Statistics and Its Interface Volume 9 (2016) 57–67

A copula approach to estimate reliability: an application to self-reported sexual behaviors among HIV serodiscordant couples

Scarlett L. Bellamy*, Seunghee Baek, Andrea B. Troxel, Thomas R. Ten Have†, and John B. Jemmott III

Copula-based approaches can be useful in multivariate modeling settings where multivariate dependency is of primary interest, such as estimating the reliability of self-reported sexual behavior assessed independently for male and female partners (dyad) in a couple-based HIV risk reduction study. Specifically, we investigate the reliability of couple reports using copulas, adjusting for key individual baseline covariates. We propose applying a copula modeling approach to measure the reliability of self-reported, shared sexual behaviors from couples where measures are assessed independently from male and female partners. In particular, we estimate measures of dependence, such as the odds ratios and binary correlations, using mixtures of max-infinitely divisible copulas with a bivariate logit model. This approach is flexible in measuring the effect of covariates on dependence parameters and for estimating marginal probabilities for multiple outcomes simultaneously. In this paper, we focus on estimating these two dependencies and explore the influences of additional covariate information on the copula parameter. We provide simulation results comparing copula-based estimates to moment estimates of the generalized estimating equation (GEE) for the correlation coefficients with respect to bias and 95% coverage probability. We illustrate that copulas have better performance in terms of bias, while their performance is similar with respect to efficiency. The estimator of the marginal probability using copula methods is robust to the choice of copula family. The choice of copula may affect the estimator of dependency when the dependency of the outcomes is very low. We apply these methods to data from the Multisite HIV/STD Prevention Trial for African American Couples (AAC) Study.

Keywords and phrases: Copula, Reliability, Bivariate.

1. INTRODUCTION

Understanding relationships among multivariate outcomes is a fundamental problem in statistical science. In longitudinal and/or cluster-randomized trials, the dependency among outcomes may not be of primary interest, but it must be accounted for in order to make valid inference. In other settings when outcome dependency is of primary interest, copulas have become an increasingly popular analysis tool particularly in biomedical research applications.

The motivation for our work comes from a clinical trial for HIV serodiscordant couples. Specifically, the Multisite HIV/STD Prevention Trial for African American Couples (AAC) was a successful behavioral modification trial for African American, heterosexual, HIV discordant couples, whose goal was to decrease risky sexual behaviors and increase health promoting behaviors among enrolled couples. The main purpose of the study was to examine the efficacy of a couple-focused HIV/STD risk reduction intervention versus an individual-focused health promotion intervention in reducing sexual risk behaviors and STD incidence. A comprehensive description of the study design and randomization process can be found in [1] and the primary intervention efficacy findings are presented for (1) sexual behavioral outcomes and (2) health promoting outcomes in [7] and [6], respectively.

Heterosexual transmission of HIV is the dominant route of infection worldwide, indicating a critical need for heterosexual couple-focused interventions [27] especially for high-risk serodiscordant couples (one partner is HIV positive and one partner is HIV negative). Reliable assessments of HIV sexual risk behaviors are critical in informing the efficacy of behavioral modification interventions. A summary of prior couple-focused interventions are presented in [2] and can provide a unique opportunity to measure the reliability of self-reported shared sexual behaviors, as each partner is measured independently and one can measure the degree to which couple responses are consistent. In AAC, both partners are assessed independently for a number of shared sexual behaviors with their study partners. Thus, estimating the magnitude of the concordancy and discordancy among couple outcomes will allow evaluation of the reliability of each individual partner's self-reported behavior. Reliable measures of self-reported sexual behaviors in high-risk populations have direct and obvious implications for estimating intervention effects. In this trial, couples assessed their condom use and other shared sexual behaviors retrospectively.

*Corresponding author.
†Deceased.

Potential issues with many self-report measures are their susceptibility to response bias, that is, a tendency for subjects to over- or under-report outcomes for a number of reasons [4]. Additionally, there is no "gold standard" for quantifying the validity of sexual behaviors since these behaviors are largely unobtainable by more objective methods. Nevertheless, in couple-based studies, examining the concordance of partner responses is a reasonable assessment of reliability for shared behaviors. A strength of the study is that each shared sexual behavior of interest is reported independently by each study partner (males and females, separately); therefore concordance of responses for these shared couple behaviors can be readily evaluated and used as a measure of reliability. Additionally, exploring the influence of individual factors on estimated concordance may also help explain sources of response biases, if they exist. The effects of demographics and the couples' relationship context on concordance of reported sexual behaviors were examined using a measure of agreement such as Kappa statistics, conditional probability and McNemar's Statistics in [8].

A few studies have quantitatively explored individual characteristics associated with concordance of partner reporting of sexual behaviors [8, 21, 23]. For the few studies that do try to estimate the association of individual factors on discordant couple responses, often a single outcome is constructed for each measure of interest that is a simple indicator of whether or not both partners had identical responses, and a gender-stratified model is used to predict the constructed couple indicator of discordance as a function of gender-specific characteristics [8]. Since factors that are related to their responses should explain both concordance and discordance and not just one, ideally we would like to employ a statistical tool to quantify dependence while simultaneously adjusting for individual factors that may be associated with this dependence.

Copulas allow us to model the dependence structure of outcomes (unadjusted and adjusted for other covariates) separately from the marginal probability in addition to constructing a joint multivariate distribution of all outcomes. A nice feature of the copula approach in settings like AAC, where each member of the couple dyad is assessed independently regarding shared behaviors (e.g., male participants provide data on condom use with their female study partner in the past 90 days and female participants provide data on condom use with their male study partner in the past 90 days) is that each pair of responses can be used to measure the reliability of these same self-reported, shared behaviors. Additionally, while we construct the dependence structure, we are able to estimate dependency, a measure of reliability in this context, adjusting for individual characteristics of interest. In AAC, we are interested in (1) measuring dependency of partner outcomes to provide a measure of reliability for a number of self-reported sexual behavioral measures and (2) adjusting these estimates of dependency for a number of potentially important couple-level characteristics. Thus, employing copulas is a feasible method for measuring reliability of self-reported data in our motivating example. Our primary outcome of interest is a correlated binary outcome, which is 'consistent condom use' at every sexual episode with study partner, for both male and female partners. Thus, we apply a multivariate logit model introduced in [19], which uses a mixture of max-infinitely divisible (max-id) bivariate copulas proposed in [14]. The application of copulas is usually limited for modeling multivariate binary outcomes primarily because of theoretical and computational limitations since the probability mass function should be obtained using finite differences for discrete data [20]. Therefore, in order to implement copula models for multivariate discrete data, we need to specify copulas with rather simple forms [20]. There are some desired properties for a parametric family of multivariate copulas applicable to discrete data described in [14, 20]. The max-id copula approach is attractive since it allows flexible positive dependence structures and has a closed form cumulative distribution function (cdf); no other copula family has both these properties [14]. However, it allows only positive dependence between random variables. Unlike other copula families where dependence parameters have joint constraints among them [14, 13], the max-id copula achieves dependency sufficiency such that we can model the dependence parameter using the covariate information. Therefore, flexibility in modeling dependency while adjusting for covariates will allow us to examine how a number of factors (e.g., sociodemographic or relationship characteristics) may be associated with the reliability of self-reported shared behaviors among couples.

