# NOVEL STATISTICAL MODELS FOR ECOLOGICAL MOMENTARY ASSESSMENT STUDIES OF SEXUALLY TRANSMITTED INFECTIONS

Fei He

Submitted to the faculty of the University Graduate School
in partial fulfillment of the requirements
for the degree
Doctor of Philosophy
in the Department of Biostatistics,
Indiana University

September 2016

Accepted by the Graduate Faculty, Indiana University, in partial fulfillment of the requirements for the degree of Doctor of Philosophy.

Jaroslaw Harezlak, Ph.D., Chair

Doctoral Committee

Ziyue Liu, Ph.D.

Patrick Monahan, Ph.D.

July 18, 2016

Devon J Hensel, Ph.D.

DEDICATION

*To my loving parents and husband*

ACKNOWLEDGMENTS

I would like to express sincere gratitude to my advisor Dr. Harezlak for his constant guidance, encouragement and support in my study. Jarek, you are such an amazing advisor that makes many students jealous about me being your student. I want to thank the members of my PhD committee, Dr. Hensel, Dr. Liu and Dr. Monahan for their helpful suggestions.

I also would like to mention few professors who are always be so supportive during my study time. Dr. Fortenberry, thank you for giving me such a great opportunity to learn so much from the research related to STIs and that turned out to be my passionate research area. You always give me time and space to think and express my ideas and respect my opinion. I couldn't expect more trust from you. Dr. Boukai, I always remember what you told me when I just started my life in the U.S. "Keep asking questions, practice English and learn how to drive." I gain so much confidence from your encouragement and wisdom from our conversation. See how far that young girl goes! Dr. Katz, I really appreciate your understanding and support during the time I took care of my mom. That means a lot to me. Dr. Teixeira-Pinto, though we haven't met in person yet, your warm greeting and smile from Sydney are indispensable part of this dissertation. Hope to see you soon!

Besides all the amazing mentors I met here, I would like to thank our wonderful department! All the faculty members, staffs and my dear colleagues, especially those I meet every Friday. I am just too lucky to have you guys around. Your intelligence, warm care and sense of humor will be the part I miss the most. I won't forget those friends who are being so helpful and thoughtful to me during the past seven years in Indy. Without you guys, who I can share my happiness and sorrow with, who can I share my sparkling thoughts and funny stories with. Jeremy, get your guitar, let's sing together. Giorgos, we still need one more basketball game to decide who will be the TRUE MVP. Hannah, how about another

museum tour and romantic dinner in Seattle? Han, the best cook should ignore all the critics.

There is a group of people at the other side of Pacific Ocean to whom I want to show my gratitude. To those family members who always be the strongest backing since my childhood, to those friends who has been taking great care of my mom when I am away, to those people who treated me with sincere hearts in my life.

Last but not the least, my dear parents, thank you for bring me to this lovely world and letting me feel and understand it in my own way. Dad, I know you are going to be so proud of me if you could watch my growth all the way to today. Mom, you are the wisest, strongest, kindest and most beautiful woman in my mind. Do you know you are my idol? My loved husband, being with you is such a splendid journey. Your love colors my life.

Words are so limited when a person full of enthusiasm is trying to show appreciation of her incredible life. I feel so proud of the profound culture of my country China and so delighted with the experience of different cultures in the U.S. I will always miss my life in beautiful Indianapolis and am looking forward to starting my new adventure in the blooming Seattle.

Fei He

NOVEL STATISTICAL MODELS FOR ECOLOGICAL MOMENTARY ASSESSMENT STUDIES OF SEXUALLY TRANSMITTED INFECTIONS

The research ideas included in this dissertation are motivated by a large sexually transmitted infections (STIs) study (IU Phone study), which is also an ecological momentary assessment (EMA) study implemented by Indiana University from 2008 to 2013. EMA, as a group of methods used to collect subjects' up-to-date behaviors and status, can increase the accuracy of this information by allowing a participant to self-administer a survey or diary entry, in their own environment, as close to the occurrence of the behavior as possible. IU Phone study's high reporting level shows one of the benefits gain from introducing EMA in STIs study. As a prospective study lasting for 84 days, participants in IU Phone study undergo STI testing and complete EMA forms with project-furnished cellular telephones according to the predetermined schedules. At pre-selected eight-hour intervals, participants respond to a series of questions to identify sexual and non-sexual interactions with specific partners including partner name, relationship satisfaction and sexual satisfaction with this partner, time of each coital event and condom use for each event. etc. STIs lab results of all the participants are collected weekly as well. We are interested in several variables related to the risk of infection and sexual or non-sexual behaviors, especially the relationship among the longitudinal processes of those variables. New statistical models and applications are established to deal with the data with complex dependence and sampling data structures. The methodologies covers various of statistical aspect like generalized mixed models, multivariate models and autoregressive and cross-lagged model in longitudinal data analysis, misclassification adjustment in imperfect diagnostic tests, and variable-domain functional

regression in functional data analysis. The contribution of our work is we bridge the methods from different areas with EMA data in the IU Phone study and also build up a novel understanding of the association among all the variables of interest from different perspectives based on the characteristic of the data. Besides all the statistical analyses included in this dissertation, variety of data visualization techniques also provide informative support in presenting the complex EMA data structure.

Jaroslaw Harezlak, Ph.D., Chair

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

# Chapter 1

## Introduction

Assessment in behaviors typically relies on global retrospective self-reports collected at research or clinic visits, which are limited by recall bias and are not well suited to address how behavior changes over time and across contexts. In order to minimize recall bias, maximize ecological validity, and allow study of instant processes that influence behavior in real-world circumstances, ecological momentary assessment (EMA) is introduced. It is a group of methods using repeated collection of real-time data on subjects' behavior and experience in their natural environments that involves repeated random sampling of subjects' up-to-date behaviors and status at periodic intervals. EMA studies use technologies ranging from written diaries and telephones to electronic diaries and physiological sensors.

Between 2008 and 2013, a large longitudinal study (IU Phone Study) with the purpose of examining sexual or non-sexual risk behaviors and incident sexually transmitted infections(STIs) was implemented in Indianapolis, which is a good example of using EMA collecting the STIs related data. EMA can increase the accuracy of this information by allowing a participant to self-administer a survey or diary entry, in their own environment, as close to the occurrence of the behavior as possible. This way of data collection garners less missing data, higher reporting levels, stronger internal data validity and low behavior reactivity. EMA also strengthens the security of sensitive or stigmatizing information and increases participant valued privacy.

IU Phone Study was a prospective study lasting for 84 days (12 weeks). Participants were recruited from the patient population of a county sexually transmitted diseases clinic but were not necessarily clinic patients at the time of enrollment. Eligibility criteria were

ages 18 to 29 years (inclusive), English speaking, and planning to reside in the area for the subsequent 84 days. The Institutional Review Board of Indiana University Purdue University Indianapolis approved this study. All participants provided informed consent.

The primary mode of data collection is via 8-hour self-reports of coital and non-coital sexual behaviors, condom use, and relationship assessments, which were recorded with project-furnished cellular telephones and service. The expected number of entries was thus 252 entries per participant. Our previous summary showed the daily diary completion rate of IU Phone Study was 87.7% from Hensel et al. (2012). At pre-selected 8-hour intervals, participants responded to a series of questions to identify sexual and non-sexual interactions with specific partners. In each eight-hour reporting period, participants identified partner name, relationship satisfaction and sexual satisfaction with this partner, time of each coital event (up to four events within the same eight-hour reporting period) and condom use for each event. On the other side, STIs lab results of all the participants were collected as well. Commercially available nucleic acid amplification test (NAAT) was used to test C trachomatis (CT), N gonorrhoeae (GC), and T vaginalis (TV). Participants received NAAT before the entry of the study. Treatments were provided to those participants who showed positive result at the enrollment test. Weekly self-obtained vaginal or urine samples were collected since the beginning time point of the study for 12 weeks. Including the STI diagnosis at the enrollment, each participant supposed to have 13 lab tests. All the samples were kept in the lab and were not tested until the end of 12-weeks in order to avoid possible intervention of natural observation. During the study time, participants were allowed to visit the clinics or take medicines if they want to. No participant received any treatment or took any medicine except the ones provided at the entry of the study. In the future analyses which involve using the STI test information, participants with any positive results in NAAT at enrollment are excluded in order to avoid a potential confounding effect due to treatment.

IU Phone Study was implemented through EMA methods as a novel design of longitudinal study in STIs area. There was a great amount of variables collected in this study with complex dependence and sampling data structures. We are interested in several variables related to the risk of infection and sexual or non-sexual behaviors, especially the relationship among the longitudinal processes of those variables. We have established new statistical models and also extended existing statistical models to deal with data from EMA study with complex dependence and sampling data structures. The methodologies covers various of statistical aspect like generalized mixed models, multivariate models and autoregressive and cross-lagged model in longitudinal data analysis, misclassification adjustment in imperfect diagnostic tests, and variable-domain functional regression in functional data analysis. Our goal is trying to explore the relationship between some variables of interest, and establish appropriate models to have a better understanding of the association between those variables from different perspectives based on the characteristic of the data set.

In the first paper, we study the dependence between the condom use and sexual satisfaction based on the EMA data reported in IU Phone Study through an extended autoregressive and cross-lagged models. Though autoregressive and cross-lagged models have been widely used to understand the relationship between bivariate commensurate outcomes in social and behavioral sciences, not much work has been done in modeling bivariate non-commensurate outcomes simultaneously. We develop a likelihood-based methodology combining ordinary autoregressive and cross-lagged models with a shared subject-specific random effect in the mixed model framework to model two correlated longitudinal non-commensurate outcomes. Inclusion of the subject-specific random effects in the proposed model accounted for between-subject variability arising from the omitted subject-level predictors.

In the second paper, we are interested in associating the changing pattern of condom use and sexual satisfaction during each observed partnership with the partner change status (with or without the partner change). Since participants in the study have different length of partnership, variable-domain functional regression (VDFR), which is a class of scalar-on-function regression models with partner-specific functional predictor domains, is introduced. Though those VDFR models are developed from the functional data analysis first, we find a strong connection between longitudinal data and functional data in terms of estimating the association between a trajectory against time and an outcome of a status. As the result of that, we apply the lagged time model in VDFR to the IU Phone Study with condom use and sexual satisfaction as two functional predictors to estimate the probability of changing a partner. Other nonfunctional covariates like gender are also included for adjustment. This is also an extension of previous author's application by including two functional predictors simultaneously rather than one.

In the third paper, we develop a method to estimate the covariate-specific probability of positivity of imperfect STIs tests. Specifically, in the IU Phone Study, we plan to use the previous and current test results to estimate the current true status of STIs based on an autoregressive (AR)(1) model. The methodology developed previously either provides large bias or are quite complex to be used in the clinical practice. We develop a likelihood based expectation-maximization algorithm to predict covariates specific estimated probability of positivity for the data with misclassified response and covariates in both cross sectional data and longitudinal data. Our model emphasizes the clinical interest in estimating the covariates specific estimated probability of positivity for certain groups of population and individuals in practical life. In the real data application, we also implement a sensitivity analysis to provide possible range of estimated probability of positivity based on different combination of sensitivity and specificity.

In the fourth paper, we are interested in assessing condom use as a function of number of coital events in newly formed sexual relationships. Statistical analyses are based on the generalized additive mixed models (GAMMs) that use smooth functions to model the mean trajectory and account for the hierarchical structure of longitudinal data. To apply GAMMs to our data, we included two nested random effects (at a partner level and a subject level respectively) to account for correlations among repeated within-partner coital events and correlations among the partners of the same subject. Specifically, a logistic additive mixed model is used to estimate the association between the event-specific condom use (coded as no/yes), cumulative number of coital events and other covariates of interest (relationship satisfaction, sexual satisfaction and gender). Instead of using parametric method of modeling condom use probability with cumulative number of coital events, we used a smoothing function as a more flexible, data-driven nonparametric approach.

This chapter provides the background information of IU Phone Study and brief introduction of all the works we cover in this dissertation. In the following chapters, all the four papers mentioned above are presented in order with more details. And the summary of this dissertation is included as the conclusion chapter.

Chapter 2

# Autoregressive and Cross-lagged Models for Bivariate Non-commensurate Outcomes

## 2.1 Introduction

In multivariate longitudinal data analysis, there are multiple questions of interest: longitudinal change of each outcome, auto-dependence on the past observations of each outcome, correlation between two or more outcomes and cross-dependence between the longitudinal changes of two outcomes. Many research papers have addressed these questions separately, but there is a scarcity of the models developed to deal with the cases that involve all four questions simultaneously. In this paper, our interest is to establish a method to jointly model correlated bivariate non-commensurate longitudinal outcomes and to estimate their auto- and cross-dependence.

In the 1960s and 1970s, Campbell (1963), Bohrnstedt (1969), Duncan (1969), Heise (1969), and Joreskog and Sorbom (1979) discussed some early social science examples about the autoregressive and cross-lagged models for two or more outcome variables in panel data. Kessler and Greenberg (1981) further developed the methodology to support those examples. These have been and continued to be popular modeling approaches for longitudinal data. Bollen and Curran (2004) also developed the model combining both autoregressive model and cross-lagged model under structural equation model framework with latent variables introduced. Several examples with application of those models can be found in the psychology and behavior literature. We find that majority of the previous research using autoregressive and cross-lagged models is based on the commensurate outcomes, but no specific algorithm has been developed to deal with non-commensurate outcomes. In order to estimate the dependence among different types of outcomes, we borrow the idea of joint

modeling of non-commensurate outcomes from multivariate methods and incorporate it in the autoregressive and cross-lagged models.

The challenge of dealing non-commensurate outcomes with multivariate methods is the nonexistence of obvious multivariate distributions. To address this problem, two general likelihood-based approaches have been proposed to avoid the direct specification of the joint distribution of the outcomes: 1) factorization of the joint distribution of the outcomes and 2) introduction of random effects to model the correlation among the multiple outcomes. Conditional models allow the joint distribution to be factorized into a marginal component and a conditional component. Such approaches have been discussed in Tate (1954), Olkin and Tate (1961), Little and Schluchter (1985), Krzanowski (1988), and Cox and Wermuth (1992)Cox and Wermuth (1994). On the other hand, Teixeira-Pinto and Normand (2009) proposed a multivariate method to analyze correlated binary and continuous outcomes using probit-type model, which was established by Catalano and Ryan (1992), with a shared random variable introduced in cross-sectional data. They have shown that this model is equivalent to the factorization model presented by Catalano and Ryan (1992) and the estimates could be obtained via mixed models equivalence.

In this paper, we present autoregressive and cross-lagged model for correlated bivariate non-commensurate longitudinal outcomes under complex dependence and sampling data structures. The path graph of autoregressive and cross-lagged model for one continuous outcome and one binary outcome is shown in Figure 2.1. Our interest is estimating the cross-lagged effect $\beta^{[c]}$ and $\beta^{[b]}$ as well as the autoregressive effects $\gamma^{[c]}$ and $\gamma^{[b]}$. This work is motivated by a sexually transmitted infection (STI) ecological momentary assessment (EMA) study (IU Phone Study). Participants in this study have undergone STI testing and completed EMA forms according to the predetermined schedules. We are interested in analyzing several variables related to the risk of infection and sexual or non-sexual behav-

iors. In particular, we are interested in the association between the longitudinal processes of two correlated variables (e.g.condom use behavior and sexual satisfaction) in the study. The design of the IU Phone Study and the data structure are introduced in Section 2.2. In Section 2.3, the construction of the proposed model and the algorithm of estimation are described. We first develop a model for two continuous outcomes, then extend it to one continuous outcome and one binary outcome. The simulation study under different data generating mechanisms are implemented to compare the proposed model with univariate generalized linear models (GLMs) in Section 2.4. We apply the proposed model and univariate GLMs to condom use (binary variable) and sexual satisfaction (continuous variable) from IU Phone Study in Section 2.5. The conclusions and limitations are discussed in Section 2.6.

### 2.2 Motivating data set: the IU Phone Study

Data were obtained from a prospective 84-day (12-week) study designed to examine sexual behaviors and incident STI. Participants were recruited from the patient population of a county sexually transmitted diseases clinic but were not necessarily clinic patients at the time of enrollment. Eligibility criteria were ages 18 to 29 years (inclusive), English speaking, and planning to reside in the area for the subsequent 84 days. The Institutional Review Board of Indiana University Purdue University Indianapolis approved this study. All participants have provided informed consent.

Ecological momentary assessment (EMA) diary technology was used in the study. The primary mode of data collection was via three-times daily self-reports of coital and non-coital sexual behaviors, condom use, and relationship assessments, recorded with project-furnished cellular telephones and service. The expected number of entries was thus 252 entries per participant. Daily diary completion rate of this EMA diary data was 87.7%.

Other methodological details were previously published in Hensel et al. (2012). At pre-selected 8-hour intervals, participants responded to a series of questions to identify sexual and non-sexual interactions with specific partners. In each eight-hour reporting period, participants identified any partner, time of each coital event (up to four events within the same eight-hour reporting period), condom use for each coital event, as well as relationship satisfaction and sexual satisfaction.

There were 254 participants who have more than one coital events with partner(s) during the reporting period included in our analysis. All the included individuals completed the whole study and were not jailed during the study period, 58% (147/254) of the participants were women and 90% (229/254) were African American.

### 2.3 Models and methodology

In our proposed autoregressive and cross-lagged model with two outcomes, we include each outcome's observations at time $(t-1)$ as predictors of the other outcome's observations at time $t$. The coefficients of autoregressive effect and cross-lagged effect are assumed to be constant between every two adjacent observation time. We start by establishing the model for two continuous outcomes case and then move to the non-commensurate setting with one continuous and one binary outcome case.

#### 2.3.1 Two continuous outcomes

Let $y_{i,t}^{[c1]}$ and $y_{i,t}^{[c2]}$ be two normally distributed correlated continuous outcomes. The structural regression models of $y_{i,t}^{[c1]}$ and $y_{i,t}^{[c2]}$ are

$$\begin{aligned} y_{i,t}^{[c1]}|y_{i,t-1}^{[c2]}, y_{i,t-1}^{[c1]}, u_i &= \alpha^{[c1]} + \beta^{[c1]} y_{i,t-1}^{[c2]} + \gamma^{[c1]} y_{i,t-1}^{[c1]} + \sigma_{\epsilon 1} u_i + \varepsilon_{i,t}^{[c1]} \\ y_{i,t}^{[c2]}|y_{i,t-1}^{[c1]}, y_{i,t-1}^{[c2]}, u_i &= \alpha^{[c2]} + \beta^{[c2]} y_{i,t-1}^{[c1]} + \gamma^{[c2]} y_{i,t-1}^{[c2]} + \sigma_{\epsilon 2} u_i + \varepsilon_{i,t}^{[c2]} \end{aligned} \tag{2.1}$$

where $\varepsilon_{i,t}^{[c1]} \sim N(0, \sigma_{\epsilon 1}^2), \varepsilon_{i,t}^{[c2]} \sim N(0, \sigma_{\epsilon 2}^2)$ and $u_i \sim N(0, \sigma_u^2)$. The random variable $u_i$ is introduced into both equations to estimate the correlation between the two outcomes; $\sigma_{\epsilon 1}$ and $\sigma_{\epsilon 2}$ are scaling parameters used to standardize the residuals. The correlation between $y_{i,t}^{[c1]}$ and $y_{i,t}^{[c2]}$ given $y_{i,t-1}^{[c1]}$ and $y_{i,t-1}^{[c2]}$ is $\frac{\sigma_u^2}{1+\sigma_u^2}$ and it is assumed that given $u_i$, $y_{i,t}^{[c1]}$ and $y_{i,t}^{[c2]}$ are independent.

We also assume the first observations for both outcomes only depend on $\alpha^{[c1]}$ and $\alpha^{[c2]}$, respectively. m

$$\begin{aligned} y_{i,1}^{[c1]}|u_i &= \alpha^{[c1]} + \sigma_{\epsilon 1} u_i + \varepsilon_{i,t}^{[c1]} \\ y_{i,1}^{[c2]}|u_i &= \alpha^{[c2]} + \sigma_{\epsilon 2} u_i + \varepsilon_{i,t}^{[c2]} \end{aligned} \tag{2.2}$$

The likelihood based on equation (2.1) and (2.2) can be written as the multiplication of conditional probability of the fixed mean part and marginal probability of the random effect $u_i$.

$$f(y_{i,1}^{[c1]}, ..., y_{i,t}^{[c1]}, y_{i,1}^{[c2]}, ..., y_{i,t}^{[c2]})$$

$$= \prod_{i=1}^{N} \int f(y_{i,1}^{[c1]}|u_i) f(y_{i,1}^{[c2]}|u_i) \prod_{t=2}^{T} f(y_{i,t}^{[c1]}|y_{i,t-1}^{[c2]}, y_{i,t-1}^{[c1]}, u_i) f(y_{i,t}^{[c2]}|y_{i,t-1}^{[c1]}, y_{i,t-1}^{[c2]}, u_i) f(u_i) du_i$$

$$= \prod_{i=1}^{N} \int \frac{1}{\sqrt{2\pi\sigma_{\epsilon 1}^2}} exp\left[ -\frac{(y_{i,1}^{[c1]} - \alpha^{[c1]} - \sigma_{\epsilon 1} u_i)^2}{2\sigma_{\epsilon 1}^2} \right] \frac{1}{\sqrt{2\pi\sigma_{\epsilon 2}^2}} exp\left[ -\frac{(y_{i,1}^{[c2]} - \alpha^{[c2]} - \sigma_{\epsilon 2} u_i)^2}{2\sigma_{\epsilon 2}^2} \right] \quad (\star)$$

$$\prod_{t=2}^{T} \frac{1}{\sqrt{2\pi\sigma_{\epsilon 1}^2}} exp\left[ -\frac{(y_{i,t}^{[c1]} - \alpha^{[c1]} - \beta^{[c1]} y_{i,t-1}^{[c2]} - \gamma^{[c1]} y_{i,t-1}^{[c1]} - \sigma_{\epsilon 1} u_i)^2}{2\sigma_{\epsilon 1}^2} \right] \quad (\star\star)$$

$$\frac{1}{\sqrt{2\pi\sigma_{\epsilon 2}^2}} exp\left[ -\frac{(y_{i,t}^{[c2]} - \alpha^{[c2]} - \beta^{[c2]} y_{i,t-1}^{[c1]} - \gamma^{[c2]} y_{i,t-1}^{[c2]} - \sigma_{\epsilon 2} u_i)^2}{2\sigma_{\epsilon 2}^2} \right] \quad (\star\star\star)$$

$$\frac{1}{\sqrt{2\pi\sigma_u^2}} exp\left[ -\frac{u_i^2}{2\sigma_u^2} \right] du_i \quad (\star\star\star\star)$$

In the likelihood formula, for the parts after the integral symbol and before the derivative symbol, $(\star)$ indicates the multiplication of probability of the first observations of $y_i^{[c1]}$ and $y_i^{[c2]}$ since they have different mean model from other observations of the same subject; $(\star\star)$

and ($\star\star\star$) indicates the multiplication of probability of all the observations except the first ones of $y_i^{[c1]}$ and $y_i^{[c2]}$; ($\star\star\star\star$) is the marginal probability of $u_i$.

In general, there is no simple closed-form solution for the joint likelihood and numerical integration techniques are required. We use Gaussian quadrature to approximate the integral by a weighted sum, where the quadrature points and weights are chosen to provide a good numerical approximation. Maximization of the likelihood is done in an iterative way.

