Measurement Error and Non-Random Sampling: Addressing Biases when Studying the Association between Behavioural Characteristics and HIV Phylogenetic Cluster Size

Nabila Parveen

Doctor of Philosophy

Department of Epidemiology, Biostatistics and Occupational Health

McGill University

Montreal, Quebec, Canada

July 24, 2017

A thesis submitted to the Faculty of Graduate Studies and Research in partial fulfillment of the requirements of the degree of Doctor of Philosophy

DEDICATION

To my Parents

ACKNOWLEDGEMENTS

First and foremost, I would like to express my heartfelt gratitude to my supervisor Dr. Erica E. M. Moodie for her continuous support, continuous encouragement, countless thoughtful discussions, patience and guidance during the preparation of this dissertation, for being an excellent teacher and mentor.

I offer my very special thanks to Dr. David Stephens to become a member of my thesis supervisory committee. I also, thank Dr. Bluma Brenner and Dr. Joseph Cox for providing the SPOT and ARGUS data and for their informative feedback on the manuscripts.

The Department of Epidemiology, Biostatistics and Occupational Health, McGill University, has always been a fantastic place. My sincere thanks go to the faculty, graduate students and staff for making an environment that has helped me to finish this research.

I would like to acknowledge the financial support provided by Dr. Erica Moodie from her research grants. I am also thankful for additional funding from the 'McGill Internal Studentships'.

My sincere thanks are for my dear husband, Muhammad Abu Shadeque Mullah, who has been a constant source of strength. He read and edited all my manuscripts, always supported me in the difficult times, and motivated me whenever I needed it. Without his support it would not have been possible to complete my PhD.

My parents, sister, uncle Md Nurul Azam Khan and aunt Jahanara Khan deserve special thanks for their continued support and encouragement they have given me over the years. I also thank my in-laws and all my friends, especially Khan Jahan who motivated me to complete my degree.

Finally, I want to thank my two lovely boys, Safwan and Saifan, for giving me inspiration to stay focused and reminding me how beautiful really life is, even with so many ups and downs.

ABSTRACT

Human immunodeficiency virus (HIV) is an infectious disease that has led to terrible losses since it was first identified over three decades ago. Although current therapies for HIV are highly effective and have dramatically reduced mortality, HIV places an immense burden on individuals as well as on society: annual costs of new HIV infections in the United States were estimated at 36.4 billion (in 2002). First identified in men who have sex with men (MSM), this population continues to be disproportionately affected by the disease.

Montreal is one of the main centres for HIV research activity in Canada. In particular, SPOT is a key study, focusing on MSM. SPOT offers rapid, free and anonymous testing to the MSM community. SPOT also collects data on socio-demographic and behavioural characteristics. This thesis is based on analysis of the SPOT data supplemented with HIV RNA sequencing information from the Quebec Genotyping Program Cohort and the Primary Infection Cohort, which provide information that informs the size of the phylogenetic cluster to which HIV-positive individuals belong. Large clusters are indicative of rapid HIV transmission. HIV researchers wish to determine the behavioural correlates of phylogenetic cluster size, as understanding the determinants of larger clusters may suggest ways to target interventions to break transmission chains. However, there is a significant number of people living with HIV in Quebec whose HIV RNA has not been sequenced. As a result, measurement error occurs in defining the cluster size in SPOT. Moreover, the measurement error in the cluster size is not mean zero, but rather exhibits systematic under-counting of the true cluster size and it is not possible to obtain a validation sample that would reveal the true cluster size for some SPOT participants. An additional challenge faced in SPOT is that the recruitment method is not based on a probability sampling technique, and as such, findings from the SPOT study may suffer from lack of generalizability. Thus, to make valid statistical inference about correlates of phylogenetic cluster size using the SPOT data, both measurement error and the sampling scheme must be taken into consideration. In this thesis, I propose and validate statistical

methods to address both of these issues.

There are several approaches to deal with measurement error, however most require validation or repeated sampling data. The SPOT study does not have a validation sample and indeed such a validation study would not be feasible in the context of phylogenetic cluster size. One measurement error approach, the simulation-extrapolation (SIMEX) method does not require such validation data, and thus represents a promising avenue for correction of the type of measurement error exhibited in the SPOT cluster size data. However, its development to date has been limited to mean zero random errors.