In sum, this paper demonstrates how to model and estimate dependency (i.e., reliability) parameters from multivariate binary data using copulas, evaluates the performance of this approach through a number of simulation studies and applies the proposed method to our motivating example. Note that GEE [15] is another commonly used method that provides estimates of correlation as well as covariates. Therefore, we conduct a simulation study to compare copula-based estimates to moment estimates of GEE and to the second order GEE2 [16] for the correlation coefficients, a common measure of dependency, and to examine how the copula approach can estimate the covariate effect on the copula parameter. Alternating logistic regression (ALR) proposed in [3] could be used to adjust for different levels of clustering in the pairwise odds ratio [3, 17]. However, one limitation of this approach is that it applies only when $n_i = n$ for all clusters, but the copula-based method proposed has no such restriction. Also, estimates of the covariate effect in the pairwise odds ratio are not directly comparable since the copula method adjusts for covariates in the copula parameter, not in the pairwise odds ratio. Based on the results from the simulation study, we apply what we believe is the most appropriate copula to AAC data.

Accordingly, the purpose of this paper is to present a novel application of copulas to quantify dependence by estimating association as a proxy for reliability in the context of a couple-based behavioral modification trial. Specifically,

Table 1. Max-id bivariate copulas and Laplace transform (LT)

Family	$C'(u_j, u_k; \theta)$	LTs: $\phi(t;\theta)$	$\theta \in$	Log transformation
Gumbel	$e^{-(\tilde{u_j}^{\theta}+\tilde{u_k}^{\theta})^{1/\theta}}$	A: $e^{-t^{1/\theta}}$	$[1,\infty)$	$\log(\theta-1)$
Kimeldorf	$(u_j^{-\theta}+u_k^{-\theta}-1)^{-1/\theta}$	B: $(1+t)^{-1/\theta}$	$(0,\infty)$	$\log\theta$
Joe	$1-(\bar{u_j}^{\theta}+\bar{u_k}^{\theta}-\bar{u_j}^{\theta}\bar{u_k}^{\theta})^{1/\theta}$	C: $1-(1-e^{-t})^{1/\theta}$	$[1,\infty)$	$\log(\theta-1)$
Frank	$-\frac{1}{\theta}\log\{1+\frac{(e^{-\theta u_j}-1)(e^{-\theta u_k}-1)}{e^{-\theta}-1}\}$	D: $-\frac{\log(1-(1-e^{-\theta})e^{-t})}{\theta}$	$(0,\infty)$	$\log\theta$
Galambos	$u_j u_k e^{(\tilde{u_j}^{-\theta}+\tilde{u_k}^{-\theta})^{-1/\theta}}$		$[0,\infty)$	$\log\theta$

Note: $\bar{u_j} = 1-u_j$ and $\tilde{u_j} = -\log u_j$ where $u_j = F_j(y_j)$; From [19]

we apply max-id copulas to the AAC project in order to estimate the dependency of couple responses and gain insight into the reliability of self-reported sexual risk behaviors among a sample of African American, serodiscordant heterosexual couples.

The following sections give more details on the copula-based approaches for bivariate binary data, parameter estimation based on the likelihood function, and estimation of the odds ratios and binary correlations.

2. MATERIALS AND METHODS

2.1 A copula approach for binary data

A copula-based model involves the generation of a multivariate joint distribution for outcomes of interest given the marginal distributions of the correlated responses. The definition of a copula $C(u_1,\ldots,u_m)$ is a multivariate distribution function defined over the unit cube linking uniformly distributed marginals $(u_1,\ldots,u_m)$ [24, 18]. Let $F_j(Y_j)$ be the cumulative distribution function (cdf) of a univariate random variable Y_j $(j = 1,\ldots,m)$. Then, $C(F_1(y_1),\ldots,F_m(y_m))$ is an m-variate distribution for $y = (y_1,\ldots,y_m)^T$ with marginal distributions F_j $(j = 1,\ldots,m)$. Sklar first showed that there exists an m-dimensional copula C such that for all y in the domain of H [24],

$$H(y_1,\ldots,y_m) = C(F_1(y_1),\ldots,F_m(y_m)). \tag{1}$$

If $F_1,\ldots,F_m$ are continuous, then the function C is unique; otherwise, there are many possible copulas as emphasized in [11]. However, all of these coincide on the closure of $Ran(H_1)\times\ldots\times Ran(H_m)$, where $Ran(H)$ denotes the range of H. While it is relatively easy to derive a joint distribution in the continuous case, it is not so simple in the case of discrete data. The latter involves 2^m finite differences of $H(y)$, thus, to compute the joint probability mass function, one needs to evaluate the copula repeatedly. Therefore, in order to be able to use copula models for multivariate discrete data, one needs to specify copulas with rather simple forms.

Joe and Hu [14] proposed multivariate parametric families of copulas that are mixtures of max-infinitely divisible (max-id) bivariate copulas, allowing flexible dependence structures, having closed form cdfs, and satisfying the closure property under marginalization. This meets three desired properties for a parametric family of multivariate copulas applicable to discrete data [20]. One property this does not satisfy is allowing negative dependence.