### 2.3.2 One continuous outcome and one binary outcome

To extend the case of two continuous outcomes in Section 2.3.1 to the case of two non-commensurate outcomes, we let $y_{i,t}^{[c]}$ and $y_{i,t}^{*[b]}$ be two normally distributed correlated continuous outcomes. Here, $y_{i,t}^{*[b]}$ is an underlying variable which is used to define the binary variable $y_{i,t}^{[b]}$. If $y_{i,t}^{*[b]} > 0$ , then $y_{i,t}^{[b]} = 1$; otherwise $y_{i,t}^{[b]} = 0$. The structural regression models of $y_{i,t}^{[c]}$ and $y_{i,t}^{*[b]}$ is

$$\begin{aligned} y_{i,t}^{[c]}|y_{i,t-1}^{[b]}, y_{i,t-1}^{[c]}, u_i &= \alpha^{[c]} + \beta^{[c]} y_{i,t-1}^{[b]} + \gamma^{[c]} y_{i,t-1}^{[c]} + \sigma_{\epsilon 1} u_i + \varepsilon_{i,t}^{[c]} \\ y_{i,t}^{*[b]}|y_{i,t-1}^{[c]}, y_{i,t-1}^{[b]}, u_i &= \alpha^{[b]} + \beta^{[b]} y_{i,t-1}^{[c]} + \gamma^{[b]} y_{i,t-1}^{[b]} + \sigma_{\epsilon 2} u_i + \varepsilon_{i,t}^{[b]} \end{aligned} \tag{2.3}$$

where $\varepsilon_{i,t}^{[c]} \sim N(0, \sigma_{\epsilon 1}^2), \varepsilon_{i,t}^{[b]} \sim N(0, \sigma_{\epsilon 2}^2)$ and $u_i \sim N(0, \sigma_u^2)$.

Based on equation (2.3), we define $y_{i,t}^{*[c]} = \frac{y_{i,t}^{[c]}}{\sigma_{\epsilon 1}}$ and $y_{i,t}^{**[b]} = \frac{y_{i,t}^{*[b]}}{\sigma_{\epsilon 2}}$, where $\sigma_{\epsilon 1}$ and $\sigma_{\epsilon 2}$ are scaling parameters such that

$$\begin{aligned} y_{i,t}^{*[c]}|y_{i,t-1}^{[b]}, y_{i,t-1}^{[c]}, u_i &= \alpha^{*[c]} + \beta^{*[c]} y_{i,t-1}^{[b]} + \gamma^{*[c]} y_{i,t-1}^{[c]} + u_i + \varepsilon_{i,t}^{*[c]} \\ y_{i,t}^{**[b]}|y_{i,t-1}^{[c]}, y_{i,t-1}^{[b]}, u_i &= \alpha^{*[b]} + \beta^{*[b]} y_{i,t-1}^{[c]} + \gamma^{*[b]} y_{i,t-1}^{[b]} + u_i + \varepsilon_{i,t}^{*[b]} \end{aligned} \tag{2.4}$$

where $\varepsilon_{i,t}^{*[c]} \sim N(0,1), \varepsilon_{i,t}^{*[b]} \sim N(0,1)$ and $u_i \sim N(0, \sigma_u^2)$. The random variable $u_i$ again is introduced into both equations to model the correlation between the outcomes. The

correlation between $y_{i,t}^{[c]}$ and $y_{i,t}^{**[b]}$ given $y_{i,t-1}^{[c]}$ and $y_{i,t-1}^{[b]}$ is $\frac{\sigma_u^2}{1+\sigma_u^2}$. It is assumed that given $u_i$, $y_{i,t}^{*[c]}$ and $y_{i,t}^{[b]}$ are independent and therefore $y_{i,t}^{[c]}$ and $y_{i,t}^{[b]}$ are also independent given $u_i$.

We can write the regression equation for the binary outcome in equation (2.4) as $P(y_{i,t}^{[b]} = 1|y_{i,t-1}^{[c]}, y_{i,t-1}^{[b]}) = P(y_{i,t}^{**[b]} > 0|y_{i,t-1}^{[c]}, y_{i,t-1}^{[b]}, u_i) = \Phi(\alpha^{*[b]} + \beta^{*[b]} y_{i,t-1}^{[c]} + \gamma^{*[b]} y_{i,t-1}^{[b]} + u_i)$.

The final model is

$$\begin{aligned} y_{i,t}^{[c]}|y_{i,t-1}^{[b]}, y_{i,t-1}^{[c]}, u_i &= \alpha^{[c]} + \beta^{[c]} y_{i,t-1}^{[b]} + \gamma^{[c]} y_{i,t-1}^{[c]} + \sigma_{\epsilon 1} u_i + \varepsilon_{i,t}^{[c]} \\ Probit[P(y_{i,t}^{[b]} = 1|y_{i,t-1}^{[c]}, y_{i,t-1}^{[b]}, u_i)] &= \alpha^{*[b]} + \beta^{*[b]} y_{i,t-1}^{[c]} + \gamma^{*[b]} y_{i,t-1}^{[b]} + u_i \end{aligned} \tag{2.5}$$

We also assume the first observations for both outcomes only depend on $\alpha^{[c]}$ and $\alpha^{[b]}$ respectively.

$$\begin{aligned} y_{i,1}^{[c]}|u_i &= \alpha^{[c]} + \sigma_{\epsilon 1} u_i + \varepsilon_{i,t}^{[c]} \\ Probit[P(y_{i,1}^{[b]} = 1|u_i)] &= \alpha^{*[b]} + u_i \end{aligned} \tag{2.6}$$

Based on the equation 2.5 and 2.6, the likelihood can be written as

$$f(y_{i,1}^{[c]}, ..., y_{i,t}^{[c]}, y_{i,1}^{[b]}, ..., y_{i,t}^{[b]})$$

$$= \prod_{i=1}^{N} \int f(y_{i,1}^{[c]}|u_i) f(y_{i,1}^{[b]}|u_i) \prod_{t=2}^{T} f(y_{i,t}^{[c]}|y_{i,t-1}^{[b]}, y_{i,t-1}^{[c]}, u_i) f(y_{i,t}^{[b]}|y_{i,t-1}^{[c]}, y_{i,t-1}^{[b]}, u_i) f(u_i) du_i$$

$$= \prod_{i=1}^{N} \int \frac{1}{\sqrt{2\pi\sigma_{\epsilon 1}^2}} exp\left[ -\frac{(y_{i,1}^{[c]} - \alpha^{[c]} - \sigma_{\epsilon 1} u_i)^2}{2\sigma_{\epsilon 1}^2} \right] \Phi(\alpha^{*[b]} + u_i)^{y_{i,1}^{[b]}} [1 - \Phi(\alpha^{*[b]} + u_i)]^{1 - y_{i,1}^{[b]}} \quad (\star)$$

$$\prod_{t=2}^{T} \frac{1}{\sqrt{2\pi\sigma_{\epsilon 1}^2}} exp\left[ -\frac{(y_{i,t}^{[c]} - \alpha^{[c]} - \beta^{[c]} y_{i,t-1}^{[b]} - \gamma^{[c]} y_{i,t-1}^{[c]} - \sigma_{\epsilon 1} u_i)^2}{2\sigma_{\epsilon 1}^2} \right] \quad (\star\star)$$

$$\Phi(\alpha^{*[b]} + \beta^{*[b]} y_{i,t-1}^{[c]} + \gamma^{*[b]} y_{i,t-1}^{[b]} + u_i)^{y_{i,t}^{[b]}} [1 - \Phi(\alpha^{*[b]} + \beta^{*[b]} y_{i,t-1}^{[c]} + \gamma^{*[b]} y_{i,t-1}^{[b]} + u_i))]^{1 - y_{i,t}^{[b]}} \quad (\star\star\star)$$

$$\frac{1}{\sqrt{2\pi\sigma_u^2}} exp\left[ -\frac{u_i^2}{2\sigma_u^2} \right] du_i \quad (\star\star\star\star)$$

In the likelihood formula, for the parts after the integral symbol and before the derivative symbol, $(\star)$ indicates the multiplication of probability of the first observations of $y_i^{[c]}$ and $y_i^{[b]}$; $(\star\star)$ and $(\star\star\star)$ indicates the multiplication of probability of all the observations except the first ones of $y_i^{[c]}$ and $y_i^{[b]}$; $(\star\star\star\star)$ is the marginal probability of $u$.

Similar as Section 2.3.1, there is no simple closed-form solution for the joint likelihood and numerical integration techniques are required. We use Gaussian quadrature to approximate the integral by a weighted sum, where the quadrature points and weights are chosen to provide a good numerical approximation. Maximization of the likelihood is done in an iterative way.

The parameters in the Probit equation of equation (2.5) and equation (2.6) are interpreted conditionally on $u_i$. Given $u_i$, $\beta^{*[b]}$ and $\gamma^{*[b]}$ are the change on the probit of the expected value of $y_t^{[b]}$ for an increase of one unit in the respective covariates. As the result of that, the parameters of the underlying model cannot be directly compared with the regression parameters of the marginal models. To obtain the marginal effects, we have to average over the distribution of $u_i$'s.

$$\int P(y_{i,t}^{[b]}|y_{i,t-1}^{[c]}, y_{i,t-1}^{[b]}, u_i) f(u_i) du_i = \Phi\left(\frac{\alpha^{*[b]} + \beta^{*[b]} y_{i,t-1}^{[c]} + \gamma^{*[b]} y_{i,t-1}^{[b]}}{\sqrt{1+\sigma_u^2}}\right)$$

Thus, $\beta^{[b]} = \frac{\beta^{*[b]}}{\sqrt{1+\sigma_u^2}}$ and $\gamma^{[b]} = \frac{\gamma^{*[b]}}{\sqrt{1+\sigma_u^2}}$ are the marginal effects associated with the covariates. In model comparison, the estimates of $\beta^{[b]}$ and $\gamma^{[b]}$ should be used for binary outcome. Estimates for the marginal effects $\hat{\beta}^{[b]}$ and $\hat{\gamma}^{[b]}$ are obtained using $\frac{\hat{\beta}^{*[b]}}{\sqrt{1+\sigma_u^2}}$ and $\frac{\hat{\gamma}^{*[b]}}{\sqrt{1+\sigma_u^2}}$.

In the following sections, we evaluate our proposed approach through a simulation study in Section 2.4 and apply the model to the IU Phone Study in Section 2.5.

## 2.4 Simulation

A Monte Carlo simulation study is used to investigate the properties of the proposed autoregressive and cross-lagged model. We are especially interested in smaller mean squared error (MSE) gained by modeling the bivariate non-commensurate processes simultaneously. Similar as the Section 2.3, we discuss the simulation procedure of two cases 1) two continuous outcomes and, 2) one continuous outcome and one binary outcome, respectively.

### 2.4.1 Simulation for two continuous outcomes

For the correlated commensurate outcomes case, two continuous outcomes $y_{i,t}^{[c1]}$ and $y_{i,t}^{[c2]}$ are generated with the initial setting

$$y_{i,t}^{[c1]} = -0.5 + 0.4y_{i,t-1}^{[c2]} + 0.6y_{i,t-1}^{[c1]} + \sigma_{\epsilon 1}u_i + \varepsilon_{i,t}^{[c1]}$$

$$y_{i,t}^{[c2]} = -0.5 + 0.6y_{i,t-1}^{[c1]} + 0.3y_{i,t-1}^{[c2]} + \sigma_{\epsilon 2}u_i + \varepsilon_{i,t}^{[c2]}$$

where $u_i \sim N(0, \sigma_u^2)$, $(\varepsilon_{i,t}^{[c1]}, \varepsilon_{i,t}^{[c2]}) \sim MVN((0,0),(\sigma_{\epsilon 1}^2, 0, 0, \sigma_{\epsilon 2}^2))$. We assume $\sigma_{\epsilon 1}^2 = 0.5$ and $\sigma_{\epsilon 2}^2 = 1$. The correlation between two continuous outcomes at time $t$ given two outcomes' observations at time $(t-1)$ is defined as $r = \frac{\sigma_u^2}{1+\sigma_u^2}$.

For the first observations of the two continuous outcomes, we define the generation mechanism as

$$y_{i,1}^{[c1]} = -0.5 + \sigma_{\epsilon 1}u_i + \varepsilon_{i,t}^{[c1]}$$

$$y_{i,1}^{[c2]} = -0.5 + \sigma_{\epsilon 2}u_i + \varepsilon_{i,t}^{[c2]}$$

We are interested in comparing the performance of the proposed method with the univariate GLMs under different data generating assumptions. In the model comparison, we test two values of the correlation: $r = 0.05$ ($\sigma_u^2 = 1/19$) as low correlation and $r = 0.3$

($\sigma_u^2 = 3/7$) as high correlation. For the same correlation assumption, we compare two models under different length of observations per subject (4 time points and 6 time points). For the data set with 6 time points, we test the balanced full data and the data with missing complete at random (MCAR) mechanism. Specifically, in each balanced full data set, we generate 6 sequential time points for both continuous outcomes for all subjects. In data with MCAR, we randomly select subjects from the full data set and restrict their length of time points to 5, 4, 3 and 2 separately. The difference between each adjacent length level is 10 subjects. If we let the $n$ represent the sample size, we end up with $n$ subjects with observations at first two time points; $n - 10$ subjects with observations at first three time points; $n - 20$ subjects with observations at first four time points and so on. For all the scenarios above, we compare models with the sample sizes $n = 100$ and $n = 200$, respectively.

In each data set, we use the proposed method to model the two outcomes simultaneously and used univariate GLMs to model each outcome separately. PROC NLMIXED in SAS version 9.3 is used to execute both models to assure that same numerical algorithms are used in likelihood maximization. Each simulation procedure is repeated 500 times. We compare the percentage of the bias and MSE of estimates from the proposed model and the univariate GLMs based on the 500 simulation results under different data settings.

The percentage of bias and MSE of estimates of two commensurate outcomes (two continuous outcomes) under different correlation (r=0.05 and r=0.3), missing mechanism (full data and MCAR), sample size (N=100 and N=200) and sequence length (p=6 and p=4) for the proposed model (red triangle) and the univariate GLMs (blue circle) are displayed in Figure 2.2. Based on the simulation results, the proposed model provide estimates with smaller percentage of bias and smaller MSE comparing to univariate GLMs. Higher correlation between two outcomes increases the bias and MSE of coefficient estimates under

univariate GLMs, especially for the intercepts in the model; but nearly have no influence on proposed model. Data missingness slightly decreases the bias and MSE of intercepts estimates and slightly increases the bias and MSE of autoregressive and cross-lagged estimates for univariate GLMs when there is a higher correlation between two outcomes, but have no influence on the proposed model. Larger sample size decreases the MSE for both models but have very small influence on bias. Shorter length of the observed sequence for each subject decreases the bias and MSE of intercepts estimates and increases the bias and MSE of autoregressive and cross-lagged effect estimates for univariate GLMs, but have no much influence on the proposed model.

### 2.4.2 Simulation for one continuous outcome and one binary outcome

To deal with the one continuous outcome and one binary outcome case, two continuous variables $y_{i,t}^{[c]}$ and $y_{i,t}^{*[b]}$ are generated with the initial setting as below. Here, $y_{i,t}^{*[b]}$ is a latent variable of $y_{i,t}^{[b]}$.

$$y_{i,t}^{[c]} = -0.5 + 0.4y_{i,t-1}^{[b]} + 0.6y_{i,t-1}^{[c]} + \sigma_{\epsilon 1}u_i + \varepsilon_{i,t}^{[c]}$$

$$y_{i,t}^{*[b]} = -0.5 + 0.6y_{i,t-1}^{[c]} + 0.3y_{i,t-1}^{[b]} + \sigma_{\epsilon 2}u_i + \varepsilon_{i,t}^{[b]}$$

where $u_i \sim N(0, \sigma_u^2)$, $(\varepsilon_{i,t}^{[c]}, \varepsilon_{i,t}^{[b]}) \sim MVN((0,0), (\sigma_{\epsilon 1}^2, 0, 0, \sigma_{\epsilon 2}^2))$. We assume $\sigma_{\epsilon 1}^2 = 0.5$ and $\sigma_{\epsilon 2}^2 = 1$.

For the first observations, we define the generation mechanism as

$$y_{i,1}^{[c]} = -0.5 + \sigma_{\epsilon 1}u_i + \varepsilon_{i,t}^{[c]}$$

$$y_{i,1}^{*[b]} = -0.5 + \sigma_{\epsilon 2}u_i + \varepsilon_{i,t}^{[b]}$$

In each simulation, we keep the $y_{i,t}^{[c]}$ as the continuous outcome and then create the binary outcome $y^{[b]}$ based on the rule that if the $P(Z < y_{i,t}^{*[b]}) > 0.5$ , then $y_{i,t}^{[b]} = 1$; otherwise $y_{i,t}^{[b]} = 0$, where $Z$ is the quantile of the standard normal distribution. The first observation of the binary outcome is generated by following the same transformation procedure. The correlation between the continuous outcome and the underlying continuous outcome of the binary outcome at time $t$ given two outcomes' observations at time $(t-1)$ equals to $\frac{\sigma_u^2}{1+\sigma_u^2}$.

We follow the same model performance comparison procedure in Section 2.4.1. The percentage of bias and MSE of estimates of two non-commensurate outcomes (one continuous outcome and one binary outcome) for the proposed model and the univariate GLMs are presented in Figure 2.3. The proposed model provides estimates with smaller percentage of bias and smaller MSE comparing to univariate GLMs. The estimates of intercepts from two models of one continuous and one binary outcome are closer than the intercept estimates of two models of two continuous outcomes. Higher correlation between two outcomes increases the bias and MSE of coefficient estimates under univariate GLMs, but nearly no influence on proposed model. Data missingness slightly decreases the bias of intercepts estimates and slightly increases the bias and MSE of the autoregressive and cross-lagged estimates for univariate GLMs when there is a higher correlation between two outcomes, but have no influence on the proposed model. Larger sample size decreases the MSE for both models but have very small influence on bias. Shorter length of observed sequence for each subject decreases the bias and MSE of intercepts estimates and increases the bias and MSE of autoregressive and cross-lagged effect estimates for univariate GLMs, but have no much influence on the proposed model.

Based on the summary of the simulation results, the correlation between two outcomes becomes the main factor that distinguishes the performance of the proposed model and the univariate GLMs. The proposed model provides consistent estimates with smaller MSE under different assumptions of the data structure.

## 2.5 Application to the IU Phone Study

In the motivating data example, we are interested in studying the relationship between condom use behavior and sexual satisfaction, which are reported by participants in each 8-hour reporting period. We model the association between the evaluation of the sexual satisfaction and condom use during the next coital event via cross-lagged effects in our model. In an analogous fashion, the association between the current condom use and the sexual satisfaction evaluation to the current event is captured via a different cross-lagged effect. In addition, we model the within-outcome associations via autoregressive effects.

In this study, we use coital event order as the time scale, since participants might not report sexual events during every reporting period. This assumption also allows us to directly utilize the proposed model, since the event-scale interval is consistent within and across the study participants. Though the information on condom use and sexual satisfaction are collected at the same time, the sexual satisfaction reflects the evaluation of the coital event that happened before the reporting time, but after the condom use. As a result, the sexual satisfaction evaluation has in practice a minor lag from the coital event reported in the same period. However, the event intervals between the adjacent events are the same for sexual satisfaction and condom use, since they are corresponding to the same coital event. The path graph of the relationship between these two outcomes of interest is illustrated in Figure 2.4. One cross-lagged regression coefficient ($\beta^{[c]}$) estimates the association between the condom use at event $t$ and sexual satisfaction at event $t$. An

analogous cross-lagged regression coefficient ($\beta^{[b]}$) estimates the association between the sexual satisfaction at event $t$ and condom use probability at event $(t+1)$. An autoregressive coefficient ($\gamma^{[b]}$) estimates the association between the condom use at event $t$ and condom use at event $(t+1)$. A similar autoregressive coefficient ($\gamma^{[c]}$) estimates the association between sexual satisfaction at event $t$ and sexual satisfaction at event $(t+1)$. From the scientific point of view, the cross-lagged associations are of primary interest for understanding reciprocal influences between condom use behaviors and sexual satisfaction over time.

All 254 participants included in our analysis have at least two coital events reported during the study period. There are very few subjects who have more than 20 coital reported. Thus, we restrict our analysis to the first 20 participant-specific events. We analyze the data using both the proposed model and the univariate GLMs. P-values are provided to show the significance of the estimates. The results of proposed model show that the cross-lagged effect ($\beta^{[c]} = -0.77$, $p< 0.0001$) of condom use on sexual satisfaction and the cross-lagged effect ($\beta^{[b]} = -0.14$, $p< 0.0001$) of sexual satisfaction on condom use are both negative and significant, which indicate the use of the condom is negatively associated the high sexual satisfaction of the same coital event and higher sexual satisfaction is negatively associated to the probability of condom use of next coital event. Both of the autoregressive effect of sexual satisfaction ($\gamma^{[c]} = 0.09$, $p< 0.0001$) and condom use ($\gamma^{[b]} = 2.07$, $p< 0.0001$) are positive and significant. The evaluation of sexual satisfaction and the behavior of condom use of next coital event have positive association with the sexual evaluation and condom use of previous coital event respectively. The estimates from the univariate GLMs are similar to the ones obtained using our method. However, the univariate GLMs have smaller absolute values of cross-lagged effect estimates ($\beta^{[c]}_{GLM} = -0.46$, $p< 0.0001$ and $\beta^{[b]}_{GLM} = -0.10$, $p< 0.0001$) than the proposed model, while larger absolute values of autoregressive effect estimates ($\gamma^{[c]}_{GLM} = 0.16$, $p< 0.0001$ and $\gamma^{[b]}_{GLM} = 2.11$, $p< 0.0001$) than the proposed model.

## 2.6 Discussion

The proposed model provides joint estimates of the relationship between two non- commensurate longitudinal outcomes(one continuous and one binary), which extends the commonly used autoregressive and cross-lagged models. By introducing a common subject-specific random effect to estimate the correlation between two correlated outcomes, we combine the univariate mixed model methodology with the cross-lagged models to model correlated bivariate longitudinal outcomes, which relaxes the independent error assumption in the univariate GLMs. Traditional panel models treat all autoregressive and cross-lagged effects as fixed without considering the variation among subjects. Inclusion of the subject-specific random effects in the proposed model accounts for between-subject variability arising from the omitted subject-level predictors. We include both cross-lagged and autoregressive effects in the model in order to minimize bias in the estimation of cross-lagged effects, which is supported by the compelling arguments made by Gollob and Reichardt (1987) and Cole and Maxwell (2003). The estimates obtained from the proposed model in the simulation studies are consistent and have smaller variability than the estimated obtained from the ordinary GLMs. Current model is developed for one continuous outcome and one binary outcome, however, this likelihood-based approach can be applied to outcomes with different measurement types. The proposed model also does not require the data to have a balanced structure and can be used when subjects contribute varying numbers of events. However, we require the inter-observation time intervals to be the same.

In the real data application, we employ the proposed model to the EMA longitudinal dataset. We are able to depict the timing and sequencing link of condom use and sexual satisfaction. In our analysis, we assume that the sexual satisfaction follows a normal distribution to simplify the illustration. In the further studies, Beta distribution might be a better assumption because the empirical distribution of the sexual satisfaction is skewed in

our data. In the IU Phone Study example, we use coital event order instead of chronological time sequence to define the "time lag". However, the chronological time between every two adjacent events might be different from event to event. This choice of time lag might influence the estimates. Till now, all the autoregressive and cross-lagged models required the outcomes share the same time lag across subjects. This assumption might not be hold in the practical life. In the future model development, we plan to combine the variable-domain functional regression method established by Gellar and et al. (2014) with the proposed model to deal with the sparse outcomes with different inter-event time intervals.