In the first manuscript, I extend the SIMEX method to non-zero mean (NZM) measurement error which better mimics the SPOT data for the HIV-infected participants. I provide a theoretical justification for the extension of the SIMEX by modifying the measurement error model in such a way that the observed cluster size will always be less than or equal to the true cluster size. The simulation step of the SIMEX approach is also modified. Simulation studies show that NZM-SIMEX considerably reduces the bias as compared to a naïve analysis that ignores measurement error. I then apply the NZM-SIMEX to the HIV-positive participants in SPOT to examine correlates of phylogenetic cluster size.

The proposed NZM-SIMEX is applicable only to the data from HIV-positive MSM whose cluster size is undercounted. However, SPOT study collects data not only from HIV-positive MSM but also from a large number of HIV-negative MSM. For HIV-negative MSM, the cluster size is zero and is not subject to any measurement error. So, the measurement error in phylogenetic cluster size depends on the HIV status. To include the data from both HIV-positive and HIV-negative MSM in an analysis, in the second manuscript, I further extend the NZM-SIMEX to the settings where the measurement error in a covariate of interest depends on the value of another correctly specified covariate. This SIMEX conditional on covariates (SIMEX-CC) performs well in simulation, typically exhibiting less biased and variability than other measurement error approaches.

Finally, in the third manuscript, I simultaneously adjust for both the non-probabilistic sampling scheme in SPOT as well as the measurement error in cluster size. Using another sample of MSM from Montreal that employed a probabilistic sampling scheme, a survey weighting approach is employed in the SPOT analysis. This analysis suggests that accounting for the recruitment (or sampling) scheme in SPOT has notable impact on the results.

Thus, in my thesis, I have developed a measurement error correction approach that can accommodate systematic under-counting without validation data, where the measurement error distribution may depend on another covariate measured in the sample. Further, I have demonstrated how external data may be leveraged to improve generalizability in a study whose sampling mechanism was not probabilistic.

ABRÉGÉ

L'infection par le virus d'immunodéficience humaine (VIH) est une maladie infectieuse qui a induit des pertes terribles depuis sa découverte, il y a un peu plus de trois décennies. Bien que les traitements actuels sont très efficaces et ont drastiquement réduit la mortalité, le VIH constitue un fardeau considérable pour les patients comme pour la société : la facture annuelle liée aux nouvelles infections par le VIH aux États-Unis est estimée à 36.4 milliards de dollars en 2002. Le virus a d'abord été identifié chez des hommes ayant des relations sexuelles avec d'autres hommes (HSH), et cette population continue d'être affectée par la maladie de manière disproportionnée.

Montréal est l'un des principaux centres pour la recherche portant sur le VIH au Canada. En particulier, SPOT est une étude clef, qui se focalise sur les HSH. SPOT offre un dépistage rapide, gratuit et anonyme à la communauté HSH. SPOT recueille également données sociodémographiques et caractéristiques comportementales. Cette thèse est basée sur l'analyse des données de SPOT, agrémentées de données de séquençge ARN du VIH provenant de la cohorte du programme de génotypage du Québec et de la cohorte de primo-infection, qui renseignent sur les tailles des groupes phylogénétiques auxquels appartiennent les sujets séropositifs: large groupe signifie transmission rapide du VIH. Les chercheurs dans le domaine du VIH souhaitent déterminer les corrélats comportementaux de la taille du groupe phylogénétique, puisque la compréhension des déterminants de certains groupes plus grands pourrait suggérer des manières de cibler des interventions permettant de briser les chaînes de transmission. Cependant, il existe au Québec un nombre significatif de personnes qui vivent avec un VIH dont l'ARN n'a pas été séquencé: par conséquent, des erreurs de mesures se produisent lors de la définition de la taille du groupe dans SPOT. De plus, l'erreur de mesure dans la taille du groupe n'est pas de moyenne zéro, mais démontre plutôt une sous-évaluation de la taille réelle du groupe, et il n'est pas possible d'obtenir un échantillon de validation permettant de révéler la taille réelle du groupe pour certains participants de SPOT. Un

défi supplémentaire rencontré avec SPOT réside dans le fait que la méthode de recrutement ne se base pas sur une technique d'échantillonnage probabiliste, et de ce fait, les résultats provenant de l'étude SPOT pourraient pâtir d'un manque de généralisabilité. Par conséquent, pour émettre une inférence statistique valide à propos des corrélats de la taille du groupe phylogénétique en se basant sur les données de SPOT, aussi bien les erreurs de mesure que la stratégie d'échantillonnage doivent être pris en considération. Dans cette thèse, je propose et valide des méthodes statistiques pour gérer ces deux problématiques.