Given that our primary outcomes of interest are binary responses collected from male and female partners within each couple, we will use mixtures of max-id copulas. The mixture of m-variate max-id copulas cdfs has the following form

$$C(\mathbf{u};\Theta) = \phi\Bigg(\sum_{j<k}\log\, C'_{jk}(e^{-p_j\phi^{-1}(u_j;\theta)}, e^{-p_k\phi^{-1}(u_k;\theta)};\theta_{jk}) + \sum_{j=1}^{m} v_j p_j \phi^{-1}(u_j;\theta);\theta\Bigg) \tag{2}$$

where $C'_{jk}(\cdot;\theta,\theta_{jk})$ is a bivariate max-id copula, $\phi(\cdot;\theta)$ is a Laplace transform (LT), $\Theta = \{\theta, \theta_{jk} : j,k = 1,...,m, j < k\}$ denotes the vector of all dependence parameters of the copula, u_j is cdf of a univariate random variable and $p_j = (v_j + m - 1)^{-1}$ where v_j is arbitrary. Specifically, the (j,k) bivariate marginal copula is

$$\begin{aligned} C_{jk}(u_j,u_k;\theta,\theta_{jk}) &= \phi\big(-\log C'_{jk}(e^{-p_j\phi^{-1}(u_j;\theta)}, e^{-p_k\phi^{-1}(u_k;\theta)};\theta_{jk}) \\ &\quad + (v_j+m-2)p_j\phi^{-1}(u_j;\theta) \\ &\quad + (v_k+m-2)p_k\phi^{-1}(u_k;\theta);\theta\big). \end{aligned} \tag{3}$$

We can simplify Equation (3) by assuming $v_j+m-2 = 0$, then Equation (2) would become max-id copula with $m\ (m-1)/2+1$ dependence parameters. In our bivariate model, we need only one copula parameter and therefore force θ_{jk} equal to θ for every pair. Some members of max-id bivariate copulas and LTs are presented in Table 1. Thus, combining the five families with the corresponding four LTs in Table 1 results in 20 parametric copula families each with flexible dependence structures.

In particular, this approach allows us to estimate the measure of association between two binary outcomes through the copula dependence parameter, θ, which represents the degree of association. Moreover, we can incorporate covariate information in estimating the copula parameter by using a log transformation; this will be explained in detail in subsection 2.2. Thus, we can obtain an estimate of correlation, adjusting for covariates of interest. Note that the GEE method could be used to model bivariate binary data by regarding the correlation between two outcomes as a nuisance parameter [15]. Since GEE is a widely used method, we will proceed to compare the estimated dependence from the copula approach with the moment estimates of correlation coefficient from GEE and GEE2. Models were fit using the R package *'copula'* ([12], [28]).

2.2 Copula-based bivariate logit model

In this section, we will discuss how the copula-based method can be integrated into a logit model and how to introduce covariate information in the copula parameter, θ. For simplicity and to relate the notation to our couples data, we will describe the bivariate logit model where $j = 1, 2$ denotes female and male responses respectively. Note that these models can be easily extended to a multivariate logit model where $j > 2$. Consider Equation (2) where $\mathbf{y} = (y_1, y_2)$ denotes the bivariate binary response for a couple and F_j the cdf of the univariate Bernoulli distribution function with probability of success π_j,

$$F_j(y_j; \pi_j) = \begin{cases} 1 - \pi_j & \text{if } y_j = 0 \\ 1 & \text{if } y_j = 1 \end{cases} \qquad j = 1, 2. \tag{4}$$

The standard logistic regression model for the probability of success π_{ij} corresponding to the copula in Equation (2) is

$$logit(\pi_{ij}) = \beta_j^T x_{ij}, \quad j = 1, 2 \tag{5}$$

where β_j is the vector of marginal regression parameters and x_{ij} is a vector of covariates for the ith couple with jth partner (female or male). As mentioned previously, we can also model the dependence structure and introduce a regression coefficient in the copula parameter θ_i by choosing the appropriate log transformation for a given family from Table 1. For example, if we specify a Frank copula with the Dth LT from Table 1 (Frank D), then we would use the following model

$$\log(\theta_i) = b_i^T x_i, \quad i = 1, ..., n, \tag{6}$$

where b_i^T is a vector of regression coefficients in the dependence measure and x_i is a vector of covariate for the ith couple. θ_i in the dependence model above will be incorporated in Equation (2) and used for joint distribution modeling of bivariate outcomes.

2.3 Parameter estimation

When marginal models are discrete, a multivariate probability function is obtained by taking the Radon-Nikodym derivative for $H(y)$ in Equation (2). Thus, for the binary case, the bivariate probability function is given by

$$\begin{aligned} P(Y_1 = y_1, Y_2 = y_2) \\ = C(u_1, u_2) - C(u_1, v_2) - C(v_1, u_2) + C(v_1, v_2) \end{aligned} \tag{7}$$

where $u_j = F_j(y_j)$ and $v_j = F_j(y_j - 1)$ [25]. It follows that the joint log-likelihood of the bivariate logit copula model with various choices of copula family and LT can be written as

$$\begin{aligned} L(\beta, b) = \sum_{i=1}^{n} \log[&C(F_1(y_{i1}), F_2(y_{i2})) \\ &- C(F_1(y_{i1}), F_2(y_{i2} - 1)) \\ &- C(F_1(y_{i1} - 1), F_2(y_{i2})) \\ &+ C(F_1(y_{i1} - 1), F_2(y_{i2} - 1)); X_{ij}, \beta_{\mathbf{j}}, b_i] \end{aligned} \tag{8}$$

where C is max-id copula, F_1, F_2 are univariate marginal cdfs, $\beta = (\beta_1, \beta_2)$ is a vector of regression coefficients in the marginal model and b_i is a vector of regression coefficients in copula parameter. In this study, we focus on the standard maximum likelihood (ML) method that maximizes the joint log-likelihood. By using the ML method, we will simultaneously obtain the estimates of both copula and marginal parameters.

2.4 Estimation of odds ratio and binary correlation

In this study, we present odds ratios and binary correlations as a measure of dependency. The copula parameter may be presented as a measure of dependency in the copula-based method. However, it is not directly comparable to other dependence measures even among the copula-based approaches since it differs according to which copula families are used. Due to this limitation, many applications involving copula methods use Kendall's τ as a measure of association. Kendall's τ is appropriate for measuring the strength of dependence between continuous outcomes, but it is less appropriate as a measure of association when applied to discrete variables. In particular, it is no longer distribution-free and has a range narrower than $[-1, 1]$, and this has to be taken into account when assessing the strength of the dependence [5]. The bounds on Kendall's τ are plotted for Bernoulli margins with success probabilities $p_1 = p_2 = p \in [0, 1]$ in [20]. Given the marginal probabilities of p_1 and p_2, we can rewrite Kendall's τ using the copula-based joint distribution of success, $C(1, 1)$,

$$\tau(Y_1, Y_2) = 2[C(1, 1) - p_1 p_2]. \tag{9}$$

Thus, even in the most favorable cases, Kendall's τ does not reach 1 or -1 and cannot be comparable to usual measures of dependence when outcomes are binary. On the other hand, the odds ratio, which is one of the common measures of the association between pairs of responses, is not