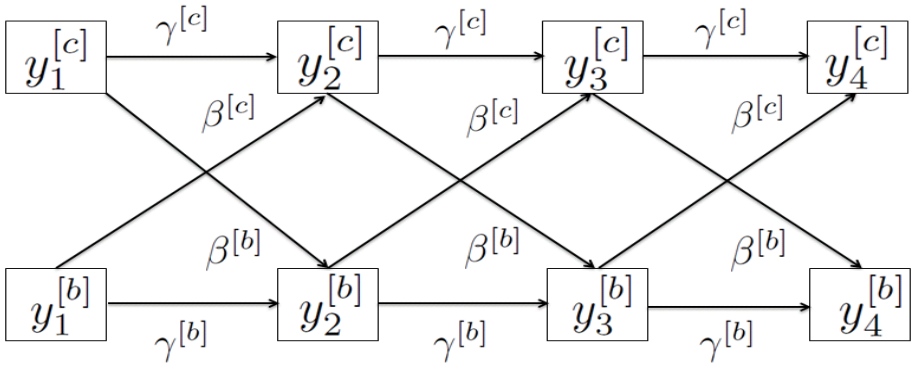


Figure 2.1: Cross-lagged autoregressive model path graph for two longitudinal outcomes (one continuous outcome and one binary outcome)

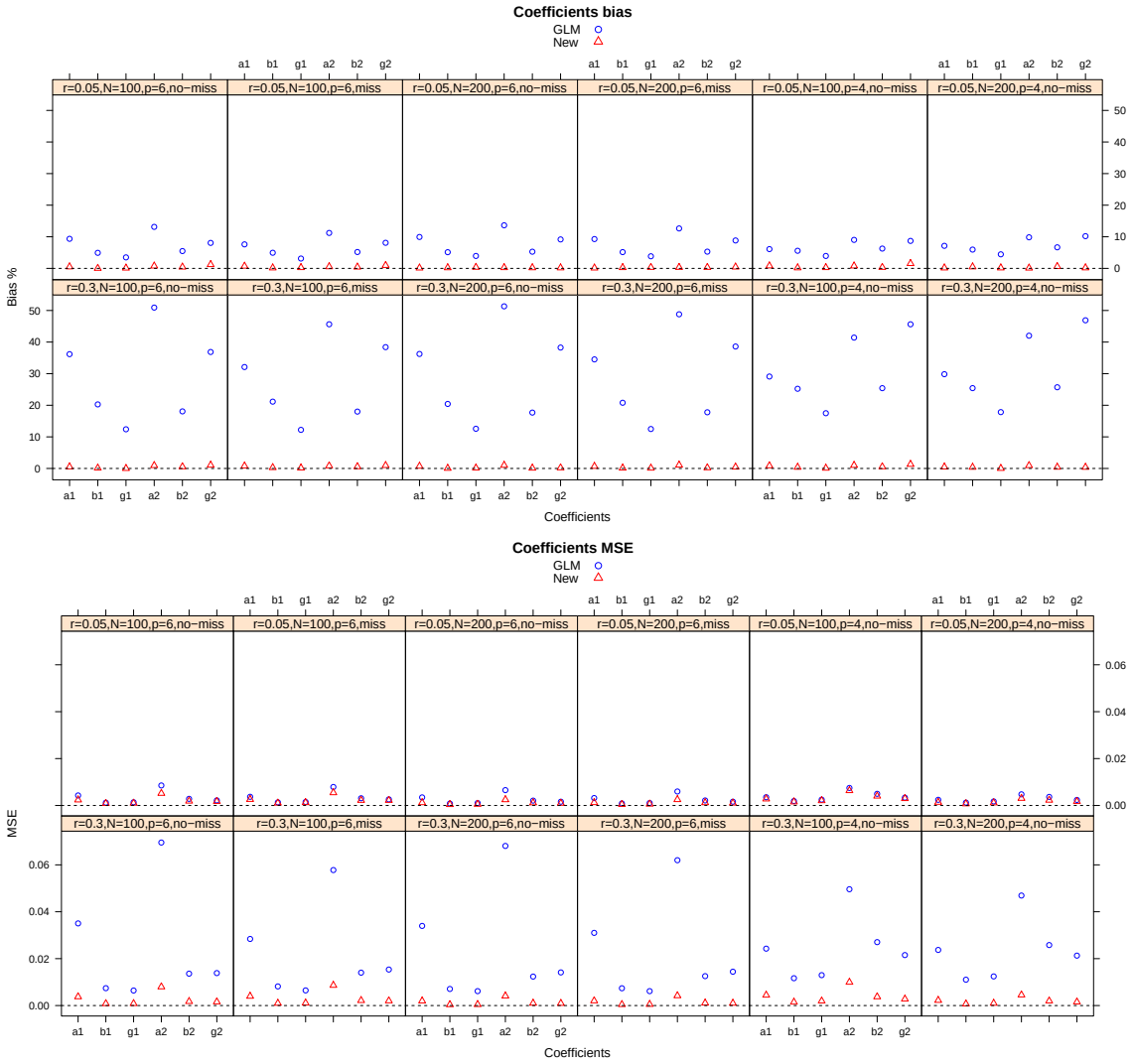


Figure 2.2: Summary of the percentage of the bias and MSE of the 500 runs of estimation between the proposed model and univariate GLMs for two continuous outcomes

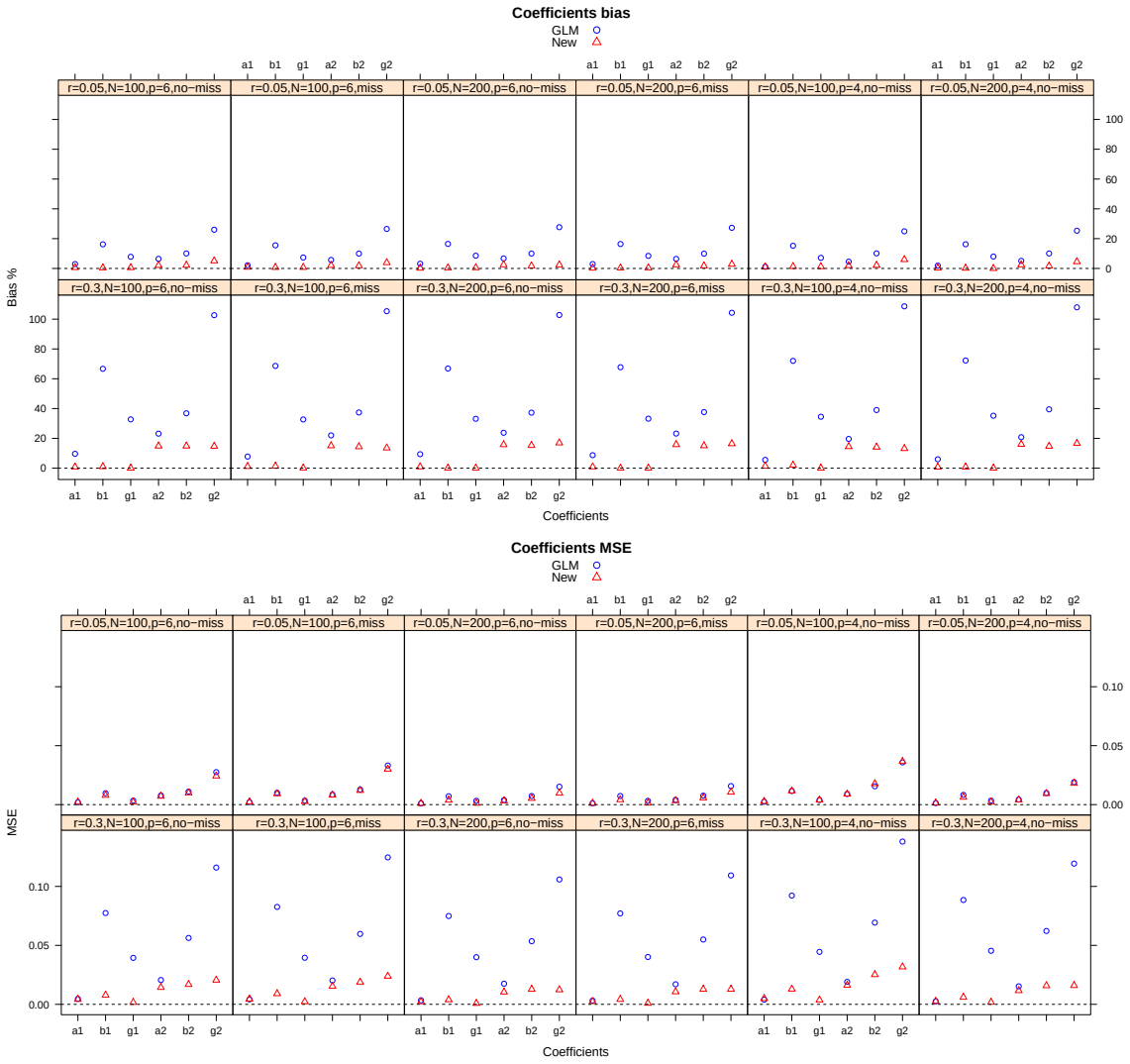


Figure 2.3: Summary of the percentage of the bias and MSE of the 500 runs of estimation between the proposed model and univariate GLMs for one continuous outcome and one binary outcome

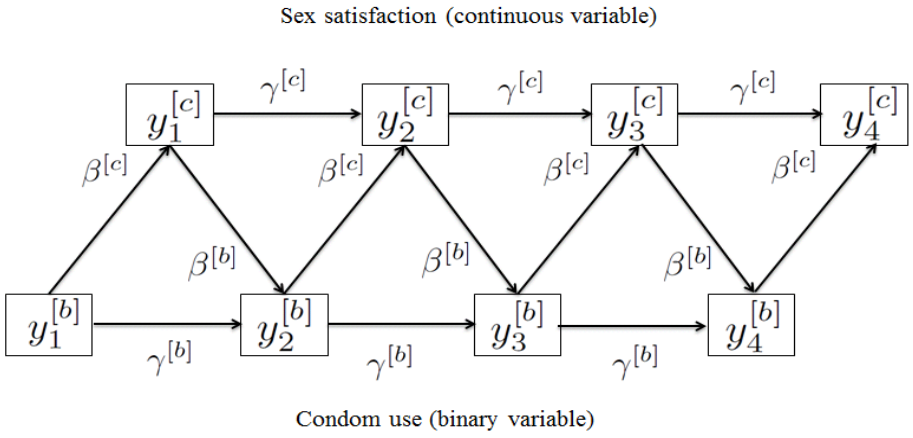


Figure 2.4: Autoregressive and cross-lagged model path graph for condom use and sexual satisfaction

Chapter 3

# An application of variable-domain functional regression models to ecological momentary assessment diary data of a sexually transmitted infections study

## 3.1 Introduction

Partner change plays an important role within sexual networks in terms of linking with the risk of sexually transmitted infections (STIs). The longitudinal associations among relationship factors, partner change, and STIs acquisition is studied by Ott et al. (2011). They find all relationship characteristics, such as lower relationship quality, shorter relationships, and less closeness between participant's partner and family or friends, are significantly and negatively associated with changing partners for adolescent women. Their findings is based on a study design that takes one measurement of relationship characteristics at one time point and partner change status at the following time point after a fixed time interval, which is not a repeated and dynamic measurement that close to the natural environment. In order to study instant processes that influence behavior in real-world circumstances, ecological momentary assessment (EMA) method is introduced into our STIs longitudinal study (IU Phone Study). Data of STIs test results as well as EMA diary data related to sexual and non-sexual behaviors are collected to study the risk factors related to STIs. Among all the variables information, we focus on the association between partner change status and sexual satisfaction and condom use. The partner change status was defined at the end of each partnership and the other two variables were repeated collection of real-time data during the same partnership. Since we are observing participants in non-laboratory conditions, the length of each partnership is not restricted. Therefore, participants could terminate their current partnership and change to a new partner within the study period. As the results of

that, each partnership is recorded with a different length of time. We would like to take the varying length of partnership into consideration when we study the association of interest. Length of partnership using days is one way to measure the stability and intimacy in a relationship, another perspective of understanding the sexual intimacy between a couple is the frequency of sexual events that happen during the partnership. In the IU Phone Study, our interest is the connection between STIs and related risk factors. Because of that, sexual related activities become the main focus in this paper. Instead of using chronological time in a partnership, we are using the order of coital events with the same partner as timeline to track the behaviors change during each partnership.

In functional data analysis (FDA), researchers study the relationship between a scalar response and a functional predictor with different length across subjects in variable-domain functional data. The common domain variable is time and each subject is followed for a different length of time. There are two traditional approaches to analyze variable-domain functional data. One consists of collapsing the trajectory of values into a summary statistic that can be used in a regression model. Common statistics include the mean, median, or maximum value, or the sum of available data. These approaches ignore the functional nature of the data by throwing away much of the available information. Additionally, the choice of summary statistic is often arbitrary, and not driven by the data. The other approach to model variable-domain functional data is to register each function to a common domain,and then apply existing functional regression techniques. This method might be less appropriate for data with large between-subject variability in the width of domain or when the original time domain is informative. Gellar and et al. (2014) developed a class of scalar-on-function regression models with subject-specific functional predictor domains. They consider a bivariate functional parameter that depends both on the functional argument and on the width of the functional predictor domain. Both parametric and nonparametric

models can be used to fit the functional coefficient. Their models have been extended to the case with lagged time, domain standardization and parametric interaction with only one predictor involved. Though those variable domain functional regression (VDFR) models are introduced to the functional data analysis first, we find a strong connection between longitudinal data and functional data in terms of estimating the association between a trajectory against time and an outcome of a status. Because of that, we plan to apply and extend the lagged time model in VDFR to our study which had one scalar outcome (partner change) and two functional predictors (sexual satisfaction and condom use).

The remainder of this article is organized as follows. In the Section 3.2, we describe our data in more detail and also include the data visualization to present the data structure. Section 3.3 introduces the lagged time VDFR model, which is used as the suitable VDFR based on the feature of our data. The results and scientific findings of the data application are displayed in Section 3.4. We conclude with a discussion of the advantage and cautions of using VDFR on EMA data in Section 3.5.

### 3.2 Motivating data set: the IU Phone Study

Data were obtained from a prospective 84-day (12-week) study which was designed to examine sexual behaviors and incident STI. Participants were recruited from the patient population of a county sexually transmitted diseases clinic but were not necessarily clinic patients at the time of enrollment. Eligibility criteria were ages 18 to 29 years (inclusive), English speaking, and planning to reside in the area for the subsequent 84 days. The Institutional Review Board of Indiana University Purdue University Indianapolis approved this study. All participants provided informed consent.

The primary mode of data collection was via scheduled three times a day self-reports of sexual and non-sexual interactions with specific partners and relationship assessments,

which was recorded with project-furnished cellular telephones and service. For example, at each pre-selected eight-hour interval, participants responded to a series of questions to identify partner name, relationship satisfaction, sexual satisfaction, time of each coital event if there was any (up to four events within the same eight-hour reporting period), as well as condom use for each coital event. The expected number of entries was thus 252 entries per participant. Our previous summary showed the daily diary completion rate of IU Phone Study is 87.70%. Other methodological details were published in the paper by Hensel et al. (2012).

In order to illustrate the data structure of IU Phone Study, we show the whole trajectories of variables of interest (partner, relationship satisfaction, sexual satisfaction and condom use for coital events) during the observational time for two subjects (ID=208 and ID=95) in Figure 3.1. The subject with ID=208 had one partner change while the subject with ID=95 had no partner change. The trend of relationship satisfaction and sexual satisfaction with the first partner of subject with ID=208 started at pretty hight level, then declined after 12 days in the study. After three weeks in the study, the subject with ID=208 switched to a new parter with consistently high satisfactions. In terms of the condom use, this subject used condom in majority of the coital events with the first partner but no condom use of the events with the second partner. The subject with ID=95 was considered as monogamous in our study. The trend of relationship satisfaction and sexual satisfaction both kept high in most of the time with few variability at the beginning in the study and when the partnership approached to the end of the study. And there was no condom use in all the coital events with the partner.

Overall, we have 348 participants who have completed the whole study and are not jailed during the study period. We summarize the reporting status of relationship satisfaction, sexual satisfaction and condom use of each subject in the lasagna plots Figure 3.2.

Comparing to relationship satisfaction, both sexual satisfaction and condom use reports are much more sparse because there might be no sexual event happen within each 8-hour reporting period. Though participants are required to report the sexual satisfaction and condom use at every coital event during the study time, some of the condom use reports are missing. Because of that, we have sparser condom use reports comparing to sexual satisfaction reports. The possible reason is detailed information like coital event time and condom use status of each specific event are required to be answered if participants select to report them in our system. Because of that, few participants might skip those questions in order to finish the report more quickly.

With consideration of variable of interest and data sparseness, we exclude those participants who have no coital event reported during the study time and make the final available subject equals to 287. Based on the sensitivity analysis, four participants with total number of coital events larger than 100 are excluded as well because of the very sparse data among higher order of coital events. Eventually, 283 participants who have at least one coital event reported with condom use status at each partnership are included in our analysis. Among those included participants, 57.95% (164/283) are women and 90.81% (257/283) are African American. There are 41.99% (116/283) non-monogamous participants who have more than one partnerships during the study time. For those participants with partner change, we only include their longest partnership in the analysis. Overall, 287 partnerships have been included in the data for analysis. Among those 283 partnerships, 20.14% (57/283) have partner change at the end of the partnership while the rest are terminated by study time window. Due to the missing data of condom use, among all the 283 partnerships, the largest number of reports with sexual satisfaction collected is 86, while the largest number of reports with condom use is 70. The lasagna plots Figure 3.3 summarize the reporting status of sexual satisfaction and condom use according to coital event order of each subject's longest

partnership. From the plots, we can see there is a large variability among the domain of subjects' partnership in terms of total number of coital events and there are some missing data of condom use for some coital events in most of the partnership.

### 3.3 Models and methodology

In the IU Phone Study, the observed data consist of $\{Y_i, Z_i, X_i(t_i) : 0 < t_i < T_i\}$, where $i$ is the index for subject and $t_i$ is the index for coital event order, $t_i = 1, 2, ..., T_i$. Here $T_i$ is the total number of coital events of the longest partnership of subject $i$. In this notation, $X_i^{[s]}(t_i)$ is the sexual satisfaction, recorded at each reported coital event during each partnership, $X_i^{[c]}(t_i)$ is the condom use status, reported at the same coital event, $Z_i$ is nonfunctional covariate gender for subject $i$, and $Y_i$ is the partner change indicator, which reflects the status of partner change at the end of the selected partnership of subject $i$. We assume that $X_i^{[s]}(t_i)$ and $X_i^{[c]}(t_i)$ are sampled from two underlying stochastic processes $\{X_i^{[s]}(t) : t \in T_i\}$ and $\{X_i^{[c]}(t) : t \in T_i\}$ respectively. We can model the association between binary outcome and two functional predictors by

$$logit[P(y_i = 1)] = \alpha + Z_i\gamma + \frac{1}{T_i}\int_0^{T_i} X_i^{[s]}(t)\beta^{[s]}(t, T_i)dt + \frac{1}{T_i}\int_0^{T_i} X_i^{[c]}(t)\beta^{[c]}(t, T_i)dt \quad (3.1)$$

If $X_i(t)$ is assumed that the most recent measurements will have a stronger effect than the earlier ones, then it makes more sense to impose smoothness based on the lagged time. In our case, the partner change status was identified at the end of each partnership and we assume the information related to the coital events that happened more close to the end of the partnership will have stronger association with the partner change status. Instead of modeling the trajectories from the beginning of the partnership, we treat the end of the

partnership as the starting point for smoothing.

Let $k = t - T_i$, $X_i^*(k) = X_i(k + T_i)$ and $\beta^*(k, T_i) = \beta(k + T_i, T_i)$, the model (3.1) becomes

$$logit[P(y_i = 1) = \alpha + Z_i\gamma + \frac{1}{T_i}\int_{-T_i}^{0} X_i^{*[s]}(k)\beta^{*[s]}(k, T_i)dk + \frac{1}{T_i}\int_{-T_i}^{0} X_i^{*[c]}(k)\beta^{*[c]}(k, T_i)dk$$

In the discussion section of Gellar and et al. (2014), they stated that currently the VDFR methods fail to account for missing observations,or sparse or unevenly sampled functional covariates. In the data application of that paper, they chose to use imputation to fill the gaps in their functions. In order to deal with the missing observation problem of condom use in our data set, we choose to fill the gap in the condom use through imputation as well. Specifically, we use the generalized additive model (GAM) to model the condom use as an outcome against the order of coital event to estimate the probability of condom use in each event. After that, we impute those events without condom use reported with the estimated probability from GAM and keep the original condom use information of other events that have condom use reported. The new condom use predictor used in the VDFR models combines the original condom use reports and estimated probability of condom use.

As we stated in Section 3.1, there are two types of models can be used to study how partner change status is associated with sexual satisfaction and condom use. On one side, the generalized linear model (GLM) can be applied to the data including the mean value of sexual satisfaction and condom use percentage of each partnership as scaler predictors and associate them with the partner change status. On the other side, we can use the trajectories of sexual satisfaction and condom use behaviors as functional predictors to predict the possible partner change decision through lagged time VDFR model. Each model is adjusted by gender. In order to compare the performance between GLM and lagged time

VDFR model, we use AUC as criteria. We also calculate the AUC statistics for both models with empirical quantiles based on 2000 bootstrapped samples.

### 3.4 Results

The statistics of total number of coital events in partnership, mean sexual satisfaction and condom use percentage are summarized in Table 3.1. Since the distributions of variables of interest are all skewed, the Mann-Whitney-Wilcoxon test is used to decide whether the population distributions are identical without assuming them to follow the normal distribution. On average, the number of coital events for partnership ended with partner change (median=5 events) is lower than the partnership ended by termination of the study (median=11 events) with p-value equals to 0.0003. The mean sexual satisfaction of group with partner change has median equals to 9, which is not significantly different from the group with no partner change with median equals to 9.33. Similarly, the median of condom use percentage across partnerships in group with partner change and no partner change are 20% and 4.27% respectively. There is no significant condom use percentage difference between partner change group and no partner change group.

We implement the GLM and VDFR model separately on our data set. According to the estimates based on GLM, we find there are no significant effects of mean sexual satisfaction and condom use percentage on partner change. But there is a significant gender difference in partner change behavior. Compared to men, women have lower odds (OR=0.42) of partner change with p-value equals to 0.004.

In the lagged time VDFR model, the plots of the estimated coefficient functions of partner change are presented in Figure 3.2 and Figure 3.4. In Figure 3.2, we show the triangular surface $\hat{\beta}(t, T_i)$ estimated by each predictor as a heat map. In Figure 3.4, we present the univariate weight functions $\hat{\beta}(t, T_0)$ for 10 different values of $T_0$ spread evenly

across the domain of $T_i$ . The top row in these figures displays these estimates, with $T_0$ indicated by color as well as the support along the t-axis, and the bottom row of plots displays the corresponding pointwise Z-scores, $\hat{\beta}(t, T_0)/SE(\hat{\beta}(t, T_0))$. The magnitude of the coefficient function at any particular point $(t, T_i) = (t_0, T_0)$ should only be interpreted conditional on the rest of the curve, the domain width $T_i$ , and the patient population under consideration. Area under curve (AUC) statistics of this lagged time VDFR model is calculated and shown in each plot.

From the left column of Figure 3.4, we see a consistent linear pattern of a positive association between partner change and high sexual satisfaction at the first 15% coital events of one's partnership, but a negative association for the rest of the events in the same partnership, regardless of total number of coital events in a partnership. Within partnerships with different $T_0$, the pointwise associations in the positive regions and most of the negative regions are not statistically significant according to a Wald test with $\alpha = 0.05$. But for the partnerships with total coital events number between 43 and 86 events, the associations are statistically significant in the middle of the negative region. This pattern is showing that subjects start with high sexual satisfaction, but those with more coital events in the partnership and low satisfaction on few events right after half of the total events in their partnership are likely to change their current partner rather than stay in the current partnership. This pattern suggests that the low sexual satisfaction right after half of the events in a partnership with large total number of coital events is very important in predicting partner change behavior. But this association is not held for partnership with small total number of coital events.

On the other side, from the right column of Figure 3.4, there is a consistent nearly linear pattern between partner change and condom use in each partnership, regardless of total number of coital events in the partnership. For partnership with total number of

coital events arranges between 34 and 86 , positive association between partner change and condom use is found during around the first 37% coital events in the partnership while negative association is found in the rest of the partnership. For partnership with number of coital events around 26, the association's change that is from positive to negative starts around the first 13 events. For the partnership with number of coital events around from 8 to 17, the negative region only starts at the last few events. In all cases, the pointwise associations are not statistically significant. But for the partnership with numbers of events less than 17, the pointwise Z-scores of positive regions are close to the significance with level equals to 0.1. This pattern suggests that the condom use in the partnership is not a major factor in predicting partner change behavior.

To compare two models through AUC values, we find VDFR model have in-sample AUC=0.730, which is larger than GLM with AUC=0.627. The AUC statistics for both models with empirical quantiles based on 2000 bootstrapped samples in Table 3.2. The bootstrap results show VDFR model has larger mean of AUC with p-value=0.01 and slightly narrower 95% empirical quantiles intervals compared to GLM.

### 3.5 Discussion

In this paper, we apply the lagged time VDFR model to the IU Phone Study by including both sexual satisfaction and condom use as two functional predictors to estimate the probability of changing a partner. We find significant negative association between partner change and sexual satisfaction trajectory only in the middle of those partnership with coital events number between 43 and 86 but there is no significant association between partner change and condom use trajectory. It is important to recognize that the models that we fit are not causal models, and we do not employ them to try to identify a causal relationship between the covariate function and outcome.

