Linear Models*, 2nd edition, Chapman and Hall, London, 1989.

[24] C. McCulloch,*Joint modelling of mixed outcome types using latent variables*, Statistical Methods in Medical Research. 17 (2008), pp. 53–73.

[25] G. Mcmillan and T. Hanson, *SAS Macro BDM for fitting the Dale regression model to bivariate ordinal response data*, Journal Of Statistical Software. 14 (2005), pp. 1–12.

[26] G. Molenberghs and E. Lesaffre, *Marginal modelling of multivariate categorical data*, Statistics in medicine, 18 (1999), pp. 2237–2255.

[27] G Molenberghs and G. Verbeke, *Models for Discrete Longitudinal Data*, Springer. & P.J., New York, 2005

[28] A. Nunn, N. Zaller, A. Cornwall, K.H.Mayer, E. Moore, S. Dickman, C. Beckwith, and H. Kwakwa, *Low perceived risk and high HIV prevalence among a predominantly African American population participating in Philadelphia's Rapid HIV testing program*, AIDS patient care and STDs. 25 (2011), pp. 229–235.

[29] N. Prata, L. Morris, E. Mazive, F. Vahidnia, and M. Stehr, *Relationship between HIV risk perception and condom use: Evidence from a population-based survey in Mozambique*, International family planning perspectives. 32 (2006), pp. 192–200.

[30] K. Pringle, R.C. Merchant, and M.A. Clark, *Is self-perceived HIV risk congruent with reported HIV risk among traditionally lower HIV risk and prevalence adult emergency department patients? Implications for HIV testing*, AIDS Patient Care STDS. 27 (2013), pp. 573–584.

[31] B.F. Qaqish and A. Ivanova, *Multivariate logistic models*, Biometrika. 93 (2006), pp. 1011–1017.

[32] S. Rabe-Hesketh and A. Skrondal, *Parameterization of multivariate random effects models for categorical data*, Biometrics. 57 (2001), pp. 1256–1264.

[33] J. -A. Røttingen, D. W. Cameron, and G.P. Garnett, *A systematic review of the epidemiologic interactions between classic sexually Transmitted Diseases: HIV How Much Really Is Known?*, Journal of Sexually transmitted diseases. 28 (2001), pp. 579–597.

[34] E.Y. Tenkorang, E. Maticka-Tyndale, and F. Rajulton, *A multi-level analysis of risk perception, poverty and sexual risk-taking among young people in Cape Town, South Africa*, Health and Place. 17 (2001), pp. 525–535.

Table 1. INSIDA survey. Variable definition and summary statistics. N is the number of households, n_i the number of eligible members of household i.

				Subset**	
Response variables (all at the individual level)		$\sum_{i=1}^{N} n_i = 10548$	%	$\sum_{i=1}^{N} n_i = 7944$	%
HIV infection status					
	HIV- (reference)	8396	79.6	6950	87.5
	HIV+	1302	12.3	994	12.5
	Missing	850	8.1	-	-
Perception of risk for HIV					
	Unknown risk	3005	28.5	2281	28.7
	No risk	3200	30.3	2335	29.4
	Low risk	2145	20.3	1777	22.4
	High risk	1942	8.4	1551	19.5
	HIV+	111	1.1	-	-
	Missing	145	1.4	-	-
Explanatory variables (partly at the individual and partly at the household level)					
Individual level		$\sum_{i=1}^{N} n_i = 10548$	%	$\sum_{i=1}^{N} n_i = 7944$	%
Condom use (last 3 sexual intercourses)					
	Consistent use	1012	9.6	872	11.0
	Inconsitent use	8103	76.8	7072	89.0
	Missing	1433	13.6	-	-
Number of sexual partners (last 12 months)					
	≤ 1	9394	89.1	6926	87.2
	≥ 2	1150	10.9	1018	12.8
	Missing	4	0.0	-	-
Sexual transmitted diseases (last 12 months)					
	No	9791	92.8	7636	96.1
	Yes	372	3.5	308	3.9
	Missing	385	3.7	-	-
Risky sexual behaviour (RSB)	Scale: Min=0 (no) and Max=3 (high)	1.05*	-	1.05*	-
Age groups (years)					
	15–24 (reference)	3139	29.8	2515	31.7
	25–34	3099	29.4	2465	31.0
	35–44	2042	19.3	1549	19.5
	$\geq$45	2268	21.5	1415	17.8
Sex					
	Male (reference)	4427	42.0	3590	45.2
	Female	6121	58.0	4354	54.8
Knowledge to reduce HIV risk infection					
	No (reference)	1063	10.1	739	9.3
	Yes	9332	88.5	7205	90.7
	Missing	153	1.4	-	-
Household level		N=5573	%	N=4912	%
Place of residence					
	Urban (reference)	2452	44.0	2155	43.9
	Rural	3121	56.0	2757	56.1
Household size	Scale: Min=1 and Max=23	4.5*	-	4.7*	-

*sample average

**subset of data used in this study

Table 2. Comparison of model fit. The column '-2ll' shows the values of -2×log-likelihood; the column '#Par' shows the number of parameters and the columns 'Rank' refers to the ranking of the models according to the AIC and BIC criterion.