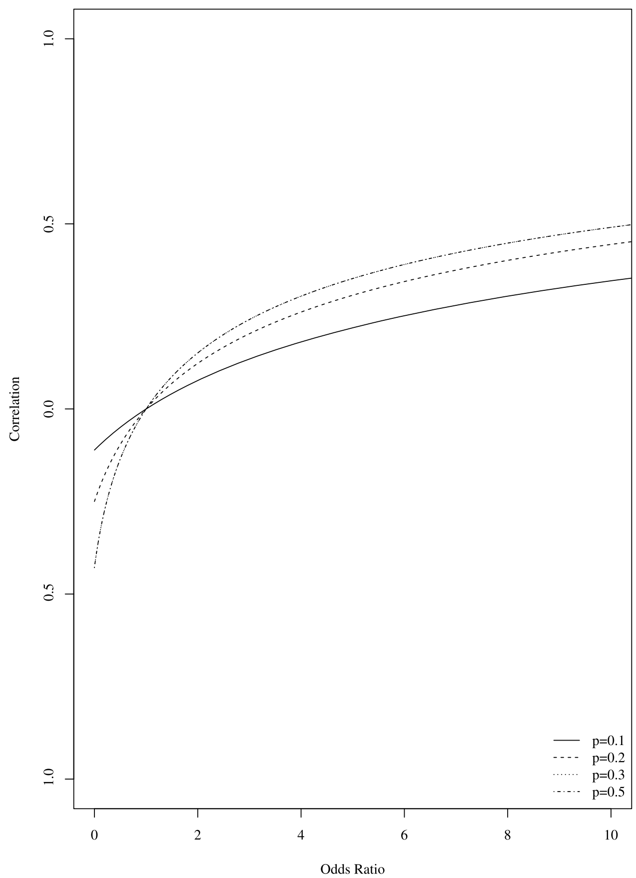

Figure 1. Relationship between the odds ratio and correlation coefficient.

constrained by marginal probabilities in the same way that Kendall's τ is constrained. The odds ratio, φ, can take any value in $(-\infty, \infty)$ with $\varphi = 1$ corresponding to no association. Figure 1 shows the relationship between the odds ratio and correlation coefficient, for examples in which the marginal probability of response for both males and females is 0.1, 0.2, 0.3, or 0.5.

We can write the odds ratio as a function of the joint probability of failure for both outcomes and their marginal probabilities, p_1 and p_2. Denote the joint probability of failure, p_{00}, for both female and male partners as

$$p_{00} = Pr\,(Y_1 = 0, Y_2 = 0) \tag{10}$$

where Y_j (Y_1 = female response, Y_2 = male response) denotes binary bivariate outcomes from female and male partners considering our motivating example. By using a copula approach, we have the following form of the joint probability of failure derived from Equation (2):

$$\begin{aligned} p_{00} &= Pr\,(Y_1 = 0, Y_2 = 0) \\ &= C(u_1, u_2 : \Theta) \\ &= \phi(-\log C'(e^{-p_1\phi^{-1}(u_1;\theta)}, e^{-p_2\phi^{-1}(u_2;\theta)}; \theta) \\ &\quad + (v_1 + m - 2)w_1\phi^{-1}(u_1;\theta) \\ &\quad + (v_2 + m - 2)w_2\phi^{-1}(u_2;\theta)) \end{aligned} \tag{11}$$

where $u_1 = F(Y_1 = 0) = 1 - \pi_1$, $u_2 = F(Y_2 = 0) = 1 - \pi_2$, $w_j = (v_j + m - 1)^{-1}$, $m = 2$, v_j is arbitrary, and $j = 1, 2$. Let φ be the odds ratio between responses Y_1 and Y_2. The odds ratio for binary responses is defined as

$$\begin{aligned} \varphi &= \frac{Pr(Y_1 = 1, Y_2 = 1)Pr(Y_1 = 0, Y_2 = 0)}{Pr(Y_1 = 1, Y_2 = 0)Pr(Y_1 = 0, Y_2 = 1)} \\ &= \frac{(p_{00} + p_1 + p_2 - 1)p_{00}}{(1 - p_1 - p_{00})(1 - p_2 - p_{00})} \end{aligned} \tag{12}$$

where $p_1 = Pr(Y_1 = 1)$ and $p_2 = Pr(Y_2 = 1)$; this is equivalent to the ratio of the odds of concordant to discordant responses. p_{11} can be simply derived by using the equation $p_{11} = p_{00} + p_1 + p_2 - 1$, where p_{00} is derived from the copula approach when estimating the joint probability.

For the purpose of comparing a measure of dependence from the copula method to a moment estimate for correlation coefficient from GEE, we also estimate binary correlation using the formula of phi correlation in [26]. The binary correlation coefficient using the above probabilities that will always lie in $[0, 1]$ is

$$Corr(\rho) = \frac{p_{11} - p_1 p_2}{\sqrt{p_1(1 - p_1)p_2(1 - p_2)}}. \tag{13}$$

3. SIMULATION STUDIES

We conducted a simulation study to explore the performance of the copula approaches in estimating correlations and covariate effects on the copula parameter, and compared copula-based correlation estimates to moment estimates for correlation coefficients obtained from the GEE methods (first- and second-order, where applicable). We use the results from the simulation study to determine which copula family might be the most appropriate for the AAC data.

3.1 Data simulation method

We created correlated bivariate binary random variables by thresholding a normal distribution using the package 'bindata' in R, which applies the algorithm presented in [22]. We can set n samples from a multivariate normal distribution with mean and variance chosen in order to get the desired marginal and common probabilities. We generated 500 simulation repetitions correlation coefficients equal to 0.05, 0.1, 0.25, 0.5 and 0.75 as well as separately with 3 sets of covariate coefficients (0.01, 0.30 and 1.00) in the dependence model (on copula parameter). The sample size was equal to 1,000 (500 pairs of correlated outcomes) in both settings. Parameter estimates and corresponding standard errors for the odds ratio and correlation coefficient from the copula method were estimated based on bootstrapped resampling (500 repetitions) within each simulated dataset.