In this data application, VDFR model estimates a weight function to capture the effect of a functional predictor and allows this weight function to vary (smoothly) based on the total follow-up time for each partnership. The advantages of introducing VDFR models here are that we do not ignore the functional nature of the data by throwing away much of the available information like the way in GLM, and it is specifically designed for data with large between-subject variability in the width of domain or when the original time domain is informative. The VDFR models are able to identify features of the association between longitudinally collected covariates and an outcome that traditional multivariate regression methods are not equipped to handle. Sometimes, the functional models might have more complex structure but lower AUC than some of the simpler, parametric models. However, these estimates from functional models may still be revealing and should not be ignored, as they are able to estimate types of associations that are not possible to be estimated by traditional approaches, and still may identify important trends in the data.

From the IU Phone data, we find an interesting connection between EMA data and functional data. Though those VDFR models are developed from the FDA first, our application shows it is appropriate and convenient to use VDFR models in EMA data. We also extend VDFR application from one functional predictor to two functional predictors in order to use more information and combined them together to make a better prediction. In our analysis, we use imputation to deal with missingness of the condom use in IU Phone Study. The systematic sparseness of sexual satisfaction information when comparing to relationship satisfaction is something new in VDFR models. In the future, we are interested in extending the VDFR method to the case with sparse or unevenly sampled functional covariates.

Table 3.1: Summary statistics regarding the distribution of number of coital events, within-partnership mean sexual satisfaction and condom use percentage in the IU Phone data

| | Partner change | | | |
|---|---|---|---|---|
| | All (N=283) | No change (N=226) | Change (N=57) | p-value (Two-sided) |
| Number of events: | | | | |
| Mean (SD) | 14.38 (14.61) | 15.94 (15.53) | 8.21 (7.54) | |
| Median (IQR) | 11 (4, 20) | 11 (4, 22.75) | 5 (3, 45) | $< 0.001$ |
| Range | (1, 86) | (1, 86) | (1, 37) | |
| Average sexual satisfaction: | | | | |
| Mean (SD) | 8.82 (1.39) | 8.87 (1.35) | 8.61 (1.55) | |
| Median (IQR) | 9.25 (8.37, 9.86) | 9.33 (8.50, 9.85) | 9.00 (7.80, 10) | 0.576 |
| Range | (4, 10) | (4, 10) | (4, 10) | |
| Condom use percentage %: | | | | |
| Mean (SD) | 30.75 (0.39) | 28.95 (0.39) | 37.89 (0.42) | |
| Median (IQR) | 5.57 (0, 64.58) | 4.27 (0, 53.91) | 20.00 (0, 85.71) | 0.263 |
| Range | (0, 100) | (0, 100) | (0, 100) | |

Table 3.2: Summary of AUC statistics from lagged time VDFR model and GLM based on 2000 bootstrapped dataset from the IU Phone Study.

| | Lagged time VDFR | GLM |
|---|---|---|
| In-sample | 0.730 | 0.627 |
| Bootstrap: | | |
| Mean | 0.755 | 0.641 |
| ESE | 0.037 | 0.041 |
| 2.5% | 0.686 | 0.563 |
| 5% | 0.694 | 0.575 |
| 10% | 0.709 | 0.590 |
| 90% | 0.801 | 0.693 |
| 95% | 0.818 | 0.708 |
| 97.5% | 0.830 | 0.720 |

[2]Both models are adjusted by gender

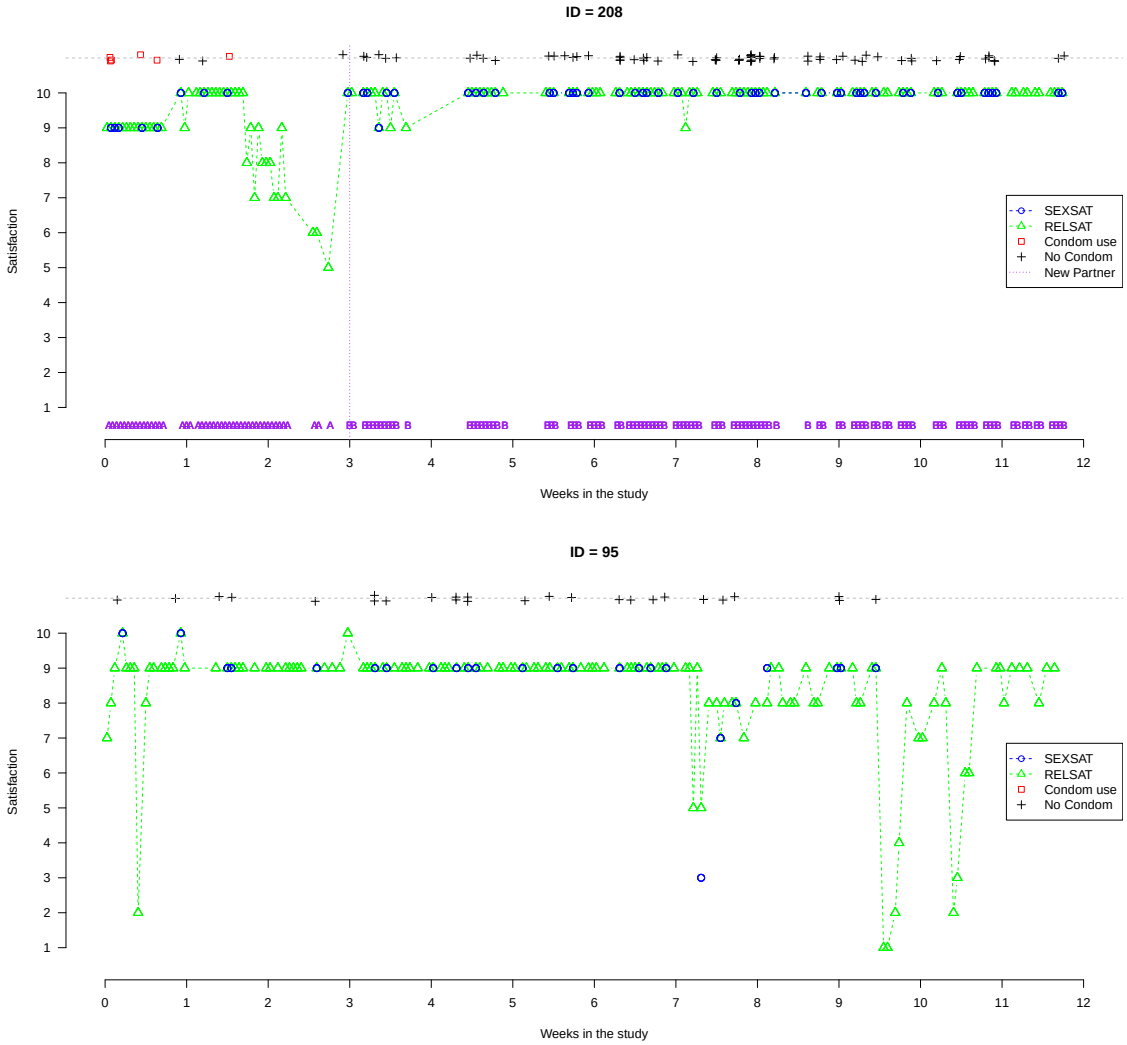


Figure 3.1: Data visualization illustration of variables of interest for one subject with partner change(top) and one subject with no partner change (bottom) during the observational time. The purple letters at the bottom indicates the different partners and the vertical dash line points out the time of changing the partner. The green triangle indicates relationship satisfaction and the blue circle indicates the sexual satisfaction. Both of them are scaled from 1 to 10. The red square shows the vaginal event with condom use, while the black cross shows the vaginal event without condom use.

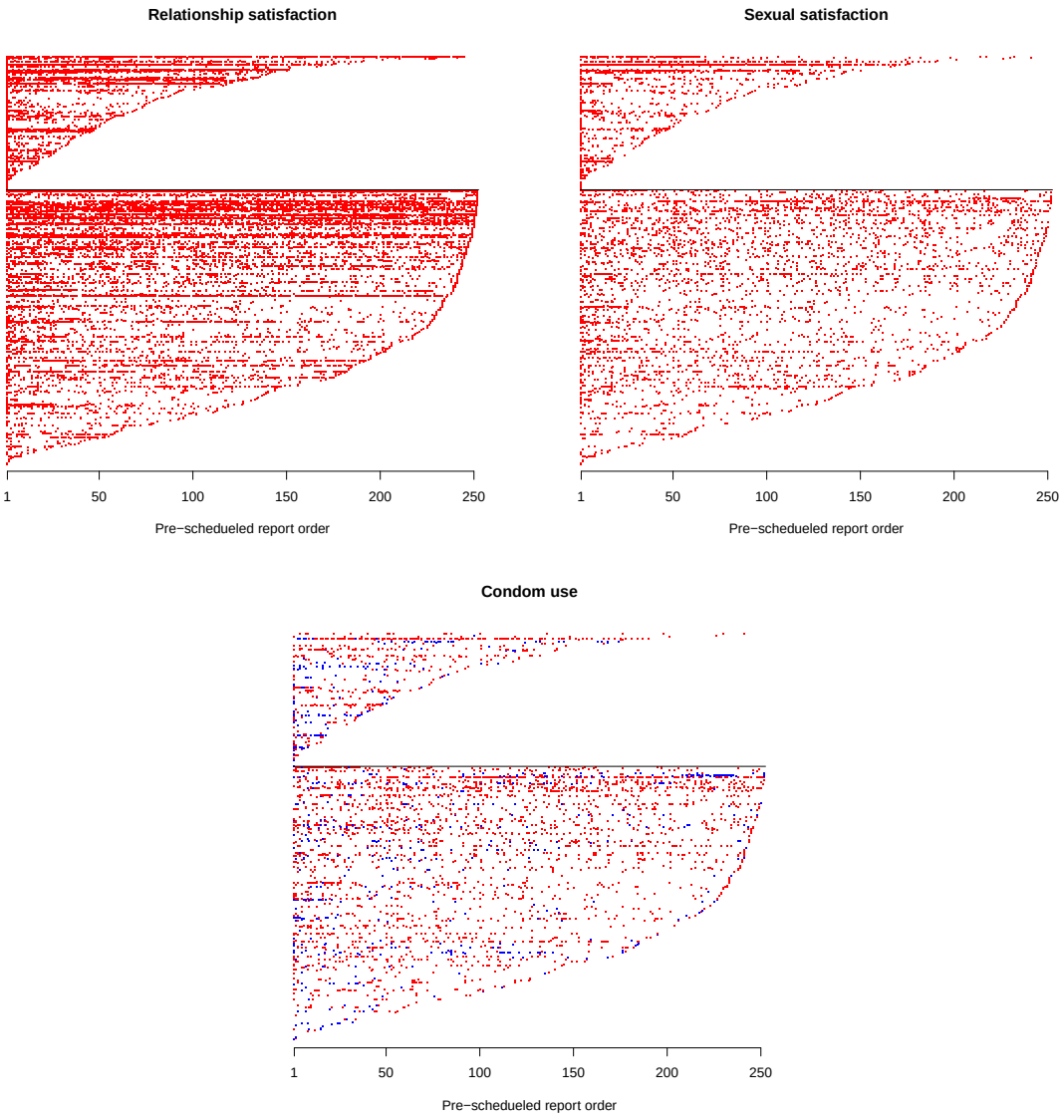


Figure 3.2: Lasagna plots of reporting status of relationship satisfaction, sexual satisfaction and condom use for 348 participants' longest partnership. Rows are correspond to individual subjects. Subjects are sorted according to the length of partnership (the time between first report and last report of each partnership). Colors in relationship satisfaction and sexual satisfaction plots are indicative of reports with reported information. Red in condom use plot indicates no condom use for that coital event and blue indicates coital events with condom use. White space are indicative of missing reports based on each subject's pre-schedules.

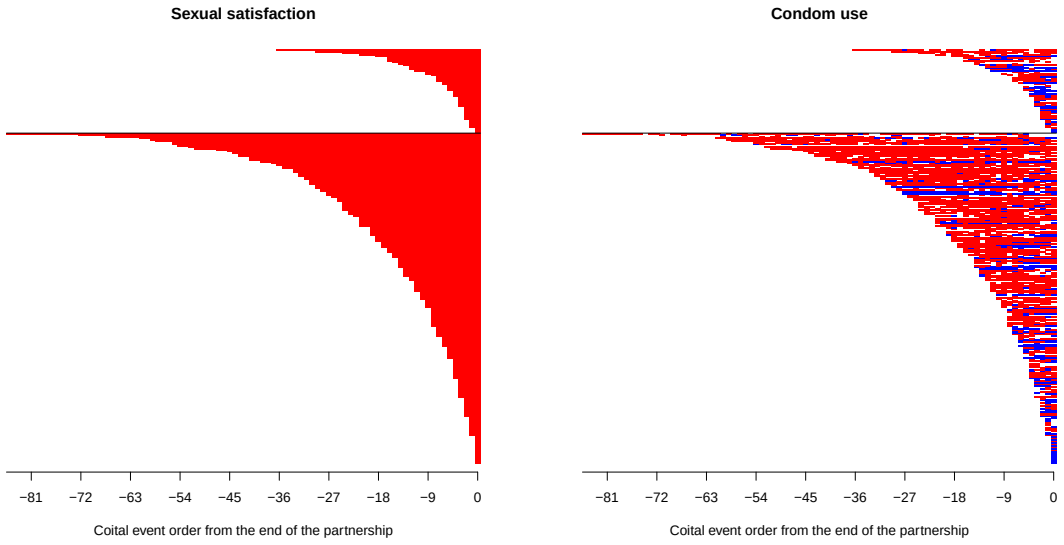


Figure 3.3: Lasagna plots of reporting status of sexual satisfaction and condom use according to coital event order for 283 participants' longest partnership. Rows are correspond to individual subjects. Subjects are sorted according to the total number of coital events in the partnership. Colors in sexual satisfaction plots are indicative of coital events with reported information. Red in condom use plot indicates no condom use for that coital event and blue indicates coital events with condom use. White space are indicative of missing information of each specific event.

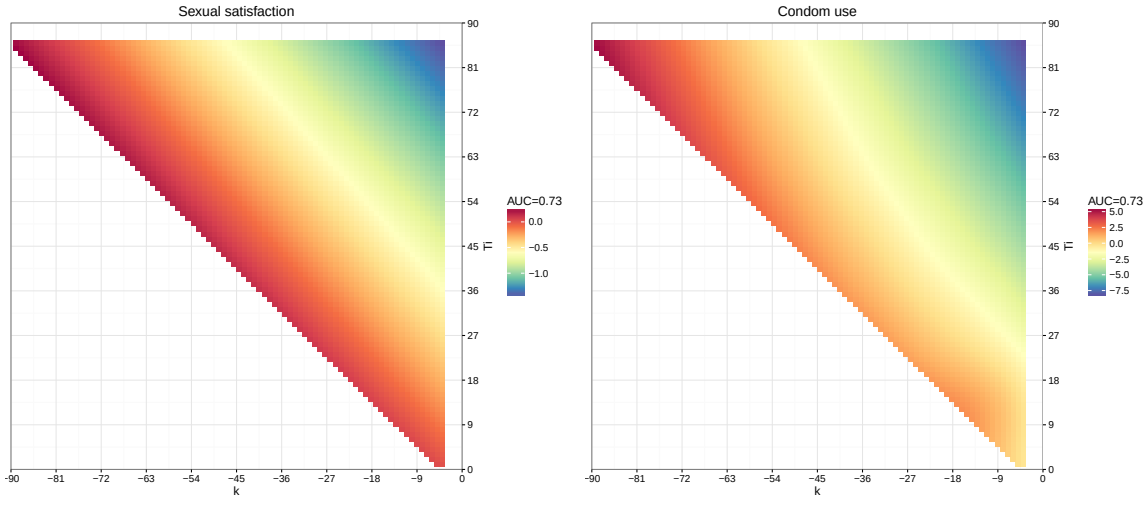


Figure 3.4: Heat maps of the estimated coefficient functions of association between parter change and sexual satisfaction(left) and condom use(right) in the IU Phone dataset. AUC value of the model is included. Here "k" indicates the coital event order from the end of the partnership (partner change or termination of the study), "Ti" indicates the length of the partnership in terms of the total reports number.

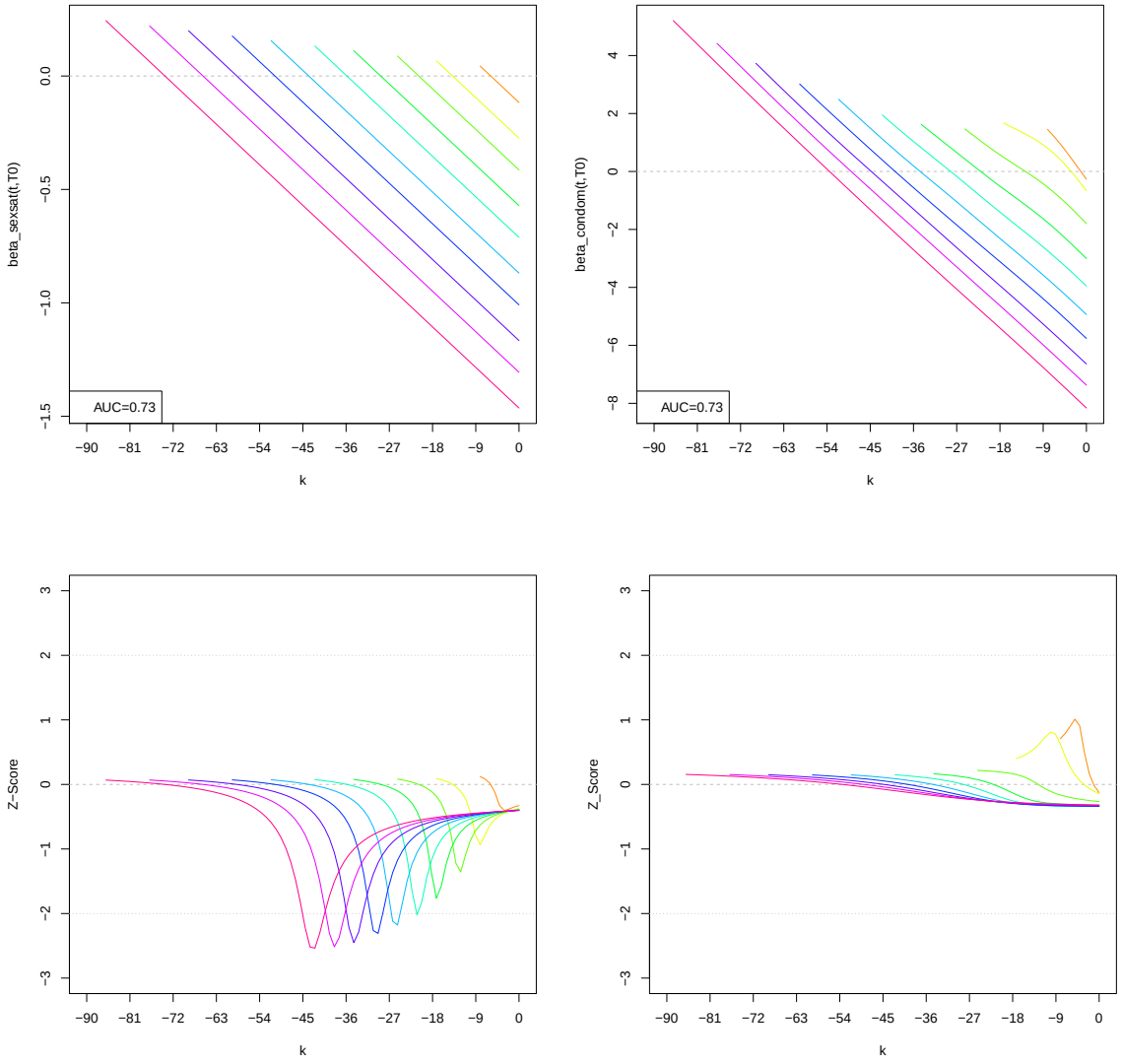


Figure 3.5: Estimated coefficient functions of association between parter change and sexual satisfaction(left column) and condom use(right column) in the IU Phone data set. In the top row of plots, estimates are depicted as $\hat{\beta}(t, T_0)$ for 10 evenly spaced values of $T_0$. AUC statistic is also provided. The bottom row displays the corresponding pointwise Z-scores,$\hat{\beta}(t, T_0)/SE(\hat{\beta}(t, T_0))$, as a function of $t$. The value of $T_0$ is indicated by color and "k" indicates the coital event order from the end of the partnership (partner change or termination of the study). The zero line is indicated with a horizontal dashed line, and dotted lines correspond to Z-scores of $\pm 1.96$.

Chapter 4

# Covariate-specific estimated probability of positivity based on imperfect diagnostic tests

## 4.1 Introduction

Misclassification of either outcomes or covariates has important implications on parameter estimates and statistical inference. In sexually transmitted infections (STIs) studies, imperfect testing tools and possible exposure (rather than real infections) would cause misclassification problem in the test results. In order to get a better estimation of the current true STIs status through previous test results and other covariates, methodology related to imperfect test need to be explored and developed. Bross (1954) is the first to investigate the effect of misclassification on binary outcome with binomial distribution. Later, Newell (1963), Koch (1969), Goldberg (1975) extended Bross's investigation to study misclassified 2 by 2 tables. More generally, the effect of misclassification on the association between exposures and outcomes has been investigated by numerous epidemiologists Copeland and et al. (1977), Lyles and et al. (2011), Magder and Hughes (1997). Edwards and et al. (2013) used multiple imputations with internal validation data to deal with the misclassified binary outcomes. Among variety of methods, maximum likelihood or quasi-likelihood estimation is feasible in many models Carrol et al. (1995). Also a direct correction for misclassification bias is available for simple models by the matrix method Kuha et al. (2005). The Bayesian literature on this topic and related ones are growing as well. For instance, Geng and Asano (1989), Evans and et al. (1996), Mendoza-Blanco et al. (1996), and Rekaya and K.A. Weige and (2001) have developed different approaches to misclassified categorical data under several sampling schemes. Thurigen and et al. (2000) offered a review of methods for measurement errors in the covariates. Paulino et al. (2003) presented a Bayesian binomial

regression analysis in which the response is subject to an unconstrained misclassification process. Among many other methods, the simulation extrapolation method (SIMEX) Cook and Stefanski (1994)Carroll and et al. (1996) has become a useful tool for correcting effect estimates in the presence of additive measurement error. Kuchenhoff et al. (2006) extended the previous work to a general method called MC-SIMEX for dealing with misclassification in regression by SIMEX. Besides those works in cross-sectional data, misclassification in the longitudinal binary response has been treated by Neuhaus (1999),Neuhaus (2002). Misclassification in both response and covariates has been discussed by Chen et al. (2014).

This paper is presenting a likelihood based method to adjust the misclassification of the binary outcome with correctly or incorrectly measured covariates. The proposed method is implemented to both cross-sectional data and longitudinal data under certain assumption. We first introduce the design of a large longitudinal STIs study in Section 4.2 as the data motivation. In Section 4.3, the procedures of estimating the parameters of covariates and probability of outcome using expectation maximization (EM) algorithm are explained under different covariate assumption. The simulation study with different sensitivity and specificity setting are provided to compare the proposed algorithm with naive model and MC-SIMEX for each covariate assumption respectively in Section 4.4. We apply the proposed method to STI test results from the IU Phone Study in Section 4.5. The limitations and future works are discussed in Section 4.6.

## 4.2 Motivating data set: the IU Phone Study

Our data were obtained from a prospective 84-day (12-week) study designed to examine sexual behaviors and incident STI. Participants were recruited from the patient population of a county sexually transmitted diseases clinic but were not necessarily clinic patients at the time of enrollment. Eligibility criteria were ages 18 to 29 years (inclusive), English

speaking, and planning to reside in the area for the subsequent 84 days. The Institutional Review Board of Indiana University Purdue University Indianapolis approved this study. All participants provide informed consent.