Type		Model	-2ll	#Par	AIC	BIC	AIC rank	BIC rank
CR	1	Marginal	26887	34	26955	27192	8	8
	2	Shared RE	26654	41	26736	27000	6	6.5
	3	Independence RE	26565	41	26647	26910	5	5
	4	Correlated RE(1)	26553	42	26637	26907	2	3
	5*	Correlated RE(2)	–	42	–	–	-	-
	6*	Correlated RE(3)	–	42	–	–	-	-
CR-PO	1	Marginal	26888	34	26956	27193	9	9
	2	Shared RE	26655	41	26737	27000	7	6.5
	3	Independence RE	26558	41	26640	26904	3.5	2
	4	Correlated RE(1)	26556	42	26640	26909	3.5	4
	5*	Correlated RE(2)	–	42	–	–	-	-
	6	Correlated RE(3)	26540	42	26624	26893	1	1

* models with convergence problems.

Table 3. Odds ratio estimates ((95% confidence interval)) of CR model 4, the best CR model as listed in Table 2.

	HIV perception of risk			HIV infection status
Effects	Odds_0^{CR}	Odds_1^{CR}	Odds_2^{CR}	Odds HIV+
Intercept	0.28(0.22 – 0.36)	1.60 (1.19 – 2.14)	4.95 (3.48 – 7.05)	0.03 (0.02 – 0.04)
Place (Rural)	1.45 (1.30 – 161)	1.97 (1.70 – 2.29)		0.41 (0.33 – 0.50)
Sex (Female)	2.14 (1.92 – 2.39)	0.72 (0.64 – 0.81)		1.67 (1.40 – 1.98)
Knowledge (Yes)	0.50 (0.42 – 0.58)	0.58 (0.46 – 0.73)		1.34 (0.97 – 1.85)
Age (25-34)	1.08 (0.95 – 1.23)	0.90 (0.79 – 1.04)		2.56 (2.04 – 3.21)
Age (35-44)	1.13 (0.98 – 1.30)	0.91 (0.78 – 1.08)		2.14 (1.64 – 2.78)
Age (≥45)	1.17 (1.01 – 1.36)	1.45 (1.22 – 1.72)		1.57 (1.19 –8.57)
RSB	1.26 (1.11 – 1.42)	0.59 (0.52 – 0.67)		1.23 (1.01 – 1.50)
	OR_0^{CR}	OR_1^{CR}	OR_2^{CR}	
Intercept	1.88 (1.28 – 2.76)	1.03 (0.68 – 1.70)	0.84 (0.39 – 1.84)	
Age (25-34)	0.50 (0.30 – 0.83)	0.87 (0.54 – 1.39)		
Age (35-44)	0.84 (0.47 – 1.48)	0.70 (0.37 – 1.36)		
Age (≥45)	1.26 (0.43 – 1.48)	0.70 (0.35 – 1.38)		
	Variance components			
σ_1^2	0.059(0.065)			
σ_2^2	2.688(0.370)*			
σ_3^2	0.074(0.123)			
ρ	-0.226(0.069)*			
γ_{11}	5.061(2.757)			
γ_{12}	4.703(3.103)			
γ_{31}	0.687(1.554)			
γ_{32}	-1.070(2.063)			

Table 4. Odds ratio estimates (95% confidence interval) of CR-PO model 6, the best CR-PO and the best overall model as listed in Table 2.

	HIV perception of risk			HIV infection status
Effects	Odds_0^{PO}	Odds_1^{PO}	Odds_2^{PO}	Odds HIV+
Intercept	0.29 (0.23 –0.36)	1.58 (1.18 – 2.13)	4.90 (3.55 – 6.78)	0.03 (0.02 – 0.04)
Place (Rural)	1.45 (1.31 – 1.60)	1.97 (1.72 – 2.26)		0.39 (0.32 – 0.49)
Sex (Female)	2.12 (1.90 – 2.36)	0.71 (0.64 – 0.80)		1.68 (1.42 – 2.00)
Knowledge (Yes)	0.50 (0.43 – 0.59)	0.58 (0.45 – 0.73)		1.35 (0.97 – 1.88)
Age (25-34)	1.08 (0.95 – 1.23)	0.89 (0.77 – 1.02)		2.66 (2.11 – 3.36)
Age (35-44)	1.13 (0.98 – 1.30)	0.91 (0.78 – 1.08)		2.16 (1.65 – 2.82)
Age (≥45)	1.17 (1.01 – 1.36)	1.43 (1.21 – 1.70)		1.60 (1.21 – 2.12)
RSB	1.25 (1.10 – 1.41)	0.59 (0.53 – 0.67)		1.26 (1.03 – 1.53)
	OR_0^{PO}	OR_1^{PO}	OR_2^{PO}	
Intercept	1.97 (1.16 – 3.35)	0.57 (0.25 – 1.29)	0.89 (0.41 – 1.90)	
Age (25-34)	0.50 (0.30 – 0.81)	0.89 (0.45 – 1.74)		
Age (35-44)	0.83 (0.47 – 1.45)	0.66 (0.30 – 1.45)		
Age (≥45)	0.76 (0.42 – 1.39)	0.75 (0.34 – 1.66)		
	Variance components			
σ_1^2	0.022(0.036)			
σ_2^2	2.421(0.288)*			
σ_3^2	0.022(0.065)			
ρ	-0.255(0.0649)*			
γ_{11}	8.877(7.538)			
γ_{12}	8.073(7.191)			
γ_{31}	-4.18(6.542)			
γ_{32}	0.189(2.174)			