Table 2. The average estimates and standard errors of odds ratios and correlation coefficient for simulated data using copula approach and GEE

True			Odds Ratio		Correlation				
Correlation	Method	Group	Est.	S.E.	Est.	S.E.	MSE	Bias	$CP_{95\%}$
$\rho = 0.05$	Gumbel D	Trt.	1.664	0.144	0.100	0.016	0.0027	0.0498	18.8
		Ctrl.	1.664	0.144	0.099	0.016	0.0026	0.0495	18.2
	Frank A	Trt.	1.325	0.323	0.050	0.046	0.0023	-0.0005	95.0
		Ctrl.	1.323	0.324	0.049	0.046	0.0023	-0.0009	94.4
	GEE		-	-	0.047	0.046	0.0023	-0.0026	94.4
$\rho = 0.1$	Gumbel D	Trt.	1.803	0.289	0.115	0.028	0.0010	0.0147	98.0
		Ctrl.	1.803	0.289	0.115	0.028	0.0010	0.0150	98.0
	Frank A	Trt.	1.684	0.406	0.098	0.047	0.0020	-0.0017	95.0
		Ctrl.	1.686	0.405	0.099	0.047	0.0020	-0.0014	95.6
	GEE		-	-	0.096	0.047	0.0020	-0.0037	95.8
$\rho = 0.25$	Gumbel D	Trt.	3.440	0.833	0.248	0.048	0.0024	-0.0024	93.4
		Ctrl.	3.445	0.836	0.247	0.048	0.0024	-0.0027	93.0
	Frank A	Trt.	3.440	0.825	0.248	0.048	0.0024	-0.0023	93.6
		Ctrl.	3.445	0.827	0.247	0.048	0.0024	-0.0027	94.4
	GEE		-	-	0.245	0.048	0.0024	-0.0048	94.2
$\rho = 0.5$	Gumbel D	Trt.	12.165	3.368	0.498	0.045	0.0019	-0.0023	95.4
		Ctrl.	12.114	3.344	0.498	0.045	0.0019	-0.0022	95.4
	Frank A	Trt.	12.157	3.383	0.498	0.045	0.0019	-0.0024	95.0
		Ctrl.	12.115	3.361	0.498	0.045	0.0019	-0.0022	95.0
	GEE		-	-	0.496	0.045	0.0020	-0.0043	95.6
$\rho = 0.75$	Gumbel D	Trt.	71.477	32.769	0.747	0.035	0.0000	-0.0033	94.4
		Ctrl.	71.057	39.021	0.747	0.035	0.0000	-0.0029	94.4
	Frank A	Trt.	71.666	32.463	0.748	0.035	0.0000	-0.0014	95.2
		Ctrl.	71.539	32.210	0.749	0.035	0.0000	-0.0015	95.4
	GEE		-	-	0.746	0.035	0.0045	0.0000	93.6

Note: '$CP_{95\%}$' = Coverage probability of 95% confidence interval

The simulated marginal probability was set to 0.25 for both treatment groups and couple measures. The marginal probability of 0.25 is a crude estimate of the marginal probability for the primary outcome of interest for both treatment and control groups at baseline in the AAC sample.

3.2 Bias and efficiency in estimating correlations

For each of the four simulation scenarios, we estimate odds ratios and correlation coefficients using the copula approach with a combination of 5 choices of max-id copula families and 4 choices of corresponding Laplace transformations (LTs A-D) (Table 2). Here we present detailed results only for Gumbel copula with Laplace transformation D (Gumbel D) and Frank copula with Laplace tranformation A (Frank A) as the other families provided similar results. To examine the performance of each estimator, we present bias, 95% coverage probability and mean squared error (MSE).

For all underlying true correlations, Frank A performs as well as or better than both Gumbel D and the GEE methods, providing a good 95% coverage probability (93.6% − 95.6%), small (absolute) bias (0.0005 − 0.0048), and small MSE (0.0019 − 0.0045). Gumbel D provides good estimates for modest and strong correlation ($\rho = 0.25, 0.5, 0.75$), but performs worse when the true correlation is small ($\rho = 0.05$). For weak correlation ($\rho = 0.1$), Gumbel D and Frank A provides estimates close to the true value with corresponding bias 0.001 − 0.002, while estimates from Gumbel D have slightly higher than optimal 95% coverage probabilities. For modest ($\rho = 0.25$) and strong ($\rho = 0.5, 0.75$) correlations, all methods provide similar results while the copula-based methods perform slightly better than GEE in terms of bias. In sum, the copula approach with Frank A performs the best or as well as the best in estimating correlations with respect to bias. Gumbel D does slightly better than Frank A when ρ is 0.5 in terms of coverage. The standard errors from all methods are similar.

3.3 Bias and efficiency in estimating covariate effects on dependence

In order to examine the performance of the copula-based methods in estimating the covariate effect on the dependence parameter, we also perform a simulation study. We create one binary variable for covariate, and adjust for it on the dependence parameter. Thus, as described in subsection 2.2, we have two regression coefficients, b_0 and b_1, on the dependence parameter, where b_1 represents a regression coefficient for the covariate. We set the value of b_0 as 0.262

Table 3. The average estimates and standard errors of covariate coefficients on the dependence parameter using copula approach

True	Copula	b_0				b_1				
$b = (b_0, b_1)$	family	Est.	S.E.	Bias	$CP_{95\%}$	Est.	S.E.	Bias	$CP_{95\%}$	p-value*
(0.262, 0.010)	Gumbel D	0.266	0.082	0.004	94.0	0.010	0.117	0.000	94.4	0.506
	Frank A	0.265	0.089	0.008	94.6	0.011	0.128	0.001	94.6	0.536
(0.262, 0.300)	Gumbel D	0.266	0.082	0.003	94.0	0.310	0.137	0.010	94.4	0.102
	Frank A	0.266	0.089	0.002	94.2	0.314	0.154	0.014	93.6	0.128
(0.262, 1.000)	Gumbel D	0.264	0.082	0.001	93.6	0.979	0.203	0.021	91.2	0.001
	Frank A	0.263	0.089	0.001	94.4	1.043	0.229	0.043	94.0	0.000

* p-value obtained from the Wald test; '$CP_{95\%}$' = Coverage probability of 95% confidence interval

which corresponds to copula parameter $\theta = 1.3$, and represents moderate level of correlation. For each simulation, we set the covariate coefficient, b_1, 0.01, 0.30 and 1.00, respectively, where the covariate effect ranges from small to large with corresponding p-value from large to small. We also use Gumbel D and Frank A. We present bias and 95% coverage probability to show their performance (Table 3). We could fit the dependence model using different levels of regression coefficients and different values of b_0 (not presented here), but the results appear consistent.