Ecological momentary assessment (EMA) diary technology was used in the study. The primary mode of data collection was via 8-hour self-reports of coital and non-coital sexual behaviors, condom use, and relationship assessments, recorded with project-furnished cellular telephones and service. The expected number of entries was thus 252 entries per participant. Our previous summary showed the daily diary completion rate of IU Phone Study is 87.7%. Other methodological details were previously published in the paper by Hensel et al. (2012). At pre-selected 8-hour intervals, participants responded to a series of questions to identify sexual and non-sexual interactions with specific partners. In each eight-hour reporting period, participants identified any partner, time of each coital event (up to four events within the same eight-hour reporting period), condom use for each coital event. Commercially available nucleic acid amplification test (NAAT) was used to test C trachomatis (CT), N gonorrhoeae (GC), and T vaginalis (TV). Participants have received NAAT before the entry of the study. Treatments have been provided to those participants who showed positive result at the enrollment test. Weekly self-obtained vaginal or urine samples are collected since study started for 12 weeks. All the samples were kept in the lab and haven't been tested until the end of 12-weeks in order to avoid possible intervention of natural observation. Participants were allowed to visit the clinics or take medicines if they want to. Including the STI diagnosis at the enrollment, each participant supposed to have 13 lab tests. Participants with any positive results in NAAT at enrollment were excluded from the analysis in order to avoid a potential confounding effect due to treatment.

We illustrate six participants' data information of lab results and condom use behaviors (use condom or not use condom) with one or more partners during the whole study time

in Figure 4.1 to show the data structure. For example, subject 1 changed partner for six times during the study time and only had one coital event without using condom. All the specimens of subject 1 were negative. Subject 3 has changed partner for five times and used the condom more frequently during the time between middle of the second and third week. All the specimens of subject 3 were negative except the specimen taken at the end of the ninth week are shown to be positive. Subject 6 only had one partner during the whole study time with few coital events protected by condom in the second week. The specimens taken after the first week were shown to be positive for 10 weeks. There was one negative test result shown after 11 weeks in the study which was followed by another positive results for the last specimen. Since it is hard to detect whether each test results is reflecting the real infection/exposure status or it is a misclassified result, we plan to build a model combining the known information of sensitivity and specificity of the test with historical test results and other covariates to estimate the probability of the positive STIs.

## 4.3 Models and methodology

Based on the type and assumption of the covariates of interest, different algorithms are used to estimate the parameter and probability of positivity. We first discuss the correctly measured continuous and binary covariates in Section 4.3.1. Following that, the case with misclassified binary variable as covariate is introduced in Section 4.3.2. Autoregressive (AR)(1) in Section 4.3.3 is considered as the extension of Section 4.3.2.

### 4.3.1 $X$ is a correctly measured continuous variable or binary variable.

Let $Y$ denoted the true binary outcome and $Y^{obs}$ is the observed value for the true $Y$ with sensitivity $P(Y^{obs} = 1|Y = 1) = sens_y$ and specificity $P(Y^{obs} = 0|Y = 0) = spec_y$. Along with the test result $Y^{obs}$ we also observe a matrix of covariates $X$. Let $\pi(x;\beta) =$

$P(Y=1|X;\beta)$ denote the probability of true $Y$ conditional on the covariates and $P(Y^{obs}=1|Y,X)=P(Y^{obs}=1|Y)=sens_y^y(1-spec_y)^{1-y}$ denote the probability of $Y^{obs}=1$ that is assumed independent of the covariates given the true $Y$. The true sensitivity and specificity are assumed known constants. Assuming the true $Y$ are i.i.d. conditional on $X$, the likelihood function given the observed data $D_i=(y_i,y_i^{obs},x_i)$, $i=1,...,n$, is

$$L(\beta;D)=\prod_{i=1}^{n}P(Y=y_i|X=x_i)P(Y^{obs}=y_i^{obs}|Y=y_i,X=x_i)$$

$$=\prod_{i=1}^{n}\pi(x_i;\beta)^{y_i}[1-\pi(x_i;\beta)]^{1-y_i}[sens_y^{y_i}(1-spec_y)^{1-y_i}]^{y_i^{obs}}[1-sens_y^{y_i}(1-spec_y)^{1-y_i}]^{1-y_i^{obs}}$$

$$\propto\prod_{i=1}^{n}\pi(x_i;\beta)^{y_i}[1-\pi(x_i;\beta)]^{1-y_i}$$

So the maximum likelihood estimate $\hat{\beta}$ can be obtained by maximizing the loglikelihood:

$$l(\beta;D)=\sum_{i=1}^{n}\{y_i\log[\pi(x_i;\beta)]+(1-y_i)\log[1-\pi(x_i;\beta)]\} \tag{4.1}$$

EM algorithm is used here to get the $\hat{\beta}$ iteratively. Conditional on the observed data $D_i^{obs}=(Y_i^{obs},X_i)$, the algorithm is:

**E-step:** In the $s$-th iteration calculate the expected loglikelihood given $D_i^{obs}$ as

$$Q(\beta;\beta^{(s-1)})=\sum_{i=1}^{n}\Big\{E_{\beta^{(s-1)}}(y_i|D_i^{obs})\log[\pi(x_i;\beta)]+[1-E_{\beta^{(s-1)}}(y_i|D_i^{obs})]\log[1-\pi(x_i;\beta)]\Big\}$$

where by the Bayes theorem

$$E_{\beta^{(s-1)}}(y_i|D_i^{obs})$$

$$=\frac{\pi(x_i;\beta^{(s-1)})sens_y^{y_i^{obs}}(1-sens_y)^{1-y_i^{obs}}}{\pi(x_i;\beta^{(s-1)})sens_y^{y_i^{obs}}(1-sens_y)^{1-y_i^{obs}}+\left[1-\pi(x_i;\beta^{(s-1)})\right](1-spec_y)^{y_i^{obs}}spec_y^{1-y_i^{obs}}} \tag{4.2}$$

**M-step:** Choose $\beta^{(s)}$ such that

$$Q(\beta^{(s)};\beta^{(s-1)}) \geq Q(\beta^{(s-1)};\beta^{(s-1)})$$

The process continuous until

$$l(\beta^{(s)};D^{obs}) - l(\beta^{(s-1)};D^{obs}) < 10^{-8}$$

### 4.3.2 $X$ is a misclassified binary variable.

The use of EM algorithm is straight forward when there is no misclassification in the binary covariate as in Section 4.3.1, but it cannot be directly extended to the case with misclassified binary covariate. When $X$ is a misclassified binary variable with observation as $X^{obs}$, let the $P(X^{obs}=1|X=1) = sens_x$ be sensitivity and $P(X^{obs}=0|X=0) = spec_x$ be specificity, which are also considered as known constant here. The maximum likelihood estimate $\hat{\beta}$ would be obtained by maximizing the equation (4.1) through EM algorithm interactively.

**E-step:** In the $s$-th iteration calculate the expected loglikelihood given $D_i^{obs}$ as

$$Q(\beta;\beta^{(s-1)}) = \sum_{i=1}^{n} E_{\beta^{(s-1)}}(y_i \log[\pi(x_i;\beta)]|D_i^{obs}) + E_{\beta^{(s-1)}}([1-y_i]\log[1-\pi(x_i;\beta)]|D_i^{obs})$$

In order to calculate the expectation of $y_i \log[\pi(x_i;\beta^{(s)})]$, we first derive the joint probability of $Y$ and $X$ as

$$P(y_i,x_i|y_i^{obs},x_i^{obs};\beta^{(s-1)}) = \frac{P(y_i^{obs}|y_i,x_i,x_i^{obs}) * P(y_i|x_i,x_i^{obs};\beta^{(s-1)}) * P(x_i^{obs}|x_i) * P(x_i)}{P(y_i^{obs},x_i^{obs})}$$

$$= \frac{P(y_i^{obs}|y_i) * P(y_i|x_i;\beta^{(s-1)}) * P(x_i^{obs}|x_i) * P(x_i)}{P(y_i^{obs},x_i^{obs})}$$

For the denominator, the $P(y_i^{obs}, x_i^{obs})$ are constant but vary from subject to subject. Based on the property of probability that $\sum_{x=0}^{1}\sum_{y=0}^{1} P(y_i, x_i|y_i^{obs}, x_i^{obs}; \beta^{(s-1)}) = 1$, $P(y_i^{obs}, x_i^{obs})$ can be calculated through normalization method in the simulation procedure. Because $Y$ and $X$ are both binary variables with value 0 and 1, for the four probabilities in multiplication in numerator, we classify each probability into several cases under different combination of $Y$ and $X$'s value.

Based on the sensitivity and specificity of $Y$, $P(y_i^{obs}|y_i)$ is defined as

$$P(y_i^{obs}|y_i = 1) = sens_y^{y_i^{obs}} * (1 - sens_y)^{1-y_i^{obs}}$$

$$P(y_i^{obs}|y_i = 0) = (1 - spec_y)^{y_i^{obs}} * spec_y^{1-y_i^{obs}}$$

According to the definition of $\pi(x; \beta)$, $P(y_i|x_i; \beta^{(s-1)})$ can be written as

$$P(y_i = v|x_i = w) = [\pi(x_i = w; \beta^{(s-1)})]^{y_i} * [1 - \pi(x_i = w; \beta^{(s-1)})]^{1-y_i}$$

where $v$ and $w$ take the value as 0 or 1.

Based on the sensitivity and specificity of $X$, $P(x_i^{obs}|x_i)$ is shown to be

$$P(x_i^{obs}|x_i = 0) = (1 - spec_x)^{x_i^{obs}} * spec_x^{1-x_i^{obs}}$$

$$P(x_i^{obs}|x_i = 1) = {sens_x}^{x_i^{obs}} * (1 - sens_x)^{1-x_i^{obs}}$$

We also estimate $P(x = 1)$ through

$$P(x = 1) = \frac{\frac{\sum_{i=1}^{n} x_i^{obs}}{n} + spec_x - 1}{sens_x + spec_x - 1}$$

Because $P(x=1)$ is a probability between 0 and 1, we let $\frac{\sum_{i=1}^{n} x_i^{obs}}{n} \leq sens_x \leq 1$ and $1 - \frac{\sum_{i=1}^{n} x_i^{obs}}{n} \leq spec_x \leq 1$ to guarantee the property of probability.

Besides the joint probability of $Y$ and $X$, in order to calculate the expectation of $y_i \log[\pi(x_i; \beta^{(s)})]$, we also derive $y_i \log[\pi(x_i; \beta^{(s)})]$ under different value of $Y$ and $X$ as

$$y_i \log[\pi(x_i; \beta^{(s)})]|_{y_i=v, x_i=w} = \{\log[\pi(x_i = w; \beta^{(s)})]\}^{y_i} * \{\log[1 - \pi(x_i = w; \beta^{(s)})]\}^{1-y_i}$$

where $v$ and $w$ take the value as 0 or 1.

The final value of $Q(\beta; \beta^{(s-1)})$ will be the summation of all the cases above.

**M-step:** Choose $\beta^{(s)}$ such that

$$Q(\beta^{(s)}; \beta^{(s-1)}) \geq Q(\beta^{(s-1)}; \beta^{(s-1)})$$

The process continuous until

$$l(\beta^{(s)}; D^{obs}) - l(\beta^{(s-1)}; D^{obs}) < 10^{-8}$$

### 4.3.3 AR(1) model of misclassified binary outcome.

We consider the AR(1) model as the extension of misclassified covariates in 4.3.2 with assumption of true $Y_t$ are i.i.d. conditional on $Y_{t-1}$. The likelihood function in Section 4.3.1 can be rewritten for AR(1) model as

$$L(\beta; D) =$$

$$\prod_{i=1}^{n} P(Y = y_{i,1}) P(Y^{obs} = y_{i,1}^{obs} | Y = y_{i,1}) \prod_{t=2}^{T} P(Y = y_{i,t} | Y = y_{i,t-1}) P(Y^{obs} = y_{i,t}^{obs} | Y = y_{i,t})$$

$$\propto \prod_{i=1}^{n} \pi(y_{i,1}; \beta_0)^{y_{i,1}} [1 - \pi(y_{i,1}; \beta_0)]^{1-y_{i,1}} \prod_{t=2}^{T} \pi(y_{i,t-1}; \beta)^{y_{i,t}} [1 - \pi(y_{i,t-1}; \beta)]^{1-y_{i,t}}$$

So the maximum likelihood estimate $\hat{\beta}$ can be obtained by maximizing the loglikelihood:

$$\begin{aligned} l(\beta; D) = \sum_{i=1}^{n} \{ & y_{i,1} \log[\pi(\beta_0)] + (1 - y_{i,1}) \log[1 - \pi(\beta_0)] \\ & \sum_{t=2}^{T} y_{i,t} \log[\pi(y_{i,t-1}; \beta)] + (1 - y_{i,t}) \log[1 - \pi(y_{i,t-1}; \beta)] \Big\} \end{aligned}$$

The estimating procedure using EM algorithm is similar like the case in Section 4.3.2 with the $Q(\beta; \beta^{(s-1)})$ in E-step as

$$Q(\beta; \beta^{(s-1)}) = \sum_{i=1}^{n} \Big\{ E_{\beta_0^{(s-1)}}(y_{i,1} | D_i^{obs}) \log[\pi(\beta_0)] + [1 - E_{\beta_0^{(s-1)}}(y_{i,1} | D_i^{obs})] \log[1 - \pi(\beta_0)] \quad (\star)$$

$$\sum_{t=2}^{T} E_{\beta^{(s-1)}}(y_{i,t} \log[\pi(y_{i,t-1}; \beta)] | D_i^{obs}) + E_{\beta^{(s-1)}}([1 - y_{i,t}] \log[1 - \pi(y_{i,t-1}; \beta)] | D_i^{obs}) \Big\} \quad (\star\star)$$

Similar like Section 4.3.1, $(\star)$ can be derived as

$$E_{\beta_0^{(s-1)}}(y_{i,1} | D_i^{obs}) =$$

$$\frac{\pi(\beta_0^{(s-1)}) sens_y^{y_{i,1}^{obs}} (1 - sens_y)^{1 - y_{i,1}^{obs}}}{\pi(\beta_0^{(s-1)}) sens_y^{y_{i,1}^{obs}} (1 - sens_y)^{1 - y_{i,1}^{obs}} + \left[1 - \pi(\beta_0^{(s-1)})\right] (1 - spec_y)^{y_{i,1}^{obs}} spec_y^{1 - y_{i,1}^{obs}}}$$

In order to calculate $(\star\star)$ , we have to derive the joint likelihood of $y_{i,t}$ and $y_{i,t-1}$ as

$$P(y_{i,t}, y_{i,t-1}|y_{i,t}^{obs}, y_{i,t-1}^{obs}; \beta^{(s-1)}) =$$

$$\frac{P(y_{i,t}^{obs}|y_{i,t}) * P(y_{i,t}|y_{i,t-1}; \beta^{(s-1)}) * P(y_{i,t-1}^{obs}|y_{i,t-1}) * P(y_{i,t-1})}{P(y_{i,t}^{obs}, y_{i,t-1}^{obs})}$$

We can follow the same procedure in Section 4.3.2 to define the probabilities under different combinations of $y_{i,t}$ and $y_{i,t-1}$'s value of 0 and 1. Because $P(Y_{t-1} = 1)$ is a probability between 0 and 1 and we assume sensitivity and specificity are not changing with time, we let $min_t(\frac{\sum_{i=1}^{n} y_{i,t-1}^{obs}}{n}) \leq sens_y \leq 1$ and $1 - min_t(\frac{\sum_{i=1}^{n} y_{i,t-1}^{obs}}{n}) \leq spec_y \leq 1$ to guarantee the property of probability.

Based on the estimation algorithm described in Section 4.3.1, Section 4.3.2 and Section 4.3.3 , this likelihood based misclassification correction method can be extended to models including more than one covariates with different type and different misclassification setting.

We only derive the method to calculate the mean estimates of the coefficients in our models. In order to get the empirical standard error and confidence interval of estimates when applying to the IU Phone Study, we use bootstrap sampling method. Samples are bootstrapped at the test event level because of the conditional independency assumption in the model. Test events are sampled from the original data set with the same sample size and this bootstrap procedure is repeated for 1000 times. Both proposed model and GLM are implemented on each bootstrapped data set.

## 4.4 Simulation studies

Corresponding to the algorithms described in Section 4.3, we implement simulations to compare the estimated parameters and probability of positivity of our model with estimates from MC-SIMEX models, naive model and true model. Similar to the case discussed

in Section 4.3, we make the comparison under three assumptions of covariates:1) $X$ is a correctly measured continuous variable or binary variable; 2) $X$ is a misclassified binary variable. 3) Use previous $Y$ as predictor in an AR(1) model.

### 4.4.1 $X$ is a correctly measured continuous/binary variable.

We perform 200 simulations with a sample size of 1000 per simulated data. Different types of $X$ are tested, as a correctly measured continuous covariate or a binary covariate with no misclassification, in the models separately. The continuous covariate are randomly drawn from a standard normal distribution. The binary covariate $X$ are generated from a Bernoulli distribution with $P(X = 0) = P(X = 1) = 0.5$. The true value $Y$ of the binary outcome are obtained from a Bernoulli distribution with the probability $P(Y = 1) = \frac{1}{1+e^{-\beta_0-\beta_1 X}}$. We apply the misclassification operation in Section 4.3.1 on $Y$ to obtain the misclassified outcome $Y^{obs}$. All the methods are evaluated under two settings of sensitivity and specificity ($sens_y$ =0.7, $spec_y = 0.9$, and $sens_y = 0.8$, $spec_y = 0.8$).

We compare the estimated $\hat{\beta}_0$ and $\hat{\beta}_1$ from the proposed method, MC-SIMEX models for the three extrapolation functions (linear, quadratic, and log-linear), the naive model, and the true model. The estimates of the true model is obtained by regressing the correctly measured $Y$ on $X$ in generalized linear model(GLM) and the estimates of the naive mode is obtained by regressing the observed $Y^{obs}$ on $X^{obs}$ in GLM. Mean, absolute bias (Abs Bias) and empirical standard error (ESE) of estimates $\beta_0$ and mean, bias percentage (Bias%) and ESE of estimates $\beta_1$ are present in the Table 4.1 under different assumption of covariate $X$ (binary variable or continuous variable) and different settings of sensitivity and specificity of $Y$ respectively.

Overall, our estimates of $\beta_0$ and $\beta_1$ in most of the cases have much smaller bias than the estimates based on naive model and MC-SIMEX models, and are very close to the

true models's estimates. The ESE of estimates from proposed method based on the 200 simulation's estimates are slightly larger than the other models.

The estimates of $\beta_0$ from naive model and MC-SIMEX models have much smaller bias when sensitivity of $Y$ increases from 0.7 to 0.8 and specificity of $Y$ decreases from 0.9 to 0.8, while estimates of the proposed method have no big difference. When the covariate is continuous variable, equal sensitivity and specificity setting of $Y$ has no big influence on estimates of $\beta_1$, except decreases the bias of estimates in the proposed model. When the covariate is binary variable, the change of the sensitivity and the specificity of $Y$ decreases the bias of the estimates of $\beta_1$ from naive model and MC-SIMEX models, especially for the MC-SIMEX(LOG) model, while increases the bias of estimates from proposed method.

Comparing the case of a binary covariate with the case of a continuous covariate, under the lower sensitivity and higher specificity setting of $Y$, the results show smaller bias of $\beta_0$ in most of the models and $\beta_1$ in naive model and MC-SIMEX(L) model in the binary covariate case. But the continuous covariate case has smaller bias of $\beta_1$ in MC-SIMEX(Q), MC-SIMEX(LOG) and proposed method. Under the equal sensitivity and specificity of $Y$, binary covariate case shows smaller bias in both $\beta_0$ and $\beta_1$ in most of the models, except a larger bias of $\beta_1$ in the proposed model .

As a general conclusion we suggest that the proposed method substantially reduces bias compared to the naive estimator and MC-SIMEX method and its performance is comparable to results of ML approach by Neuhaus (2002) in the Table 1 of Kuchenhoff et al. (2006) with the same simulation setting in our simulation. ML approach requires running package called Numerical Algorithms Group in Fortran or calling Fortran routines from R, while the proposed method implement the whole procedure in R.

For clinical interest, we also calculate the estimated $P(Y = 1)$ according to equation (4.2) using the estimates of $\beta_0$ and $\beta_1$ in 200 simulations of each model under difference

scenarios. Specifically, four combinations of observed binary covariate $X$ (0 or 1) and observed binary $Y^{obs}$ (0 or 1) are assumed. We summary the percentage of bias under two scenarios with $(sens_y, spec_y)$=(0.7, 0.9) and $(sens_y, spec_y)$=(0.8, 0.8) separately. From Figure 4.2, we can see in majority of the scenarios, the estimated probability based on the proposed model in most of the cases has smaller percentage of bias comparing to naive models and MC-SIMEX models. The mean percentage of bias of estimated probabilities of $Y$ from our models are all below 2%.

### 4.4.2 $X$ is a misclassified binary variable.

After compare the simulation results under correctly measured covariates, we perform another 200 simulations each time with a sample size of 1000 to discuss the case with misclassified binary covarite. Here, $X$ are assumed as a misclassified binary covariate. The true $X$ are generated from a Bernoulli distribution with $P(X = 0) = P(X = 1) = 0.5$. The true value $Y$ of the binary outcome are generated from a Bernoulli distribution with the probability $P(Y = 1) = \frac{1}{1+e^{-\beta_0-\beta_1 X}}$. We apply the misclassification operation in Section 4.3.2 on $X$ and $Y$ to obtain the misclassified covariate and outcome. The misclassified covariate $X^{obs}$ are obtained based on two settings of sensitivity and specificity ($sens_x$ =0.8, $spec_x$ = 0.8, and $sens_x$ = 0.7, $spec_x$ = 0.9). The misclassified outcome $Y^{obs}$ are generated under two settings of sensitivity and specificity ($sens_y$ =0.8, $spec_y$ = 0.8, and $sens_y$ = 0.7, $spec_y$ = 0.9).

$\hat{\beta}_0$ and $\hat{\beta}_1$ are estimated from the proposed method, MC-SIMEX models for the three extrapolation functions (linear, quadratic, and log-linear), the naive model, and the true model. Mean, absolute bias (Abs Bias) and ESE of estimates $\beta_0$ and mean, bias percentage and ESE of estimates $\beta_1$ are present in the Table 4.2 under different settings of sensitivity and specificity of $Y$.

Overall, our estimates of $\beta_0$ and $\beta_1$ in most of the cases have much smaller bias than the estimates based on naive model and MC-SIMEX models, and are very close to the true models's estimates. The ESE of estimates from proposed method based on the 200 simulation's estimates are slightly larger than the other models.

The estimates of $\beta_0$ from MC-SIMEX models have slightly larger bias when sensitivity of $Y$ increases from 0.7 to 0.8 and specificity of $Y$ decreases from 0.9 to 0.8, while estimates of the proposed method have no big difference. Under the unequal sensitivity and specificity of $X$ , the change of the sensitivity and the specificity of $Y$ does not influence the bias of the estimates of $\beta_1$ from all the models, except that it decreases the bias of the MC-SIMEX(LOG). We have the same findings under the equal sensitivity and specificity of $X$.

Under the unequal sensitivity and specificity setting of $Y$, comparing the case of lower sensitivity and higher specificity of $X$ with equal sensitivity and specificity of $X$, the bias of estimates of $\beta_0$ have no big change in all models, while the bias of estimates of $\beta_1$ have no big change in naive model, MC-SIMEX(L), MC-SIMEX(Q) model and the proposed model, but smaller bias in MC-SIMEX(LOG) . Under the equal sensitivity and specificity setting of $Y$, comparing the case of lower sensitivity and higher specificity of $X$ with equal sensitivity and specificity of $X$, the bias of estimates of $\beta_0$ have no big change in all models, while the bias of estimates of $\beta_1$ have no big change in naive model, MC-SIMEX(L) and MC-SIMEX(Q) model, but smaller bias in MC-SIMEX(LOG) and the proposed model.

For clinical interest, we also calculate the estimated $P(Y = 1)$ according to equation (4.2) by using the estimates of $\beta_0$ and $\beta_1$ in 200 simulations of each model under four combinations of observed binary covariate $X^{obs}$ (0 or 1) and observed binary $Y^{obs}$ (0 or 1). We summarized the Bias% under different settings of sensitivity and specificity of $X$ in Figure 4.3 with $(sens_y, spec_y)$=(0.7, 0.9) and Figure 4.4 with $(sens_y, spec_y)$=(0.8, 0.8)

respectively. From Figure 4.3 and Figure 4.4, we can see in majority of the scenarios, the estimated probability from the proposed model has smaller percentage of bias compared to naive models and MC-SIMEX models.