Il existe plusieurs approches pour traiter les erreurs de mesure, mais la plupart nécessitent une validation des données ou l'échantillonnage répété. L'étude SPOT ne dispose pas d'un échantillon de validation, et ce type de validation ne pourrait en effet pas être faisable dans un contexte de taille de groupe phylogénétique. La méthode de simulation-extrapolation (SIMEX), une autre approche pour la correction des erreurs de mesure, ne nécessite pas de données de validation, et représente donc une voie prometteuse pour la correction du type d'erreurs de mesure retrouvé dans les données de taille de SPOT. À ce jour, son développement a cependant été restreint aux erreurs aléatoires de moyenne zéro.

Dans mon premier manuscrit, j'étends la méthode SIMEX à la mesure d'erreurs de moyenne non-zéro (MNZ), qui reproduit mieux les données de SPOT pour les sujets séropositifs. Je justifie cette extension de SIMEX par la modification du modèle de mesure d'erreurs de façon à ce que la taille du groupe observé soit toujours inférieure ou égale à la taille réelle de ce groupe. L'étape de simulation de l'approche SIMEX est également modifiée. Les études de simulation démontrent que MZN-SIMEX réduit considérablement le biais, en comparaison avec une analyse naïve ne tenant pas compte des erreurs de mesure. J'applique par la suite MZN-SIMEX aux participants séropositifs de SPOT pour examiner les corrélats de la taille de groupe phylogénétique.

La méthode MZN-SIMEX proposée n'est applicable qu'aux données provenant de séropositifs HSH dont la taille de groupe est sous-évaluée. Cependant, les études SPOT répertorient des données provenant non-seulement d'HSH séropositifs, mais aussi d'un grand nombre d'HSH

séronégatifs. Pour ces derniers, la taille de groupe est zéro, et n'est sujette à aucune erreur de mesure : l'erreur de mesure dans la taille de groupe phylogénétique dépend donc de la sérologie VIH. Dans le second manuscrit, et pour inclure dans l'analyse les données provenant des HSH séropositifs et séronégatifs, j'étends l'approche MZN-SIMEX aux situations où l'erreur de mesure dans une covariable d'intérêt dépend de la valeur d'une autre covariable correctement spécifiée. Cette approche SIMEX conditionnelle aux covariables (SIMEX-CC) donne de bons résultats en simulation, montrant habituellement moins de biais et de variabilité comparativement à d'autres approches de mesure d'erreurs.

Finalement, dans le troisième manuscrit, j'ajuste simultanément pour la stratégie d'échantillonnage non-probabiliste dans SPOT, et pour la mesure d'erreurs dans la taille de groupe. En utilisant un autre échantillon d'HSH (de Montréal) qui a utilisé une approche d'échantillonnage probabiliste, une approche de pondération est employée dans l'analyse de SPOT. Cette analyse suggère que la stratégie de comptabilisation pour le recrutement (ou échantillonnage) dans SPOT a un impact notable sur les résultats.

En résumé, j'ai donc développé dans ma thèseune approche de correction des erreurs de mesure qui peut s'adapter aux sous-évaluations systématiques sans validation de données, quand la distribution des erreurs de mesure peut dépendre d'autres covariables mesurées dans l'échantillon. Par ailleurs, j'ai aussi démontré comment les données externes peuvent être mises à profit pour améliorer la généralisabilité dans une étude où la stratégie d'échantillonnage n'est pas probabiliste.

PREFACE

Format of the thesis

This is a manuscript-based thesis that is formatted following the McGill University Guidelines for the Thesis Preparation. It consists of a series of three research manuscripts each of which corresponds to a chapter.

There are a total of seven chapters. The first chapter includes the introduction that explains the rationale of the study. A comprehensive literature review has been carried out in Chapter 2. All objectives are stated in Chapter 3. Chapters 4-6 contain three research manuscripts that are linked, each building on the developments of the previous chapter with the aim of addressing a specific gap in the literature motivated by a particular question and applied to a single dataset. Combined these chapters form a consistent unit that addresses the main objectives of this dissertation research. A preamble to each manuscript elucidates its rationale and its connection to the other manuscripts as well as to the overall objectives of the thesis. A summary of the contributions of this thesis and point to future directions for research are presented in Chapter 7. Finally, all references included in different chapters are combined into an overall 'Bibliography', at the end of the thesis.