Both methods provide good estimators of both regression coefficients on the dependence parameter. Gumbel D performs slightly better than Frank A with respect to bias, while Frank A does better than Gumbel D with respect to coverage. As expected, mean p-values for b_1 decrease when true value of b_1 gets bigger. Significant p-value (<0.05) represents a significant difference in dependency among two groups we adjust for on the dependence parameter. We do not present the performance of copula-based approach in estimating marginal probability, but it appears to provide unbiased estimates (bias $= 0.000 - 0.002$).

4. APPLICATION: HIV RISK REDUCTION STUDY FOR SERODISCORDANT COUPLES (AAC)

In this section, we analyze data from AAC, a randomized trial of HIV serodiscordant African-American couples applying the copula approach. In the case of the AAC couples data, we will model a bivariate joint distribution considering the correlated responses from male and female partners as bivariate outcomes. Specifically, each partner is asked to report on shared sexual behaviors with his/her study partner in the past 90 days. By construction, partner responses are expected to be correlated and a measure of the magnitude of this correlation serves as a measure of reliability of these self-reported, sexual behavior outcomes.

As noted, we focus more on the dependency parameters than the marginal parameters, so that the estimated dependency can serve as a proxy for reliability of self-reported sexual behaviors, which are shared in this couple-focused context. The following subsections describes the study design, data, characteristics of study participants and outcomes of interest. In the last subsection, the results of the analysis applying the copula approach are summarized.

4.1 Study design and data

We use baseline data from AAC, a two-arm, couple-based randomized controlled intervention trial of HIV serodiscordant African-American couples from four cities in the US (Atlanta, GA; Los Angeles, CA; New York, NY; and Philadelphia, PA). The study was designed to test the efficacy of a couple-focused HIV/STD risk reduction intervention compared to an individual-focused health promotion intervention in reducing sexual risk behaviors and STD incidence [9, 10].

The study included 535 couples (1,070 individuals) recruited from HIV care clinics, HIV testing and counseling sites, primary care clinics, AIDS services organizations, substance abuse treatment programs, churches and HIV/AIDS ministries, HIV/AIDS providers and community-based coalitions and advocacy organizations. Data were obtained from three sources. First, participants completed a 90-minute Audio Computer-Assisted Self Interview (ACASI), which assessed sociodemographic and relationship characteristics and sexual behaviors, including condom use. Items assessing sexual behaviors were worded so that they were appropriate for each gender.

4.2 Characteristics of study participants

Study partners were asked to indicate their age (in years), education, income, health insurance status and incarceration history. HIV status at baseline was determined via biological testing in order to confirm that couples were HIV serodiscordant. Study participants were also asked questions that addressed relationship characteristics including length of relationship with their study partner, whether or not participants were married to their study partner (yes/no), and items assessing sexual dysfunction (yes/no). To illustrate previously described copula methods in this data, we created categorized couple variables with 3 levels for the following items: high school graduate, income (over \$850/month), insurance, incarceration history indicating whether each characteristic was observed in neither, one or both partners

Table 4. Dependence parameter estimates adjusting for each covariate using max-id copula with Frank A and GEE2

	Frank A			GEE2		
Covariate	Est.	S.E.	p-value	Est.	S.E.	p-value
Age	-0.002	0.010	0.874	0.000	0.007	0.996
Education	0.225	0.133	0.092	0.186	0.099	0.061
Income	0.128	0.102	0.206	0.094	0.070	0.179
Insurance	0.176	0.081	0.029	0.167	0.079	0.035
Incarceration	0.091	0.100	0.358	-0.014	0.104	0.891
HIV	0.007	0.134	0.959	-0.050	0.100	0.617
Duration	-0.171	0.135	0.205	-0.146	0.099	0.140
Married	-0.021	0.136	0.876	-0.014	0.104	0.891
Sexual dysfunction	-0.087	0.131	0.505	-0.068	0.100	0.499

within each couple. HIV status refers to whether the female partner was the HIV positive partner. These 9 items were considered covariates of interest in measuring the dependence parameter. Because our primary interest was to estimate the dependence parameter, the only covariate used in estimating the marginal probability of the outcome of interest was randomized treatment assignment.

4.3 Primary outcome

Participants provided data on the use of male and female condoms during sex they had engaged in with study partners over the past 90 days and the proportion of condom-protected sex was constructed using self-reported number of sexual episodes and number of self-reported times condoms were used during these episodes. For the purpose of illustrating the present copula methodology and because it is a common primary endpoint in HIV/STD risk modification trials, we constructed an outcome, 'consistent condom use', that equals one when condom use was reported at every sexual episode with study partner and zero otherwise, for both male and female partners.

5. RESULTS

5.1 Correlations according to different level of covariates

Our main purpose in this study is to determine how the associations between male and female consistent condom use responses might vary across different sub-populations and to estimate the reliability of self-reported outcomes in the unique context of a couple-based behavioral study. We fit the model that incorporates the covariates of interest in the copula parameter as described in subsection 4.2. Our preliminary crude analysis using `PROC CORR` showed that the estimated couple correlation of reported consistent condom use was 0.34. From the previous simulation study with modest correlation, both Gumbel D and Frank A performed similarly. We arbitrarily fit our model using the Frank A copula since we expect which copula family we choose does not affect the results based on simulations in the previous section.

Table 4 summarizes the results of the estimated regression coefficients of covariates on the dependence parameter, θ. The results suggest a statistically significant difference in the copula dependence parameter between couples where both have health insurance compared to those where neither or either has health insurance (copula approach: $p = 0.029$; GEE2 approach: $p = 0.035$). Couples where both partners are insured have more correlated reports of consistent condom use than those where neither or either has insurance. In addition, education, income and duration of relationship have arguably marginal p-values indicating there may be differences in the correlation of male and female partner responses to reported consistent condom use between subgroups defined by these covariates. Also, there were no observed differences on estimated dependence based on age, incarceration history, HIV or marital status, or reported sexual dysfunction.

Table 5 summarizes the odds ratios and correlation across different levels of the covariates estimated from its corresponding dependence parameter. The average odds ratio and correlation between female and male partners are 5.59 and 0.34, respectively. As expected, as sociodemographic indices such as education, income, insurance status and no incarceration history increase, the correlation between couples response increases. The correlation where both have insurance is 0.401, while the correlation where neither have insurance is 0.099, demonstrating the impact of these covariates on correlation.