### 4.4.3 AR(1) model of misclassified binary outcome.

We perform another 200 simulations each time with a sample size of 1000 to discuss the case of AR(1) model with longitudinal data. Here, the true value $Y_t$ of the binary outcome are generated from a Bernoulli distribution with the probability $P(Y_t = 1) = \frac{1}{1+e^{-\beta_0-\beta_1 Y_{t-1}}}$, where $t = 2, 3, ..., T$. The initial observation $Y_1$ are generated with the probability $P(Y_1 = 1) = \frac{1}{1+e^{-\beta_0}}$. The misclassified outcome $Y^{obs}$ are generated under two settings of sensitivity and specificity ($sens_y$ =0.8, $spec_y$ = 0.8, and $sens_y$ = 0.7, $spec_y$ = 0.9).

$\hat{\beta}_0$ and $\hat{\beta}_1$ are estimated from the proposed method, the naive model, and the true model. The estimates in true model are obtained by regressing the correctly measured $Y_t$ on $Y_{t-1}$. Mean, absolute bias and ESE of estimates $\beta_0$ and mean, bias percentage and ESE of estimates $\beta_1$ are present in the Table 4.3 under different settings of sensitivity and specificity of $Y$.

Overall, our estimates of $\beta_0$ and $\beta_1$ in most of the cases have much smaller bias than the estimates based on naive model. The ESE of estimates from proposed method based on the 200 simulation's estimates are slightly larger than the other two models. Different settings of sensitivity and specificity have no much influence on the estimates from the proposed model, but affect the estimates of $\beta_0$ in the naive model.

For clinical interest, we also estimate $P(Y = 1)$ according to the equation (4.2) using the estimates of $\beta_0$ and $\beta_1$ in 200 simulations of each model under different scenarios. Specifically, four combinations of $Y_t^{obs}$ (0 or 1) and $Y_{t-1}^{obs}$ (0 or 1) are assumed. We summarize the percentage of bias under two scenarios with ($sens_y, spec_y$)=(0.7, 0.9) and

$(sens_y, spec_y)$=(0.8, 0.8) separately. From Figure 4.5, we can see in majority of the scenarios, the estimated probability based on the proposed model has smaller percentage of bias comparing to naive models in all the cases. The mean percentage of bias of estimated probabilities of $Y$ from our models are all below 9%.

### 4.5 Application to IU Phone Study

Two hundreds and forty participants with negative result of NAAT at the baseline are included in this analysis. Among those participants,88% (212/240) are African-American; 55% (133/240) are women; 9.2% (22/240) participants acquire positive result of CT+, GC+, or TV+ NAAT in their last test; 40% (96/240) report at least 1 partner change.

Our interest is adjusting the NAAT based on the previous test result in the model with known sensitivity and specificity. In our analysis, NAAT+ is defined as a positive result in any NAAT of CT, GC or TV. For each subject, among the 13 lab test results, we include the 12 test results after enrollment as outcome and first 12 test results since the baseline test as one of the predictors. Because we define the infection status as any positivity of CT, GC and TV and specimens are collected from male's urine samples and female's vaginal samples, based on the studies implemented by Taylor and et al. (2012), Pol and et al. (2012)Pol and et al. (2013) and Gaydos and et al. (2013), we decide to use 0.975 and 0.99 as sensitivity and specificity respectively to be more conservative in our data analysis.

The proposed model and naive model are applied to the original data set. According to the GLM based on the observed test results, the estimates of intercept and autoregressive effect are $-3.857$ (p-value$<$0.001) and 5.015 (p-value$<$0.001) respectively. The estimates from the proposed model under the setting of sensitivity 0.975 and specificity to 0.99 are $-4.581$ for intercept and 6.942 for autoregressive effect respectively. The results of both models show that the current test result had positive association with previous test result.

Table 4.4 shows the summaries of the 1000 bootstrap results based on both proposed model and GLM, with the same sample size (n=2796) as the original data set. Bootstrap sampling provides the mean, empirical standard error (ESE) and empirical quantiles of all the estimates. From Table 4.4, we find the absolute value and ESE of estimates of intercept and autoregressive effect are both larger in proposed model comparing to GLM. The 95% empirical confidence intervals of both models do not cover "0", while the proposed model has wider confidence interval than GLM. We compare the estimated probability of positivity from the proposed model and GLM in Figure 4.6. In the cases with consistent previous and current test results (case 1 and 4), both model give very similar results. In the cases with different previous and current test results (case 2 and 3), the estimated probability provided by the proposed model is more influenced by the previous results, which is reasonable since the proposed model on average provides larger estimates of autoregressive effect.

In the IU Phone data, only 14.17% (34/240) of the participants acquired positivity with 5.72%(160/2796) positivity test results during the observing time. Because of this high imbalance between positive events and negative events, the the value of sensitivity and specificity have influence on the estimates of $\beta_0$ and $\beta_1$ in the AR(1) model. In our data application, we make a conservative choice of sensitivity and specificity based on the literature related to NAAT. But in case there might be other more appropriate choices, we establish a sensitivity analysis to summarize the estimates of coefficients under different combination of sensitivity (ranges from 0.975 to 1) and specificity (ranges from 0.99 to 1) in the Figure 4.7. As it is shown in the Figure 4.7, when the sensitivity and specificity increase to 1, the estimates of coefficients are approaching to the GLM's estimates, which is the case without adjustment of misclassification.

For clinical interest, we also display the estimated probability of true positive test result based on observed test results at $t$ and $t-1$ with the estimates of coefficients from the

sensitivity analysis above under different sensitivity and specificity setting in Figure 4.8. We can tell from the Figure 4.8 that the variability of the estimated probability is very small when $Y_t^{obs} = Y_{t-1}^{obs}$, but gets larger when $Y_t^{obs} \neq Y_{t-1}^{obs}$. We also find the value of specificity influences the estimated $P(Y = 1)$ quite much when the $Y^{obs} = 1$, while the value of sensitivity affects the estimated $P(Y = 1)$ more when the $Y^{obs} = 0$.

## 4.6 Discussion

The proposed models provide the adjusted probability of true status based on different scenarios of covariates in both cross-sectional data and longitudinal data with AR(1) model. The estimates from those models have smaller bias than the MC-SIMEX models and naive models for cross-sectional data and much smaller bias than the naive model for longitudinal data. Our method can be applied to different clinical studies with imperfect diagnosis test for two main different purposes. At the population level, more accurate covariate specific probability of positivity can be estimated with correction of the misclassification, which is helpful in public health intervention. At the individual level, before the subject taking any test samples or doing any test, he/she can get an estimates of the probability of disease based on his/her information of test history and other covariates in the model in advance. This would be really helpful for the people who do not get the chance to visit the clinics for test but concern about their health status.

In the real data application, we implement a sensitivity analysis to provide possible range of estimated probability of positivity based on different combination of sensitivity and specificity, which also provides a tool for analysis when there is limited amount of information about sensitivity and specificity of the test. Based on the way we are building the joint probability in EM part, our AR(1) model can be extended to the case with value of sensitivity and specificity changing with time.

In our discussion of model with misclassified covariates in Section 4.3.2, extra covariates with correct measurement can be added into the model with no cost. If we are trying to add more misclassified binary covariates simultaneously, the dimension of steps in calculating the joint probability will increase geometrically. But it is still within reasonable range when the number of covariate is less than 50. Similarly, we allow more historical test information (less than 50) involved in our autoregressive model setting as the AR(k) model.

In our model, we didn't discussion the case with continuous variable with measurement error as covariates, because our interest is using the previous and current test results to estimate the true probability of positivity of current STIs status. This can be considered as the limitation in term of the generalization of the method, but we propose a straightforward method which is efficient and easy to be implemented in clinical practice.

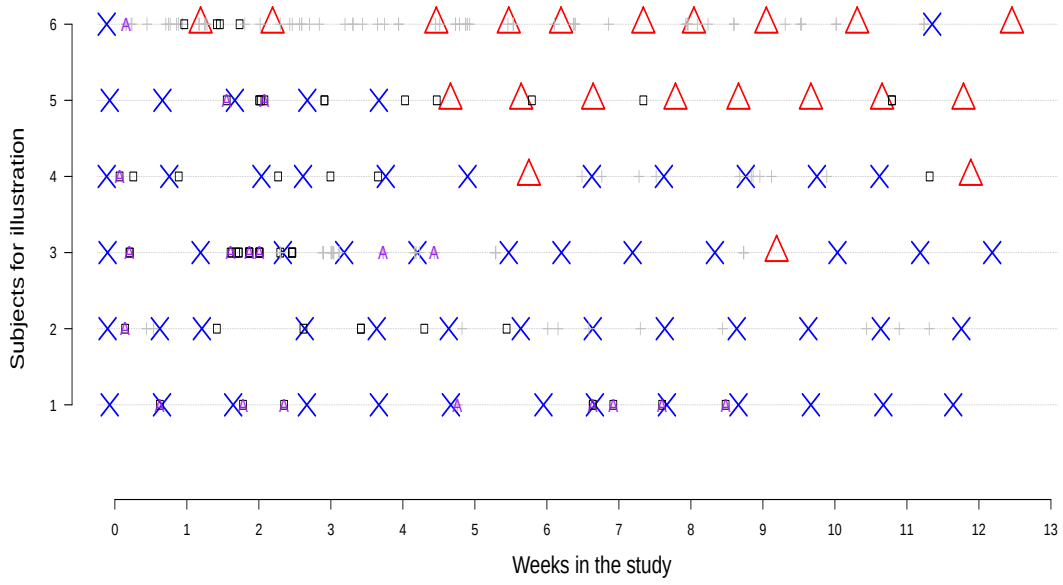


Figure 4.1: Summary of STIs test results and condom use for participants with negative STIs test results at baseline. Cross indicates negative test and triangle indicates positive test. The event with a new reported partner is indicated by a "A". The event with condom use is square, otherwise, is gray cross.

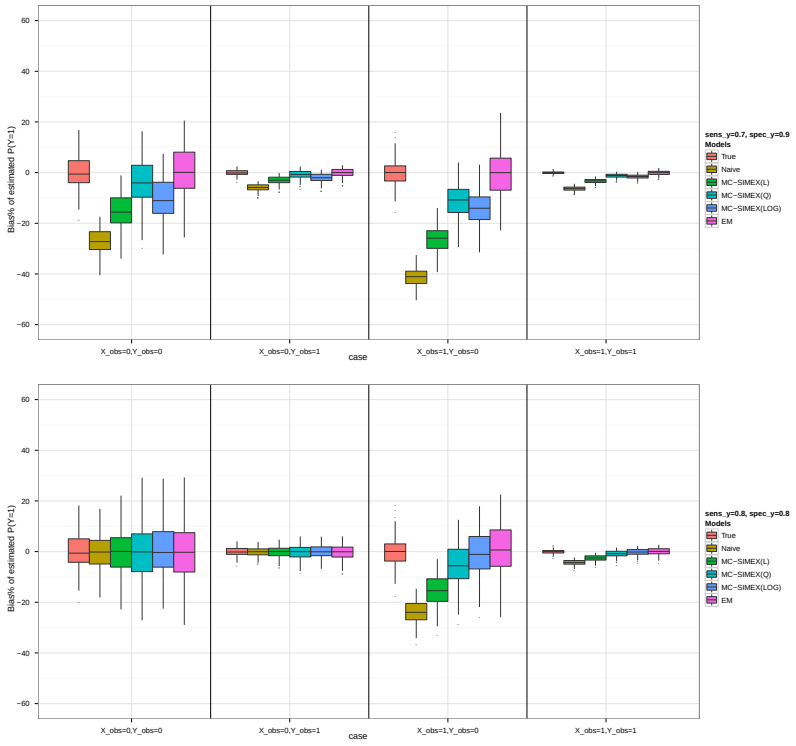


Figure 4.2: Summary of the percentage of the bias of P(Y=1) with misclassified $Y$

Table 4.1: Simulation results of logistic regression of a misclassified response $Y$ on a binary or a continuous covariate $X$. The true regression coefficients are $\beta_0 = 0$ and $\beta_1 = 1$ (based on 200 simulations each with sample size =1000).

| | | $(sens_y, spec_y) = (0.7, 0.9)$ | | | $(sens_y, spec_y) = (0.8, 0.8)$ | | |
|---|---|---|---|---|---|---|---|
| Models | | Mean | Abs Bias | ESE | Mean | Abs Bias | ESE |
| | | | | X is a binary covariate | | | |
| True model | b0 | 0.002 | 0.002 | 0.089 | 0.002 | 0.002 | 0.089 |
| Naive model | b0 | −0.404 | 0.404 | 0.081 | −0.005 | 0.005 | 0.085 |
| MC-SIMEX(L) | b0 | −0.218 | 0.218 | 0.101 | −0.005 | 0.005 | 0.108 |
| MC-SIMEX(Q) | b0 | −0.055 | 0.055 | 0.122 | −0.006 | 0.006 | 0.134 |
| MC-SIMEX(LOG) | b0 | −0.149 | 0.149 | 0.118 | 0.007 | 0.007 | 0.125 |
| EM | b0 | 0.003 | 0.003 | 0.129 | −0.008 | 0.008 | 0.142 |
| | | | | X is a continuous covariate | | | |
| True model | b0 | 0 | 0 | 0.082 | 0 | 0 | 0.082 |
| Naive model | b0 | −0.431 | 0.431 | 0.067 | 0.003 | 0.003 | 0.066 |
| MC-SIMEX(L) | b0 | −0.253 | 0.253 | 0.085 | 0.004 | 0.004 | 0.085 |
| MC-SIMEX(Q) | b0 | −0.091 | 0.091 | 0.104 | 0.005 | 0.005 | 0.107 |
| MC-SIMEX(LOG) | b0 | −0.190 | 0.190 | 0.098 | 0.019 | 0.019 | 0.102 |
| EM | b0 | 0.008 | 0.008 | 0.118 | 0.005 | 0.005 | 0.123 |
| | | Mean | Bias% | ESE | Mean | Bias% | ESE |
| | | | | X is a binary covariate | | | |
| True model | b1 | 0.994 | 0.599 | 0.131 | 0.994 | 0.599 | 0.131 |
| Naive model | b1 | 0.555 | 44.549 | 0.115 | 0.579 | 42.115 | 0.121 |
| MC-SIMEX(L) | b1 | 0.697 | 30.284 | 0.145 | 0.738 | 26.248 | 0.154 |
| MC-SIMEX(Q) | b1 | 0.838 | 16.203 | 0.179 | 0.913 | 8.737 | 0.196 |
| MC-SIMEX(LOG) | b1 | 0.873 | 12.724 | 0.181 | 0.976 | 2.400 | 0.205 |
| EM | b1 | 0.993 | 0.656 | 0.216 | 1.022 | 2.158 | 0.220 |
| | | | | X is a continuous covariate | | | |
| True model | b1 | 1.004 | 0.372 | 0.088 | 1.004 | 0.372 | 0.088 |
| Naive model | b1 | 0.549 | 45.082 | 0.073 | 0.525 | 47.499 | 0.070 |
| MC-SIMEX(L) | b1 | 0.693 | 30.679 | 0.093 | 0.670 | 33.048 | 0.090 |
| MC-SIMEX(Q) | b1 | 0.845 | 15.469 | 0.122 | 0.844 | 15.616 | 0.121 |
| MC-SIMEX(LOG) | b1 | 0.878 | 12.204 | 0.121 | 0.892 | 10.808 | 0.124 |
| EM | b1 | 1.004 | 0.407 | 0.156 | 0.999 | 0.096 | 0.161 |

Table 4.2: Simulation results of logistic regression of a misclassified response $Y$ on a misclassified binary covariate $X$. The true regression coefficients are $\beta_0 = 0$ and $\beta_1 = 1$ (based on 200 simulations each with sample size =1000).

| | | $(sens_y, spec_y) = (0.7, 0.9)$ | | | $(sens_y, spec_y) = (0.8, 0.8)$ | | |
|---|---|---|---|---|---|---|---|
| Models | | Mean | Abs Bias | ESE | Mean | Abs Bias | ESE |
| | | | | $(sens_x, spec_x)$=(0.7, 0.9) | | | |
| True model | b0 | 0.006 | 0.006 | 0.085 | 0.006 | 0.006 | 0.085 |
| Naive model | b0 | −0.261 | 0.261 | 0.079 | 0.143 | 0.143 | 0.080 |
| MC-SIMEX(L) | b0 | −0.043 | 0.043 | 0.101 | 0.179 | 0.179 | 0.104 |
| MC-SIMEX(Q) | b0 | 0.079 | 0.079 | 0.138 | 0.149 | 0.149 | 0.145 |
| MC-SIMEX(LOG) | b0 | 0.046 | 0.046 | 0.120 | 0.206 | 0.206 | 0.128 |
| EM | b0 | 0.007 | 0.007 | 0.197 | 0.012 | 0.012 | 0.201 |
| | | | | $(sens_x, spec_x)$ = (0.8, 0.8) | | | |
| True model | b0 | 0.006 | 0.006 | 0.085 | 0.006 | 0.006 | 0.085 |
| Naive model | b0 | −0.287 | 0.287 | 0.091 | 0.118 | 0.118 | 0.091 |
| MC-SIMEX(L) | b0 | −0.078 | 0.078 | 0.119 | 0.145 | 0.145 | 0.117 |
| MC-SIMEX(Q) | b0 | 0.037 | 0.037 | 0.163 | 0.108 | 0.108 | 0.167 |
| MC-SIMEX(LOG) | b0 | 0.003 | 0.003 | 0.14 | 0.167 | 0.167 | 0.144 |
| EM | b0 | 0.011 | 0.011 | 0.212 | 0.017 | 0.017 | 0.210 |
| | | Mean | Bias% | ESE | Mean | Bias% | ESE |
| | | | | $(sens_x, spec_x)$ = (0.7, 0.9) | | | |
| True model | b1 | 1.002 | 0.182 | 0.133 | 1.002 | 0.182 | 0.133 |
| Naive model | b1 | 0.345 | 65.462 | 0.128 | 0.353 | 64.73 | 0.133 |
| MC-SIMEX(L) | b1 | 0.453 | 54.712 | 0.167 | 0.462 | 53.762 | 0.175 |
| MC-SIMEX(Q) | b1 | 0.666 | 33.355 | 0.247 | 0.700 | 30.028 | 0.273 |
| MC-SIMEX(LOG) | b1 | 0.949 | 5.125 | 0.498 | 0.983 | 1.740 | 0.394 |
| EM | b1 | 1.008 | 0.753 | 0.400 | 1.004 | 0.407 | 0.402 |
| | | | | $(sens_x, spec_x)$ = (0.8, 0.8) | | | |
| True model | b1 | 1.002 | 0.182 | 0.133 | 1.002 | 0.182 | 0.133 |
| Naive model | b1 | 0.328 | 67.234 | 0.130 | 0.331 | 66.886 | 0.130 |
| MC-SIMEX(L) | b1 | 0.429 | 57.125 | 0.170 | 0.433 | 56.714 | 0.170 |
| MC-SIMEX(Q) | b1 | 0.653 | 34.659 | 0.259 | 0.676 | 32.363 | 0.273 |
| MC-SIMEX(LOG) | b1 | 0.894 | 10.639 | 0.389 | 0.972 | 2.773 | 0.418 |
| EM | b1 | 0.994 | 0.602 | 0.412 | 0.989 | 1.074 | 0.409 |

Table 4.3: Simulation results of logistic regression of a misclassified response $Y_t$ on $Y_{t-1}$. The true regression coefficients are $\beta_0 = 0$ and $\beta_1 = 1$ (based on 200 simulations each with sample size =1000).

| | | $(sens_y, spec_y) = (0.7, 0.9)$ | | | $(sens_y, spec_y) = (0.8, 0.8)$ | | |
|---|---|---|---|---|---|---|---|
| Models | | Mean | Abs Bias | ESE | Mean | Abs Bias | ESE |
| True model | b0 | −0.003 | 0.003 | 0.048 | −0.003 | 0.003 | 0.048 |
| Naive model | b0 | −0.318 | 0.318 | 0.043 | 0.060 | 0.060 | 0.045 |
| EM | b0 | 0.014 | 0.014 | 0.074 | 0.014 | 0.014 | 0.081 |
| | | Mean | Bias% | ESE | Mean | Bias% | ESE |
| True model | b1 | 1.004 | 0.350 | 0.078 | 1.004 | 0.350 | 0.078 |
| Naive model | b1 | 0.420 | 57.997 | 0.080 | 0.419 | 58.085 | 0.080 |
| EM | b1 | 0.902 | 9.768 | 0.196 | 0.916 | 8.428 | 0.199 |

Table 4.4: The comparison of estimates between proposed model and GLM on 1000 bootstrap data of IU Phone Study

| | EM | | GLM | |
|---|---|---|---|---|
| | beta0 | beta1 | beta0 | beta1 |
| $Mean$ | −4.604 | 7.242 | −3.858 | 5.025 |
| 2.5% | −5.297 | 5.943 | −4.131 | 4.568 |
| 5% | −5.148 | 6.065 | −4.083 | 4.653 |
| 10% | −4.978 | 6.186 | −4.035 | 4.718 |
| 90% | −4.257 | 8.206 | −3.692 | 5.357 |
| 95% | −4.176 | 9.491 | −3.647 | 5.448 |
| 97.5% | −4.123 | 12.834 | −3.614 | 5.550 |

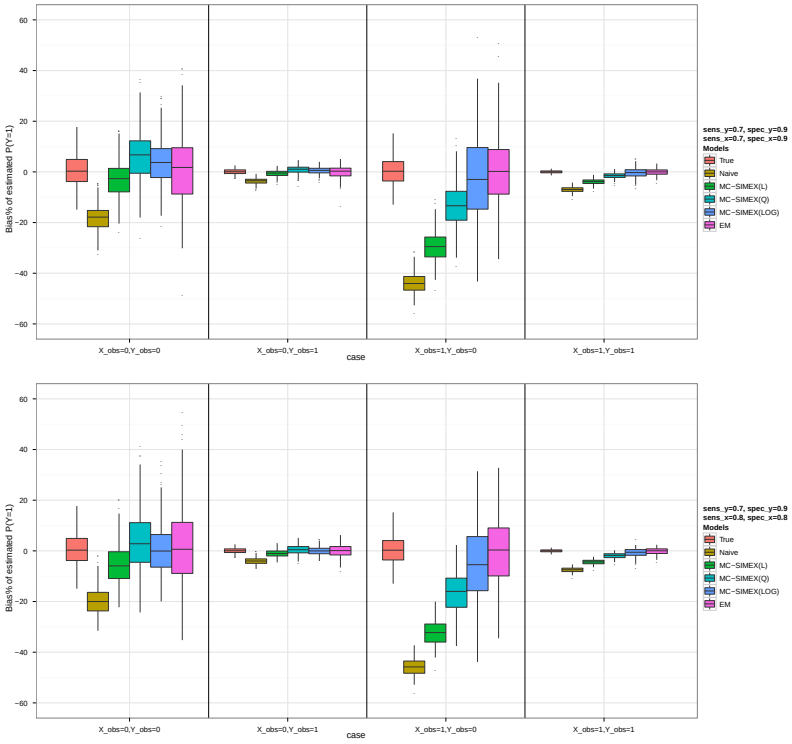


Figure 4.3: Summary of the percentage of the bias of P(Y=1) with misclassified $X$ and $Y$ $[(sens_y, spec_y)=(0.7, 0.9)]$