Contribution of authors

This thesis is based on innovative ideas that were selected, developed and finalized in a series of discussion with my supervisor Dr. Erica Moodie. Fully guided by Dr. Moodie, I determined the overall scope of my thesis and selected the specific objectives and methods.

I conducted the literature review, developed analytical strategies in collaboration with my supervisor, designed and carried out simulation studies. I also performed data analyses and wrote all three manuscripts along with the other chapters of the thesis. Dr. Moodie provided guidance and feedback on the methods, simulation studies, data analyses, interpretation of results, and consecutive drafts of the manuscripts and other chapters. Dr. Bluma Brenner provided the access to the SPOT data that are analyzed in all three manuscripts. She also offered informative feedback on all three manuscripts. Dr. Joseph Cox provided

access to the ARGUS data that are used in the third manuscript. Moreover, he provided feedback on the third manuscript.

As a PhD candidate, I am fully responsible for the scientific quality of the research, originality of the ideas and truthfulness of the results contained in this thesis.

Statement of originality

The research in this thesis constitutes original scholarship and advances knowledge in the domain of statistical methodology for correcting measurement error in covariates and adjusting for a non-probabilistic sampling scheme.

The simulation-extrapolation (SIMEX) method for measurement error correction is well developed, but not for settings where measurement error occurs due to systematic *under-counting*. In such cases measurement error distribution does not have mean zero. Moreover, error distribution may depend on other correctly measured covariates. Further, improving the generalizability of results from a study which has non-probabilistic sampling mechanism is challenging, especially when no internal information is available.

This thesis addresses all of the issues outlined above. The main new contributions in this thesis are contained in three manuscripts presented in Chapters 4, 5 and 6. In the first manuscript (Chapter 4), I extended and validated the SIMEX to the non-zero mean measurement errors and called this non-zero mean SIMEX (NZM-SIMEX). In the second manuscript (Chapter 5), I further extended the NZM-SIMEX to the settings where measurement error in a covariate of interest depends on the value of another correctly specified covariate, and called this SIMEX conditional on covariates (SIMEX-CC).

In the third manuscript (Chapter 6), I proposed a novel approach that uses an external source of information to predict sampling weights that can be used to improve the generalizability of a study whose recruitment mechanism was non-probabilistic. The estimated sampling weights were then used to fit weight adjustment model to improve the generalizability of the results.

All the proposed methods were applied to the data from SPOT study (in each manuscript separately) to reveal the (lack of) association of HIV phylogenetic cluster size with demographic and sexual behavioural characteristics of men who have sex with men in Montreal.

The developed methods provide new insight into correcting measurement error in covariates, especially in HIV studies, where interest lies in correlating HIV phylogenetic clustering data with epidemiological data. To the best of my knowledge, the ideas, methods, simulation plans and data analysis strategies considered in this research were not adopted or published in any previous publication.

TABLE OF CONTENTS

DEDICATION ii

ACKNOWLEDGEMENTS iii

ABSTRACT iv

ABRÉGÉ vii

PREFACE x

LIST OF TABLES xvi

LIST OF FIGURES xviii

LIST OF ABBREVIATIONS xix

1 Introduction 1

2 Literature Review 7

2.1 Introduction 7
2.2 HIV and AIDS 7
2.3 Phylogenetic Clustering: Offering New Insights into Viral Epidemics . . . 8
2.4 HIV Research in Montreal 9
2.4.1 The SPOT Study 10
2.4.2 The ARGUS Study 12
2.5 Measurement Error 12
2.5.1 Differential Measurement Error 13
2.5.2 Non-differential Measurement Error 14
2.5.2.1 Classical Additive Measurement Error 14
2.5.2.2 Berkson Measurement Error 15
2.5.3 Effects of Measurement Error 15
2.5.3.1 Measurement Error in a Predictor Variable 15
2.5.3.2 Measurement Error in the Outcome Variable 18
2.6 Measurement Error Correction Methods 19
2.6.1 Sources of Data 19
2.6.2 Functional and Structural Approaches to Measurement Error . . . 20
2.6.3 Regression Calibration 21
2.6.4 Simulation-Extrapolation (SIMEX) 22
2.6.5 Multiple Imputation 25
2.6.6 Likelihood Method 25