6. CONCLUSIONS AND DISCUSSION

In this work, we applied a bivariate copula-based logit model to data from the HIV serodiscordant couples study and estimated the correlation of the couples' responses according to the different covariates. We have illustrated a novel application of estimating dependency, while adjusting for covariates, as an estimator of reliability of self-reported shared sexual behavior from a couple-based study. Prior to the application, we conducted a simulation study to examine how copula-based models perform relative to GEE (first- and second-order, where applicable), and to determine which copula family works well in a number of settings

Table 5. Estimated odds ratio of consistent condom use and corresponding correlation of covariate levels on dependence parameter using max-id, Frank A copula

	Odds ratio		Correlation	
	Treatment group Est. (S.E.)	Control group Est. (S.E.)	Treatment group Est. (S.E.)	Control group Est. (S.E.)
Age	5.566 (1.689)	5.572 (1.723)	0.341 (0.056)	0.341 (0.056)
One or more partner HS grad/GED				
No	3.515 (1.334)	3.525 (1.342)	0.247 (0.073)	0.247 (0.074)
Yes	8.107 (3.286)	8.142 (3.276)	0.416 (0.067)	0.416 (0.066)
Income > $850 per month				
Neither partner	4.120 (1.484)	4.117 (1.455)	0.279 (0.064)	0.280 (0.065)
One partner	6.667 (2.439)	6.657 (2.434)	0.377 (0.062)	0.378 (0.062)
Both partners	10.202 (8.625)	10.174 (8.651)	0.458 (0.103)	0.460 (0.102)
Insurance Status				
Both uninsured	1.709 (1.587)	1.708 (1.491)	0.099 (0.378)	0.100 (0.360)
One uninsured	3.958 (1.173)	3.941 (1.167)	0.269 (0.059)	0.271 (0.059)
Both insured	7.645 (2.762)	7.588 (2.732)	0.401 (0.060)	0.404 (0.060)
Incarceration history				
Neither partner	3.902 (2.677)	3.911 (2.702)	0.269 (0.108)	0.269 (0.109)
One partner	5.545 (1.544)	5.558 (1.533)	0.341 (0.050)	0.342 (0.051)
Both partners	7.629 (4.037)	7.644 (4.049)	0.404 (0.084)	0.405 (0.08)5
HIV positive female partner				
No	5.383 (2.395)	5.363 (2.371)	0.333 (0.076)	0.335 (0.076)
Yes	5.537 (2.001)	5.516 (1.953)	0.339 (0.065)	0.341 (0.065)
Relationship w. study partner > 5 yrs				
No	7.654 (3.182)	7.599 (3.156)	0.402 (0.071)	0.404 (0.072)
Yes	4.046 (1.539)	4.029 (1.520)	0.274 (0.070)	0.276 (0.070)
Married				
No	5.636 (1.792)	5.614 (1.774)	0.343 (0.059)	0.344 (0.060)
Yes	5.185 (2.638)	5.167 (2.581)	0.326 (0.086)	0.327 (0.086)
One or more partner sexual dysfunction				
No	6.297 (2.333)	6.275 (2.294)	0.365 (0.065)	0.366 (0.066)
Yes	4.573 (2.007)	4.562 (1.971)	0.300 (0.077)	0.301 (0.077)

with varying levels of underlying correlations and covariates on the copula parameter. We also compared the results of the copula-based method with those of GEE. In prior work with copulas, many have used the Kendall's τ or copula parameters as the measure of concordance. Since neither the Kendall's τ nor copula parameter is directly comparable to moment estimates for the correlation coefficient from GEE, we introduced a copula-based estimator of binary correlation and odds ratio to compare the moment estimates for the correlation coefficient from GEE. Based on the results from the simulation study, we found that most of the models with copula families perform well and provide similar results for moderate and strong correlation. Frank A performed well for the weak correlation, however, Gumbel D did not do well when true ρ is 0.05. In terms of small bias, both copula models performed better than GEE when true correlations are 0.25 and 0.5. For correlation 0.1, Frank A worked the best. In terms of 95% coverage rate, the results from all methods were similar for all levels of correlation except for Gumbel D when true $\rho = 0.05$. Both methods with Gumbel D and Frank A also perform well in estimating the regression coefficient on the dependence parameter. Gumbel D performs better in terms of bias, while Frank A does better in terms of coverage.

Finally, we fitted copula-based models to our data and focused on estimation of the correlation between responses of consistent condom use from couples adjusting for couple-level covariate information. We selected 9 different couple-based covariates. The findings show that there is a statistically significant difference in the correlation between couples where both have insurance and those where neither or either has insurance. Couples where both have insurance have more correlated outcomes than those where neither or either has insurance. Among couples where females and males have high school diplomas compared to couples where neither or either has a diploma, the responses are more highly correlated. Couples where both have high income (>$850) are likely to have more correlated outcomes than those where neither or either has high income. Interestingly, in terms of relationship duration, the responses from couples with more than a 5-year relationship seem to be less correlated, whether the female partner is HIV infected or not does not affect the correlation. These findings suggest that we need to pay attention to those couples with covariates such as no insurance, low income and low education, indicating low correlation, to improve the reliability of self-reports.

To summarize, this work provide a good measure of reliability of self-reported sexual behaviors among HIV serodiscordant couples by estimating correlation using a copula-based method. This work also introduces systematic research on the influence of the factors on the responses of self-reported sexual behaviors. Thus, we can see the magnitude of matching responses based on the estimated correlation adjusting for important covariates of interest, which can tell us the reliability of paired or couple-based self-reported data.

This approach has the advantages of constructing separate models for the marginal probabilities and the dependence parameters, which is more efficient. In addition, this method is fully specified allowing joint and conditional probabilities to be derived easily, and is straightforward to apply using a standard and direct maximum likelihood inference procedure. R code is available from the authors. Also, it allows us to model the dependence parameter with covariate information of interest without computational difficulty. Therefore, it leads to a better understanding of couple-level issues related to self-reported sexual behaviors.

7. TRIAL REGISTRATION

clinicaltrials.gov Identifier: NCT00644163.

ACKNOWLEDGEMENTS

We gratefully acknowledge the helpful comments provided by Reviewers and Editors.

Received 31 January 2014

REFERENCES

[1] Bellamy, S. and the NIMH Multisite HIV/STD Prevention Trial for African American Couples Study Group (2005). A dynamic block-randomization algorithm for group-randomized clinical trials when the composition of blocking factors is not known in advance. *Contemporary Clinical Trials* **26** 469–479.