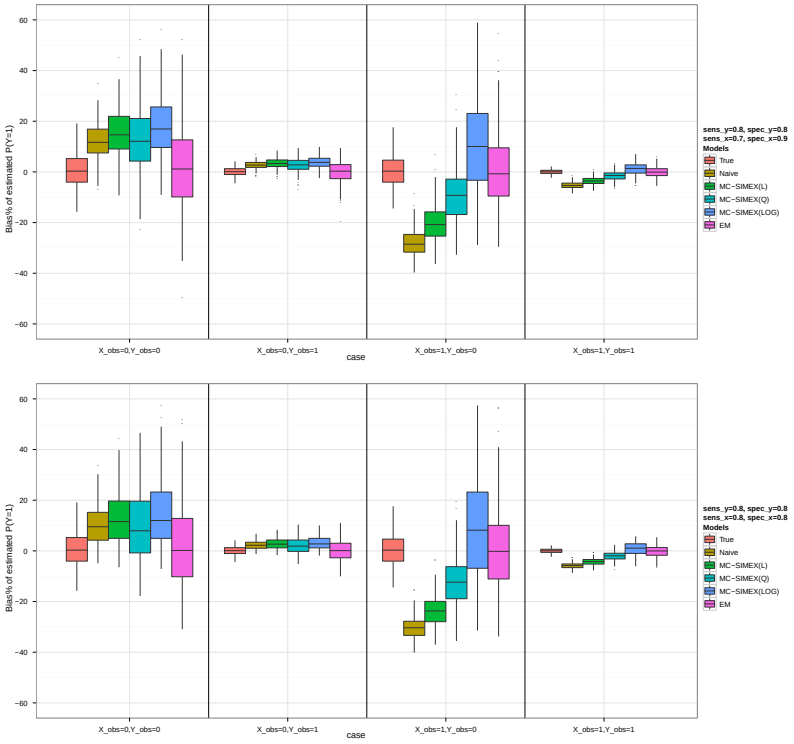


Figure 4.4: Summary of the percentage of the bias of P(Y=1) and P(Y=0) with misclassified $X$ and $Y$ $[(sens_y, spec_y)$=(0.8, 0.8)]

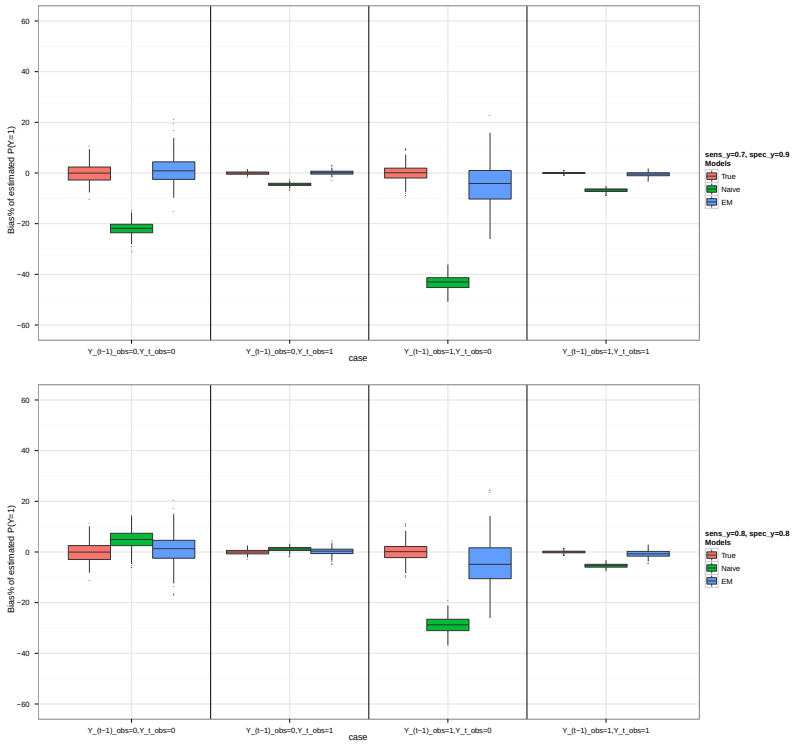


Figure 4.5: Summary of the percentage of the bias of P(Y=1) with misclassified $Y$ in AR(1) model

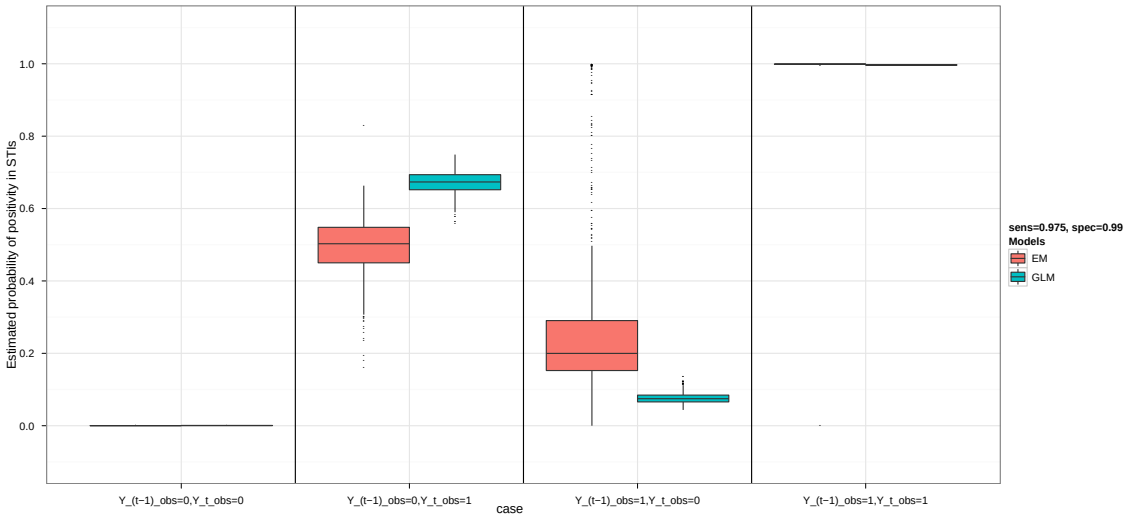


Figure 4.6: The comparison of estimated probability of positivity of STIs between proposed model and GLM based on the estimates from 1000 bootstrap data of IU Phone Study

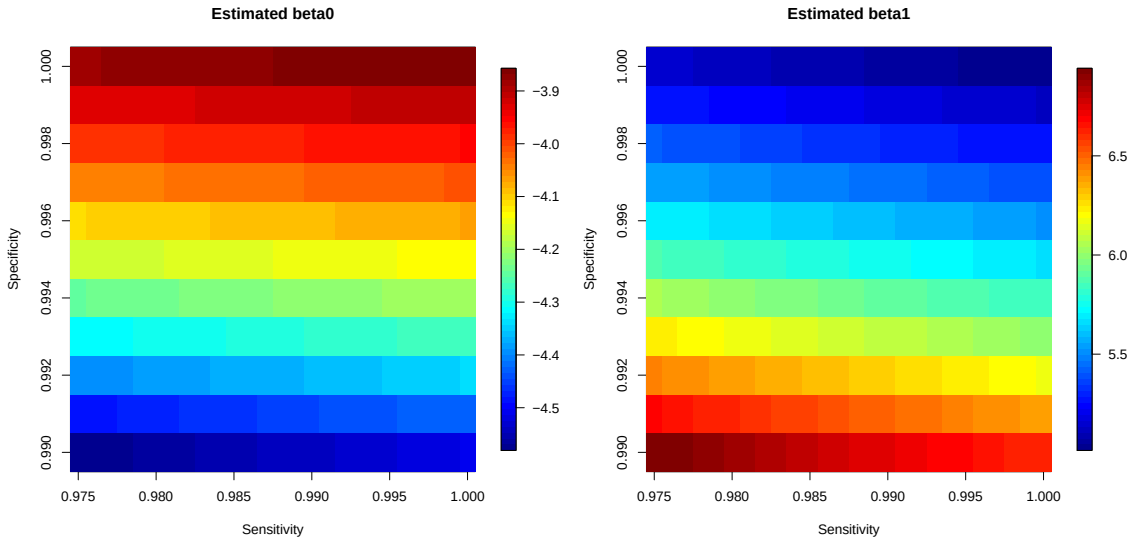


Figure 4.7: Summary of the estimates in AR(1) model of IU Phone Study under different sensitivity and specificity

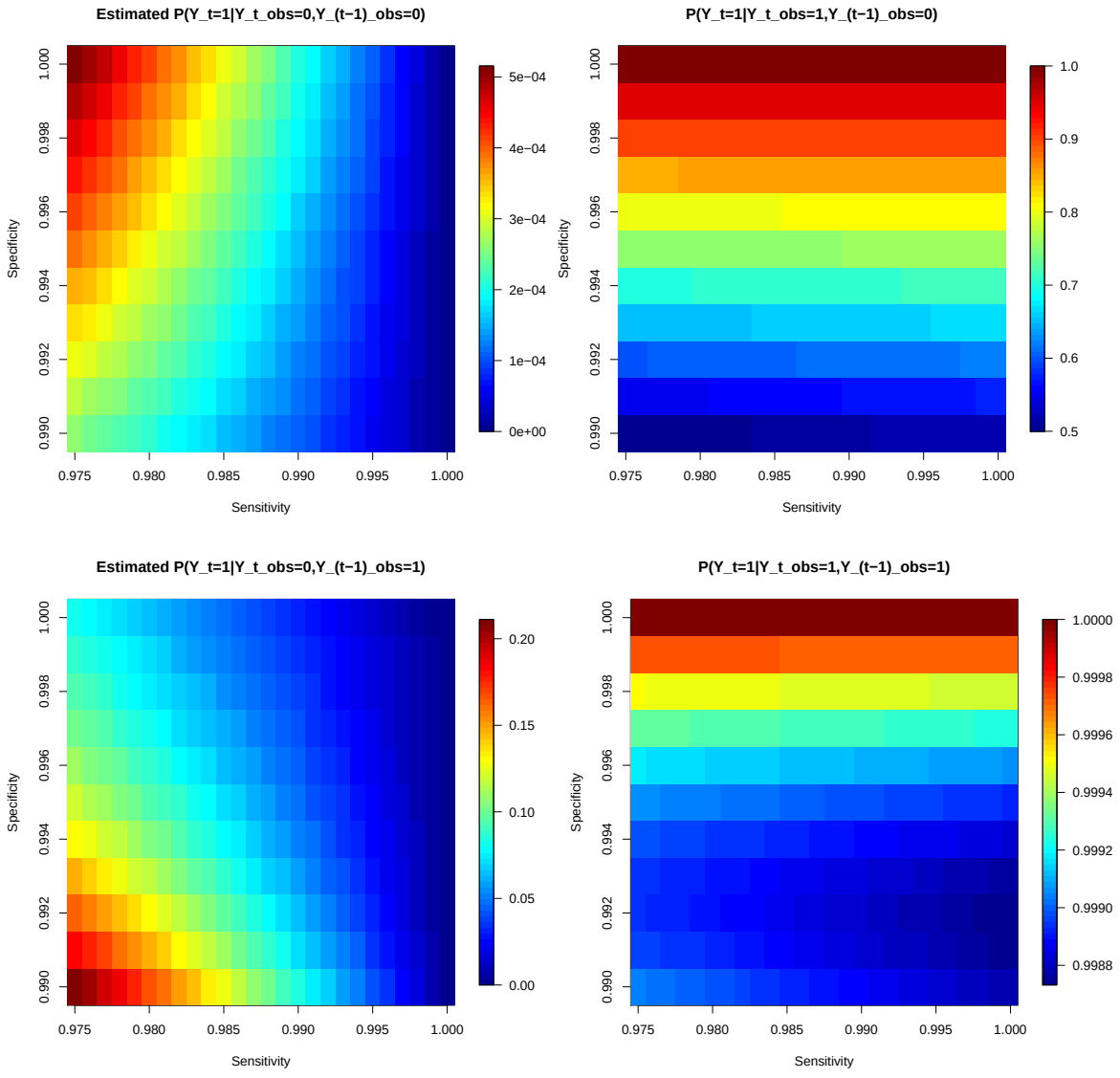


Figure 4.8: Summary of the estimated probability of positivity in AR(1) model of IU Phone Study under different sensitivity and specificity

# Chapter 5

## Condom use as a function of number of coital events in new relationships

### 5.1 Introduction

Condom use follows changes in the larger interpersonal and sexual relationship, with the proportion of condom-protected coital events declining in new relationships within a few weeks of first sex between two partners. (Ku et al. (1994) Fortenberry et al. (2002) Bauman and Berman (2005)) Reasons for decline in condom use with increased relationship duration include diminished perceived sexually transmitted infection (STI) risk, increased within-dyad trust, and shifts to non-barrier contraception.(Manning et al. (2009)) Dyad members' subjective assessments of sexual satisfaction, relationship quality, and relationship satisfaction are all related to relationship durability, which in turn affects condom use through decreased likelihood of partner change and increased coital frequency.(Sayegh et al. (2006))

Understanding the pace of decline of condom use in relationships is relevant for STI prevention efforts because the duration of infectiousness for an STI acquired in a previous sexual relationship may be several weeks or months, potentially extending past a period of relatively higher condom use in newer dyads.(Shew et al. (2006) Anderson (1991)) Concurrent sexual partners, as well as sequential partners for whom the interval between partners is less than the duration of infectiousness, could therefore be exposed to infection if condom use is irregular or ceases.(Kraut-Becher and Aral (2003) Matson et al. (2012) Manhar et al. (2002)) This may explain - at least in part - the often-observed association of STI and "new" sexual partners.(Ott et al. (2011)) Relationship duration thus frames a number of issues of relevance to understanding of STI transmission and prevention.

However, relationship duration as a reflection of changes in condom use is potentially incomplete in that coitus is the exposure of interest, and some dyads - particularly adolescents - may have substantial intervals of non-coital sexual interaction that precede first coitus. Moreover, substantial between-dyad variability exists in the number of exposures per unit time (i.e., in coital frequency).(Brewis and Meyer (2005)) An alternative possibility is that the need for condoms is assessed by dyad members according to a metric such as the accrual of sexual exposures within the dyad, rather than by the time interval over which those events are dispersed. First coital exposure with a partner is an easily recognized signal for condom use.(Shafii et al. (2007)) Dyads' evaluations of condom use for second (and subsequent) coital exposures is much less clearly understood, as these events may occur within a few hours or days afterward. The interpersonal and neurohormonal reward effects of partnered sex accrue based on sexual experiences, contributing to development of interpersonal trust.(Cacioppo et al. (2012)) Trust is among the most commonly cited reasons for discontinuation of condom use.(Bauman and Berman (2005) Manning et al. (2009) Hattori (2014) Bolton et al. (2010) Willig (1997)) Thus, perception of the need for condoms may be quite different for dyads whose second coital exposure occurs within 24 hours of the first, as compared to those whose subsequent coital exposure occurs after an interval of several weeks.(Ott et al. (2010))

The purpose of this paper, then, is to explore an alternative understanding of factors associated with condom discontinuation by prospectively assessing condom use as a function of the number of coital exposures reported with a specific partner. Because decisions to use condoms may also be influenced by interpersonal and sexual aspects of relationships, we assessed differences in condom use trajectories as a function of relationship quality, relationship satisfaction, and sexual satisfaction.(T. Ein-Dor (2012)) Because men and women may

differ in the relative weight given to emotional and sexual characteristics of relationships, analyses were done separately for men and women.(Thompson and O'Sullivan (2012))

## 5.2 Materials and Methods

Data were obtained from a prospective 84-days (12-weeks) study designed to examine sexual behaviors and incident STI. Participants were recruited from the patient population of a county sexually transmitted diseases clinic but were not necessarily clinic patients at the time of enrollment. Eligibility criteria were ages 18 to 29 years (inclusive), English speaking, and planning to reside in the area for the subsequent 84 days. The Institutional Review Board of Indiana University Purdue University Indianapolis approved this study. All participants provided informed consent.

The primary mode of data collection was via three-times daily self-reports of coital and non-coital sexual behaviors, condom use, and relationship assessments, recorded with project-furnished cellular telephones and service. The expected number of entries was thus 252 entries per participant. Daily diary completion rate was 87.7%. Other methodological details are previously published (Hensel et al. (2012)). At pre-selected 8-hour intervals, participants responded to a series of questions to identify sexual and non-sexual interactions with specific partners. In each eight-hour reporting period, participants identified any partner, time of each coital event (up to four events within the same eight-hour reporting period), condom use for each coital event, as well as relationship satisfaction and sexual satisfaction. Relationship and sexual satisfaction were measured by single item rankings from 1 ('very low') to 10 ('very high'). Coital events were analyzed on the basis of sequences of coital events (not necessarily on successive days) with a specific sexual partner. Each new sequence of events formed a separate analytic frame, even if the partner had been identified earlier. Number of sequences of exposures did not necessarily equal number of partners,

because sexual exposures with different partners could be interspersed.(Fortenberry et al. (2002))

Statistical analyses were based on the generalized additive mixed models (GAMMs, an extension of generalized linear models) that uses smooth functions to model the mean trajectory and account for the hierarchical structure of longitudinal data. To apply GAMMs to our data, we included two nested random effects (at a partner level and a subject level respectively) to account for correlations among repeated within-partner coital events and correlations among the partners of the same subject. Specifically, a logistic additive mixed model was used to estimate the association between the event-specific condom use (coded as no/yes), cumulative number of coital events and other covariates of interest. As the dependence of the condom use on cumulative number of coital events was of primary interest, this predictor was always kept in the model. Instead of using parametric method of modeling condom use probability with cumulative number of coital events (e.g., linear models with quadratic or polynomial forms, which would be inappropriate for our data), we used a smoothing function as a more flexible, data-driven nonparametric approach.

Relationship satisfaction, sexual satisfaction and gender were included as additional covariates. Because of a substantial positive correlation between the relationship satisfaction and sexual satisfaction, models including either satisfaction score were considered separately. To study the association between event-specific condom use probability and each related covariate, we first considered models including age, gender, relationship satisfaction and sexual satisfaction separately. Age was not associated with condom and was not included in subsequent analyses. Multivariable models including gender and either relationship satisfaction or sexual satisfaction were established consecutively to study the interaction of those covariates. R-2.15.3 (www.r-project.org) was used to conduct the data

analysis. Level of statistical significance was set at p¡0.05 and 95% confidence intervals of estimates were reported.

### 5.3 Results

The sample consisted of 115 participants (55/115 [48%] women; 103/115 [90%] African-American). Median number of lifetime partners was 31 and 22 for women and men, respectively. About 24% of women and 18% men had chlamydia, gonorrhea, or trichomonas at enrollment.

Participants reported 676 intervals of sex (419 for men; 257 for women) with a new partner. Preliminary analyses showed that less than one percent of sexual sequences consisted of more than 40 coital events. To reduce risk of biases analyses due to this extreme skew, number of coital events was truncated at 40 per participant.

Overall relationship satisfaction and sexual satisfaction scores were high, with 67.7% and 74.2% of relationship satisfaction and sexual satisfaction scores, respectively, at 9 or 10. To highlight potential influences of very high relationship and sexual satisfaction on condom use, and to reduce bias due to the skewed distribution of scores, we recoded both relationship satisfaction and sexual satisfaction: low satisfaction was defined as less than or equal to 8 and high satisfaction as more than 8.

Exploratory data analysis using a simple summary of the condom use percentage defined as the total number of condom-protected events divided by the total number of coital events vs. the cumulative number of coital events is presented in Figure 5.1. The estimated percentages for men and women are displayed in the left and right panels, respectively. Both men and women experienced a sharp decline in condom use percentage during the first few coital events. Men started at a higher average condom use percentage of 56% and quickly declined to 26% during the first 17 coital events, with condom use stabilizing

around 25% for the subsequent coital events. Women started at a lower average condom use percentage of 43% and sharply dropped to 6% during the first 17 coital events, remaining at this low level during subsequent events.

The univariable analyses show that there is no significant difference in condom use probability by gender (odds ratio OR = 0.81, 95% confidence interval CI= [0.25, 2.64]). In addition, neither the dichomtomized relationship satisfaction (OR = 1.08, CI = [0.76, 1.54]) nor sexual satisfaction (OR = 0.85, CI = [0.60,1.19] ) scores were associated with condom use probability.

Separate multivariable analyses were conducted to include effects of gender relationship satisfaction, and cumulative coital events, and gender, sexual satisfaction, and cumulative coital events on probability of condom use. The multivariable models showed that the interaction effect of gender by relationship satisfaction was a significant predictor of condom use probability with women reporting high relationship satisfaction being the least likely to use condoms. Men with higher relationship satisfaction had significantly higher odds of condom use (OR = 1.53, CI = [1.02, 2.30]) than the men with lower relationship satisfaction, while women with higher relationship satisfaction have significantly lower odds of condom use (OR = 0.40, CI = [0.21, 0.79] than the women with lower relationship satisfaction.

Similarly, the interaction of gender by sexual satisfaction was significant with women reporting high sexual satisfaction being less likely to use condoms, while men with higher sexual satisfaction were not significantly different in condom use probability (OR = 1.16, CI = [0.78, 1.74]) from the men with lower sexual satisfaction, while women with higher sexual satisfaction had significantly lower odds of condom use (OR = 0.38, CI = [0.21, 0.72] than the women with lower sexual satisfaction.

We tested the possibility of gender differences in the shape of the condom use probability curve as coital exposures accrued. Based on the adjusted R2 comparison between the

models, assuming different gender smoothing function showed trivial improvement of model fit (relationship satisfaction: 0.058[same shape] vs 0.057[different shape]; sexual satisfaction: 0.066 [same shape] vs 0.064 [different shape]). Therefore, we used the same smoothing function for both genders in final GAMM analysis of condom use probability. Figure 5.2 shows the estimated condom use probability as a function of the cumulative number of coital events for participants with high relationship satisfaction. From the trajectory of the predicted curve, women's condom use probability shows a rapid decrease from 36% to 8% during the first 9 coital events, followed by a low level between 3% and 8% afterwards. Men's condom use probability in the high relationship satisfaction group also decreases rapidly from 55% to 16% during the first 9 coital events and stays between 7 and 15% during the following coital events.

## 5.4 Discussion

We showed that condom use declines sharply for both men and women during the first 9 coital exposures in a relationship, then remains stable at much lower levels. This suggests that dyads evaluate the need for condom use based - at least in part - on accrual of exposures rather than relationship duration per se. It may be that decisions to continue a relationship with second and subsequent sexual exposures incorporate assessments of familiarity, trust and intimacy that mitigate perceptions of risk.(Matson et al. (2011)) These decisions may contribute to condom non-use in the face of continued objective STI risk. In addition to providing data on condom use in an adult sample of both men and women, these data add to existing literature by shifting focus to specifically sexual aspects of relationships rather than relationship duration.

We also showed that higher levels of both relationship satisfaction and sexual satisfaction are associated with even more rapid declines in condom use, after very few coital

exposures, particularly for women. The association of relationship satisfaction with decline in condom use suggests that differential investment in relationships, particularly in terms of the relationship affirming functions of sex are associated with different experiences of condom use as relationships progress.(Edwards et al. (2014)) The association of sexual satisfaction and condom use may reflect influence of higher levels of coital frequency in dyads with high levels of sexual satisfaction, with consequent rapid decline in condom use.(Senn et al. (2014)) Alternatively, common perceptions that condoms interfere with pleasure and sexual function may lead dyads to abandon condoms in order to preserve high levels of sexual satisfaction.(Randolph et al. (2007) Lehmiller et al. (2014))

Taken together, our findings have several implications for enhancing condom use for STI prevention. The data call attention to condom use as a dynamic prevention behavior enacted (or not) in the immediate context of a sexual event. Each coital exposure after the first builds on dyad members' accrued experience as partners. These experiences may generate trust and intimacy in addition to fulfilment of sexual needs.(Corbett et al. (2009) Denes (2012)) The finding that condom use is not a fixed characteristic of a given dyad's sexual relationship means that approaches to teaching condom use negotiation skills may change as well.(Zukoski et al. (2009)) Public health messages that emphasize associations of STI and condom use with risky "casual" sex means that dyads with ongoing sexual relations - by definition no longer casual - feel out of danger as the number of sexual exposures increases.(Royer et al. (2009) Lane and Viney (2002)) It may be that supplementing the long-established "risk" paradigm of STI prevention with a "sexual health" perspective could help dyads better align long-term condom use with the interpersonal demands of close relationships.(Fortenberry (2013) Harvey et al. (2009))

Inferences drawn from these data should be considered in light of several issues of the study design and research methods. First, the sample is of young adults with relatively

high rates of STI reflective of an STI clinic population. Young adults - particularly those less than age 25 - have very high STI rates although condom maintenance is relatively less studied. This means, however, that adolescents under age 18 - also at high STI - are not represented in these data. We also measured coitus three times daily rather than daily or retrospectively as in previous research. This methodological difference necessarily emphasizes coital events over elapsed time. Finally, we included a small number sequences of coital events with a previously identified partner when those events were separated by coitus with another partner. We treated these sequences as unrelated, although condom use may decline after even fewer coital events in such concurrent relationships.

Number of coital events may serve as an important source of sexual risk evaluation and inform decisions about condom use. Women's condom use, in particular, is associated with relationship satisfaction and sexual satisfaction. A relationship focused approach to condom use and condom maintenance may be particularly important in STI prevention among relatively well-established couples.

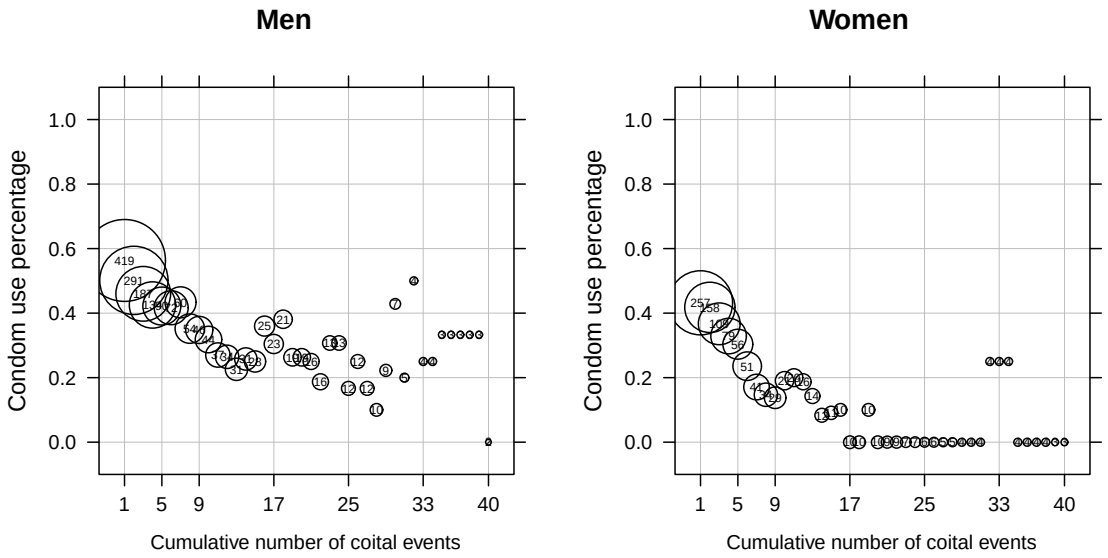


Figure 5.1: Condom use percentage as a function of the cumulative number of coital events for men (left panel) and women (right panel). The center of each circle indicates the average condom use percentage for all 676 intervals of sex with a new partner. The radius of each circle reflects the numbers of intervals included in each ordered coitus event.

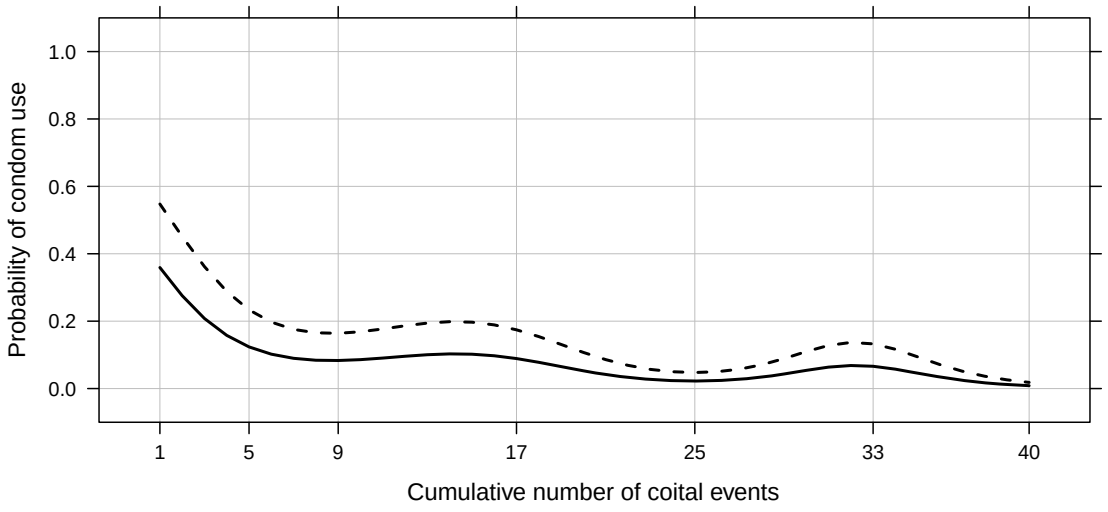


Figure 5.2: Estimated condom use probability trajectory (solid curve) for women and (thick dash curve) men with high level of relationship satisfaction based on the multivariable GAMM.

# Chapter 6

## Conclusions

In this dissertation, we present two new developed statistical models and also extend two existing statistical models to deal with data from the IU Phone Study, which is an EMA study with complex dependence and sampling data structures related to the risk of STIs and sexual or non-sexual behaviors.

In the first paper, the proposed autoregressive and cross-lagged model for bivariate non-commensurate outcomes can be applied to the scenarios that are trying to understand the cross-dependence between two correlated non-commensurate longitudinal outcomes and auto-dependence within each outcome, which extends the commonly used autoregressive and cross-lagged models. By introducing a common subject-specific random effect to estimate the correlation between two correlated outcomes, we combine the univariate mixed model methodology with the cross-lagged models to model correlated bivariate longitudinal outcomes, which relaxes the independent error assumption in the univariate GLMs. Traditional panel models treat all autoregressive and cross-lagged effects as fixed without considering the variation among subjects. Inclusion of the subject-specific random effects in the proposed model accounts for between-subject variability arising from the omitted subject-level predictors. We include both cross-lagged and autoregressive effects in the model in order to minimize bias in the estimation of cross-lagged effects. The estimates obtained from the proposed model in the simulation studies are consistent and have smaller variability than the estimated obtained from the ordinary GLMs. In the real data application, we employ the proposed model to the EMA longitudinal dataset. We are able to depict the timing and sequencing link of condom use and sexual satisfaction. We find negative cross-lagged effect

between these two outcomes and positive autoregressive effect within each outcome. Current model is developed for one continuous outcome and one binary outcome; however, this likelihood-based approach can be applied to outcomes with different measurement types. The proposed model also does not require the data to have a balanced structure and can be used when subjects contribute varying numbers of events. However, we require the inter-observation time intervals to be the same. Till now, all the autoregressive and cross-lagged models required the outcomes share the same time lag across subjects. This assumption might not be hold in the practical life. In the future model development, we plan to combine the VDFR method with the proposed model to deal with the sparse outcomes with different inter-event time intervals.

In the second paper, we apply the lagged time model in VDFR to the IU Phone Study by including condom use and sexual satisfaction as two functional predictors to estimate the probability of partner change. We find significant negative association between partner change and sexual satisfaction trajectory only in the middle of those partnerships with many coital events but there is no significant association between partner change and condom use trajectory. It is important to recognize that the models that we fit are not causal models, and we do not employ them to try to identify a causal relationship between the covariate function and outcome. The advantages of introducing VDFR models here are that we do not ignore the functional nature of the data by throwing away much of the available information and it is specifically designed for data with large between-subject variability in the width of domain or when the original time domain is informative. From IU Phone data, we find an interesting connection between EMA data and functional data. Though those VDFR models are developed from the functional data analysis first, our application shows it is appropriate and convenient to use VDFR models in EMA data. We also extend VDFR application from one functional predictor to two functional predictors in order to

use more information and combined them together to make a better prediction. In our analysis, we use imputation to deal with missingness of the condom use in IU Phone Study. The systematic sparseness of sexual satisfaction information when comparing to relationship satisfaction is something new in VDFR models. In the future, we are interested in extending the VDFR method to the case with sparse or unevenly sampled functional covariates.

In the third paper, the proposed models provide the adjusted probability of true status based on different scenarios of covariates in both cross-sectional data and longitudinal data with autoregressive (AR)(1) model. The estimates from those models have smaller bias than the MC-SIMEX models and naive models for cross-sectional data and much smaller bias than the naive model for longitudinal data. Our method can be applied to different clinical studies with imperfect diagnosis test for two main different purposes. At the population level, more accurate covariate specific probability of positivity can be estimated with correction of the misclassification, which is helpful in public health intervention. At the individual level, before the subject taking any test samples or doing any test, she/he can get an estimates of the probability of disease based on his/her information of test history and other covariates in the model in advance. This would be really helpful for the people who do not get the chance to visit the clinics for test but concern about their health status. In the real data application, we implement a sensitivity analysis to provide possible range of estimated probability of positivity based on different combination of sensitivity and specificity, which also provides a tool for analysis when there is limited amount of information about sensitivity and specificity of the test. Based on the way we are building the joint probability in EM part, our AR(1) model can be extended to the case with value of sensitivity and specificity changing with time. In our proposed models, extra covariates with correct measurement can be added into the model with no cost. If we are trying to add more misclassified binary covariates simultaneously, the dimension of steps in calculating

the joint probability will increase geometrically. But it is still within reasonable range when the number of covariate is less than 50. Similarly, we allow more historical test information (less than 50) involved in our autoregressive model setting as the AR(k) model.

The forth paper is an example of employing GAMMs to the EMA data with hierarchical structure by including nested random effects. Analyses utilizing GAMMs to show that the likelihood of condom use declines sharply for both men and women after the early accrual experience with a partner and the smooth shapes of estimated condom use probabilities are similar for both sexes. Relatively higher condom use percentage was followed by a sharp decline during the first 9 coital events decreasing to 16% for men and 8% for women. Relationship satisfaction and sexual satisfaction also influence declines in condom use, especially among women. More rapid decline in condom use among women is highly associated with higher levels of relationship and sexual satisfaction.

The contribution of our work is to bridge the algorithms from different areas with the special EMA data structure of IU Phone Study and also to build up a novel understanding of the association among all the variables of interest from different perspectives based on the characteristic of the data. Besides the statistical methodologies developed in this dissertation, variety of plots included for data visualization also provide informative support in clearly presenting complicated EMA data structure.

# BIBLIOGRAPHY


Anderson, R. (1991). The transmission dynamics of sexually transmitted diseases: the behavioral component. *Research Issues in Human Behavior and Sexually Transmitted Diseases in the AIDS Era, Wasserheit JN, Aral SO, Holmes KK (eds) American Society for Microbiology: Washington DC*, 38–60.

Bauman, L. and R. Berman (2005). Adolescent relationships and condom use: Trust, love and commitment. *AIDS and Behavior 9*, 211–222.

Bohrnstedt, G. (1969). Observations on the measurement of change. *Sociological methodology 1*, 113–133.

Bollen, K. and P. Curran (2004). Autoregressive latent trajectory (alt) models a synthesis of two traditions. *Sociological Methods & Research 32*, 336–383.

Bolton, M., A. McKay, and M. Schneider (2010). Relational influences on condom use discontinuation: A qualitative study of young adult women in dating relationships. *Can J Hum Sexual 19*, 91–104.

Brewis, A. and M. Meyer (2005). Marital coitus across the life course. *J Biosoc Sci 37*, 499–518.

Bross, I. (1954). Misclassification in 2 by 2 tables. *Biometrics 10*, 478–486.

Cacioppo, S., F. Bianchi-Demicheli, C. Frum, J. Pfaus, and J. Lewis (2012). The common neural bases between sexual desire and love: a multilevel kernel density fmri analysis. *Sex Transm Dis 9*, 1048–1054.

Campbell, D. (1963). From description to experimentation: Interpreting trends as quasiexperiments. *Problems in measuring change*, 212–242.

Carrol, R., D. Ruppert, and L. Stefanski (1995). *Nonlinear Measurement Error Models*. New York: Chapman and Hall.

Carroll, R. and et al. (1996). Asymptotics for the simex estimator in structural measurement error models. *Journal of the American Statistical Association 91*, 242–250.

Catalano, P. and L. Ryan (1992). Bivariate latent variable models for clustered discrete and continuous outcomes. *Journal of the American Statistical Association 87*, 651–658.

Chen, Z., G. Yi, and C. Wu (2014). Marginal analysis of longitudinal ordinal data with misclassification in both response and covariates. *Biometrical Journal 56*, 69–85.

Cole, D. and S. Maxwell (2003). Testing mediational models with longitudinal data: Questions and tips in the use of structural equation modeling. *Journal of Abnormal Psychology 112*, 558–577.

Cook, J. and L. Stefanski (1994). Simulation-extrapolation estimation in parametric measurement error models. *Journal of the American Statistical Association 89*, 1314–1328.

Copeland, K. and et al. (1977). Bias due to misclassification in the estimation of relative risk. *Am J Epidemiol 105*, 488–495.

Corbett, A., M. Dickson-Gomez, J. Hilario, H. Weeks, and R. Margaret (2009). A little thing called love: condom use in high-risk primary heterosexual relationships. *Perspect Sexual Reprod Health 41*, 218–224.

Cox, D. and N. Wermuth (1992). Response models for mixed binary and quantitative variables. *Biometrika 79*, 441–461.

Cox, D. and N. Wermuth (1994). A note on the quadratic exponential binary distribution. *Biometrika 81*, 403–408.

Denes, A. (2012). Pillow talk : Exploring disclosures after sexual activity. *West J Communication 76*, 91–108.

Duncan, O. (1969). Some linear models for two-wave, two-variable panel analysis. *Psychological Bulletin 72*, 177–182.

Edwards, G., B. Barber, and S. Dziurawiec (2014). Emotional intimacy power predicts different sexual experiences for men and women. *J Sex Res 51*, 340–350.

Edwards, J. and et al. (2013). Accounting for misclassified outcomes in binary regression models using multiple imputation with internal validation data. *American Journal of Epidemiology 146*, 195–203.

Evans, M. and et al. (1996). Bayesian analysis of binary data subject to misclassification. *Bayesian Analysis in Statistics and Econometrics: Essays in Honor of Arnold Zellner, K. D.Berry, Chaloner, and J.G. eke, Editors, New York: North Holland*, 67–77.

Fortenberry, D. (2013). The evolving sexual health paradigm: transforming definitions into sexual health practices. *AIDS 27*, 127–133.

Fortenberry, D., W. Tu, J. Harezlak, B. Katz, and D. Orr (2002). Condom use as a function of time in new and established adolescent sexual relationships. *Am J Public Health 92*, 211–213.

Gaydos, C. A. and et al. (2013). Performance of the cepheid ct/ng xpert rapid pcr test for detection of chlamydia trachomatis and neisseria gonorrhoeae. *Journal of Clinical Microbiology 51*, 1666–1672.

Gellar, J. and et al. (2014). Variable-domain functional regression for modeling icu data. *Journal of the American Statistical Association 109*, 1425–1439.

Geng, Z. and C. Asano (1989). Bayesian estimation methods for categorical data with misclassification. *Communications in Statistics 8*, 2935–2954.

Goldberg, J. (1975). The effect of misclassification on the bias in the difference in two proportions and the relative odds in the fourfold table. *Journal of the American Statistical Association 70*, 561–567.

Gollob, H. and C. Reichardt (1987). Taking account of time lags in causal models. *Child Development 28*, 80–92.

Harvey, S., J. Kraft, S. West, A. Taylor, K. Pappas-Deluca, and L. Beckman (2009). Effects of a health behavior change model–based hiv/sti prevention intervention on condom use among heterosexual couples: a randomized trial. *Health Educ Behav 36*, 878–894.

Hattori, M. (2014). Trust and condom use among young adults in relationships in dar es salaam, tanzania. *J Biosoc Sci 46*, 651–668.

Heise, D. (1969). Separating reliability and stability in test-retest correlation. *American Sociological Review 34*, 93–101.

Hensel, D., D. Fortenberry, J. Harezlak, and D. Craig (2012). The feasibility of cell phone based electronic diaries for sti/hiv research. *BMC Medical Research Methodology 12*.

Joreskog, K. and D. Sorbom (1979). *Separating reliability and stability in test-retest correlation*. Abt Books.

Kessler, R. and D. Greenberg (1981). *Linear Panel Analysis*. New York: Academic Press.

Koch, G. (1969). The effect of non-sampling errors on measures of association in 2 by 2 tables. *Journal of the American Statistical Association 64*, 852–863.

Kraut-Becher, J. and S. Aral (2003). Gap length: An important factor in sexually transmitted disease transmission. *Sex Transm Dis 30*, 221–225.

Krzanowski, W. (1988). *Principles of Multivariate Analysis*. Oxford University Press.

Ku, L., F. Sonenstein, and J. Pleck (1994). The dynamics of young men's condom use during and across relationships. *Fam Plann Perspect 26*, 246–251.

Kuchenhoff, H., S. Mwalili, and E. Lesaffre (2006). A general method for dealing with misclassification in regression: the misclassification simex. *Child Development 62*, 85–96.

Kuha, J., C. Skinner, and J. Palmgren (2005). Misclassification error. *Encyclopedia of Biostatistics*.

Lane, L. and L. Viney (2002). Toward better prevention: Constructions of trust in the sexual relationships of young women. *J Appl Soc Psychol 32*, 700–718.

Lehmiller, J., L. Vanderdrift, and J. Kelly (2014). Sexual communication, satisfaction, and condom use behavior in friends with benefits and romantic partners. *J Sex Res 51*, 74–85.

Little, R. and M. Schluchter (1985). Maximum likelihood estimation for mixed continuous and categorical data with missing values. *Biometrika 72*, 497–512.

Lyles, R. and et al. (2011). Validation data-based adjustments for outcome misclassification in logistic regression: an illustration. *Epidemiology 22*, 589–597.

Magder, L. and J. Hughes (1997). Logistic regression when the outcome is measured with uncertainty. *American Journal of Epidemiology 146*, 195–203.

Manhar, L., S. Aral, K. Holmes, and B. Foxman (2002). Sex partner concurrency: Measurment, prevalence, and correlates among urban 18 - 39-year olds. *Sex Transm Dis 29*, 133–143.

Manning, W., C. Flanigan, P. Giordano, and M. Longmore (2009). Relationship dynamics and consistency of condom use among adolescents. *Perspect Sexual Reprod Health 41*, 181–190.

Matson, P., N. Adler, S. Millstein, J. Tschann, and J. Ellen (2011). Developmental changes in condom use among urban adolescent females: influence of partner context. *J Adolesc Health 48*, 386–390.

Matson, P., S. Chung, and J. Ellen (2012). When they break up and get back together: length of adolescent romantic relationships and partner concurrency. *Sex Transm Dis 39*, 281–285.

Mendoza-Blanco, J., X. Tu, and S. Iyengar (1996). Bayesian inference on prevalence using a missing-data approach with simulation-based techniques: applications to hiv screening. *Stat Med 15*, 2161–2176.

Neuhaus, J. (1999). Bias and efficiency loss due to misclassified responses in binary regression. *Biometrika 86*, 843–855.

Neuhaus, J. (2002). Analysis of clustered and longitudinal binary data subject to response misclassification. *Biometrics 58*, 675–683.

Newell, D. (1963). Errors in the interpretation of errors in epidemiology. *American Journal of Public Health 11*, 1925–1928.

Olkin, I. and R. Tate (1961). Multivariate correlation models with mixed discrete and continuous variables. *Annals of Mathematical Statistics 32*, 448–465.

Ott, M., A. Katschke, W. Tu, and D. Fortenberry (2011). Longitudinal associations among relationship factors, partner change, and sexually transmitted infection acquisition in adolescent women. *Sex Transm Dis 38*, 153–157.

Ott, M., S. Ofner, W. Tu, and D. Fortenberry (2010). Characteristics associated with sex after periods of abstinence among sexually experienced young women. *Perspect Sexual Reprod Health 42*, 43–48.

Paulino, C., P. Soares, and J. Neuhaus (2003). Binomial regression with misclassification. *Biometrics 59*, 670–675.

Pol, B. V. D. and et al. (2012). Clinical evaluation of the bd probetectm neisseria gonorrhoeae qx amplified dna assay on the bd vipertm system with xtrtm technology. *Sexually Transmisted Disease 39*, 147–153.

Pol, B. V. D. and et al. (2013). Vaginal swabs are the optimal specimen for detection of genital chlamydia trachomatis or neisseria gonorrhoeae using the cobas 4800 ct/ng test. *Sexually Transmisted Disease 40*, 247–250.

Randolph, M., S. Pinkerton, L. Bogart, H. Cecil, and P. Abramson (2007). Sexual pleasure and condom use. *Arch Sex Behav 36*, 844–848.

Rekaya, R. and D. G. K.A. Weige and (2001). Threshold model for misclassified binary responses with applications to animal breeding. *Biometrics 57*, 1123–1129.

Royer, H., M. Keller, and S. Heidrich (2009). Young adolescents' perceptions of romantic relationships and sexual activity. *Sex Educ 9*, 395–408.

Sayegh, M., D. Fortenberry, M. Shew, and D. Orr (2006). The developmental association of relationship quality, hormonal contraceptive choice and condom non-use among adolescent women. *J Adolesc Health 39*, 388–395.

Senn, T., L. Scott-Sheldon, and M. Carey (2014). Relationship-specific condom attitudes predict condom use among std clinic patients with both primary and non-primary partners. *AIDS Behav 18*, 1420–1427.

Shafii, T., K. Stovel, and K. Holmes (2007). Association between condom use at sexual debut and subsequent sexual trajectories: a longitudinal study using biomarkers. *Am J Pub Health 97*, 1090–1095.

Shew, M., D. Fortenberry, W. Tu, B. Juliar, B. Batteiger, B. Qadadri, and D. Brown (2006). Association of condom use, sexual behaviors, and sexually transmitted infections with the duration of genital human papillomavirus infection among adolescent women. *Arch Pediatr Adolesc Med 160*, 151–156.

T. Ein-Dor, G. H. (2012). Sexual healing: Daily diary evidence that sex relieves stress for men and women in satisfying relationships. *J Soc Pers Relationships 29*, 126–139.

Tate, R. (1954). Correlation between a discrete and a continuous variable. *Annals of Mathematical Statistics 25*, 603–607.

Taylor, S. and et al. (2012). Evaluation of the roche cobas ct/ng test for detection of chlamydia trachomatis and neisseria gonorrhoeae in male urine. *Sexually Transmisted Disease 39*, 543–549.

Teixeira-Pinto, A. and S. Normand (2009). Correlated bivariate continuous and binary outcomes: issues and applications. *Statistics in Medicine 28*, 1753–1773.

Thompson, A. and L. O'Sullivan (2012). Gender differences in associations of sexual and romantic stimuli: do young men really prefer sex over romance? *Arch Sex Behav 41*, 949–957.

Thurigen, D. and et al. (2000). Measurement error correction using validation data: a review of methods and their applicability in case-control studies. *Stat Methods Med Res 9*, 447–474.

Willig, C. (1997). The limitations of trust in intimate relationships: constructions of trust and sexual risk taking. *Br J Soc Psychol 36*, 211–221.

Zukoski, A., S. Harvey, and M. Branch (2009). Condom use: exploring verbal and non-

verbal communication strategies among latino and african american men and women. *AIDS Care 21*, 1042–1049.

# CURRICULUM VITAE

Fei He

## EDUCATION

- Ph.D. in Biostatistics, Indiana University, Indianapolis, IN, 2016 (minor in Epidemiology)
- M.S. in Applied Statistics, Purdue University, Indianapolis, IN 2013
- B.S. in International Economics and Business, Xiamen University, China 2008

## WORKING EXPERIENCE

- Research Scientist, June 2016, Washington University
- Instructor, July 2015 - Dec 2015, Indiana University, Indianapolis
- Research Assistant, July 2014 - June 2015, Indiana University, Indianapolis
- Teaching Assistant, July 2013 - June 2014, Indiana University, Indianapolis
- Research Assistant, July 2011 - June 2013, Indiana University, Indianapolis

## SELECT PUBLICATIONS

- Fei He, Jaroslaw Harezlak, Devon Hensel, J. Dennis Fortenberry (2016). Condom use as a function of number of coital events in new relationships. Sexual Transmitted Diseases 43(2), 67-70.
- Fei He (2009). The analysis and equilibrium of multidimensional game model of merchant's marketing strategies under the ban for free plastic bags. The International Conference on Computational Intelligence and Software Engineering, China.