[2] Burton, J., Darbes, L. and Operario, D. (2008). Couple-focused behavioral interventions for prevention of HIV: systematic Review of the state of evidence. *AIDS and Behavior* **14** 1–10.

[3] Carey, V., Zeger, S. and Diggle, P. (1993). Modelling multivariate binary data with alternating logistic regressions. *Biometrika* **80** 517–526.

[4] Catania, J., Gibson, D., Chitwood, D. and Coates, T. (1990). Methodologic problems in AIDS behavioral research: influences on measurement error and participation bias in studies of sexual behavior. *Psychol Bull* **108** 339–362.

[5] Denuit, M. and Lambert, P. (2005). Constraints on concordance measures in bivariate discrete data. *Journal of Multivariate Analysis* **93** 40–57. MR2119763

[6] El-Bassell, N., Jemmott III, J., Landis, J., Pequegnat, W., Wingood, G., Wyatt, G., Bellamy, S. and the NIMH Multisite HIV/STD Prevention Trial for African American Couples Group (2011). Intervention to influence behaviors linked to risk of chronic diseases: a multisite randomized controlled trial with African-American HIV-serodiscordant heterosexual couples. *Archives of Internal Medicine* **171** 728–736.

[7] El-Bassell, N., Jemmott III, J., Wingood, G., Wyatt, G., Pequegnat, W., Landis, J., Bellamy, S. and the NIMH Multisite HIV/STD Prevention Trial for African American Couples Group (2010). NIMH multisite eban HIV/STD prevention intervention for African American HIV serodiscordant couples: a cluster randomized trial. *Archives of Internal Medicine* **17** 1594–1601.

[8] NIMH Multisite HIV/STD Prevention Trial for African American Couples Group (2010). Concordant and discordant reports on shared sexual behaviors and condom use among African American serodiscordant couples in four cities. *AIDS and Behavior* **14** 1011–1022.

[9] NIMH Multisite HIV/STD Prevention Trial for African American Couples Group(a) (2008). Eban risk reduction intervention: conceptual basis and procedures. *Journal of Acquired Immune Deficiency Syndromes* **49** Suppl 1:S15–27.

[10] NIMH Multisite HIV/STD Prevention Trial for African American Couples Group(b) (2008). Eban health promotion intervention: conceptual basis and procedures. *Journal of Acquired Immune Deficiency Syndromes* **49** Suppl 1:S28–34.

[11] Genest, C. and Nešlehová, J. (2007). A primer on copulas for count data. *The Astin Bulletin* **37** 475–515. MR2422797

[12] Hofert, M., Kojadinovic, I., Maechler, M. and Yan, J. (2014). copula: multivariate dependence with copulas R package version 0.999-10.

[13] Joe, H. (1997). *Multivariate Models and Dependence Concepts.* Chapman & Hall, London. MR1462613

[14] Joe, H. and Hu, T. (1996). Multivariate distributions from mixtures of max-infinitely divisible distributions. *Journal of Multivariate Analysis* **57** 240–265. MR1391171

[15] Liang, K. and Zeger, S. (1986). Longitudinal data analysis using generalized linear models. *Biometrika* **73** 13–22. MR0836430

[16] Liang, K., Zeger, S. and Qaqish, B. (1992). Multivariate regression analyses for categorical data. *Journal of the Royal Statistical Society, Series B* **54** 3–40. MR1157713

[17] Lipsitz, S., Laird, N. and Harrington, D. (1991). Generalized estimating equations for correlated binary data: using the odds ratio as a measure of association. *Biometrika* **78** 153–160. MR1118240

[18] Nelsen, R. (2006). *Introduction to Copulas.* Springer, New York, NY. MR2197664

[19] Nikoloulopoulos, A. and Karlis, D. (2008). Multivariate logit copula model with an application to dental data. *Statistics in Medicine* **27** 6393–6406. MR2655124

[20] Nikoloulopoulos, A. and Karlis, D. (2008). Finite normal mixture copulas for multivariate discrete data modeling. *Journal of Statistical Planning and Inference* **139** 3878–3890. MR2553774

[21] Ochs, E. and Binik, Y. (1999). The use of couple data to determine the reliability of self-reported sexual behavior. *Journal of Sex Research* **36** 374–384.

[22] Qaqish, B. (2003). A family of multivariate binary distributions for simulating correlated binary variables with specified marginal means and correlations. *Biometrika* **90** 455–463. MR1986660

[23] Seal, D. (1997). Interpartner concordance of self-reported sexual behavior among college dating couples. *Journal of Sex Research* **34** 39–55.

[24] Sklar, A. (1959). Fonctions de répartition à n dimensions et leurs marges. *Publications de L'Institut de Statistique de L'Université de Paris* **8** 229–231. MR0125600

[25] Song, P. (2000). Multivariate Dispersion models generated from Gaussian copula. *Scandinavian Journal of Statistics* **27** 305–320. MR1777506

[26] Streiner, D. and Norman, G. (1995). *Health Measurement Scales: Practical Guide to Their Development and Use*, 3rd ed. Oxford University Press, New York, NY.

[27] Witte, S., El-Bassel, N., Gilbert, L., Wu, E. and Chang, M. (2007). Predictors of discordant reports of sexual and HIV/sexually transmitted infection risk behaviors among heterosexual couples. *Sexually Transmitted Diseases* **34** 302–308.

[28] Jun Yan (2007). Enjoy the joy of copulas: with a package copula. *Journal of Statistical Software* **21** 1–21.

Scarlett L. Bellamy
Department of Biostatistics and Epidemiology
Perelman School of Medicine
University of Pennsylvania
Philadelphia, PA
USA
E-mail address: bellamys@mail.med.upenn.edu

Seunghee Baek
Department of Clinical Epidemiology and Biostatsitics
Asan Medical Center
Ulsan College of Medicine
Seoul
Korea
E-mail address: sbaek@amc.seoul.kr

Andrea B. Troxel
Department of Biostatistics and Epidemiology
Perelman School of Medicine
University of Pennsylvania
Philadelphia, PA
USA
E-mail address: atroxel@mail.med.upenn.edu

Thomas R. Ten Have
Department of Biostatistics and Epidemiology
Perelman School of Medicine
University of Pennsylvania
Philadelphia, PA
USA

John B. Jemmott III
Annenberg School for Communication
Center for Health Behavior and Communication Research
University of Pennsylvania
Philadelphia, PA
USA
E-mail address: jjemmott@asc.upenn.edu