2.6.7 Bayesian Method . . . 26
2.6.8 Method of Moments . . . 27
2.7 Sampling Techniques for Recruiting Hard-to-Reach Populations . . . 27
2.7.1 Snowball Sampling . . . 28
2.7.2 Targeted Sampling . . . 28
2.7.3 Respondent-Driven Sampling . . . 29
2.7.4 Venue-Based Sampling . . . 30
2.8 Summary . . . 32

3 Objectives . . . 33

4 Manuscript I: The Non-Zero Mean SIMEX: Improving Estimation in the Face of Measurement Error . . . 35

4.1 Introduction . . . 38
4.2 The Simulation-Extrapolation (SIMEX) Method . . . 40
4.3 Simulation Study . . . 44
4.3.1 Design of the Simulation Study . . . 44
4.3.2 Results of the Simulation Study . . . 45
4.4 SPOT Analysis . . . 46
4.4.1 Measurement Error Cluster Size . . . 49
4.4.2 Results . . . 50
4.4.3 Limitations and Discussion of the Analysis . . . 51
4.5 Discussion . . . 55
4.6 Appendix for Manuscript I . . . 58
4.6.1 Appendix A: Proofs . . . 58
4.6.1.1 Proof of Theorem 1 . . . 58
4.6.1.2 Proof of Theorem 2 . . . 61
4.6.2 Appendix B: Details of Simulation Study . . . 64
4.6.2.1 Design of the Simulation Study . . . 64
4.6.2.2 Simulation Results . . . 64
4.6.3 Appendix C: Additional Results . . . 72

5 Manuscript II: Correcting Covariate-Dependent Measurement Error with Non-Zero Mean . . . 76

5.1 Introduction . . . 79
5.2 Measurement Error Correction . . . 80
5.2.1 Two Common Approaches: Regression Calibration and Multiple Imputation for Measurement Error . . . 81
5.2.2 Proposed Approach: SIMEX Conditional on Covariates . . . 82
5.3 Simulation Study . . . 86
5.3.1 Design of the Simulation Study . . . 86
5.3.2 Analysis of the Simulated Data . . . 87
5.3.3 Results of the Simulation Study . . . 89
5.4 Analysis of the SPOT Data . . . 91

5.5 Discussion . . . 93
5.6 Appendix for Manuscript II . . . 96
5.6.1 Appendix D: Proof of Theorem . . . 96
5.6.2 Appendix E: Additional Results . . . 99
5.6.2.1 Additional Numerical Results . . . 99
5.6.2.2 Additional Details on the SPOT Analysis . . . 102

6 Manuscript III: New Challenges in HIV Research: Combining Phylogenetic Cluster Size and Epidemiologic Data . . . 103

6.1 Introduction . . . 106
6.2 Methods . . . 107
6.2.1 Data Sources . . . 107
6.2.1.1 The SPOT Study . . . 107
6.2.1.2 The ARGUS Study . . . 109
6.2.2 Statistical Methods . . . 109
6.2.2.1 Venue-Based Sampling . . . 109
6.2.2.2 Simulation-Extrapolation Conditional on Covariates . . . 111
6.3 The SPOT Analysis . . . 113
6.4 Discussion . . . 117
6.5 Appendix for Manuscript III . . . 120

7 Discussion . . . 124

Bibliography . . . 133

LIST OF TABLES

Table page

4–1 Characteristics of HIV-positive MSM in SPOT study 52

4–2 Simulation scenarios 64

4–3 Simulation results for a continuous outcome and a correctly specified error distribution 65

4–3 (cont.) Simulation results for a continuous outcome and a correctly specified measurement error distribution 66

4–3 (cont.) Simulation results for a continuous outcome and a correctly specified measurement error distribution 67

4–4 Simulation results for a continuous outcome and a mis-specified measurement error distribution 68

4–4 (cont.) Simulation results for a continuous outcome and a mis-specified measurement error distribution 69

4–5 Simulation results for a Poisson outcome and a correctly specified measurement error distribution 70

4–6 Simulation results for a Bernoulli outcome and a correctly specified measurement error distribution 71

4–7 Results from the simple linear regression model 72

4–8 Results from the simple logistic regression model 73

4–9 Results from the log-linear model of number of sex partners on cluster size . . 74

4–10 Results from the multinomial model of number of one night partners on cluster size 75

5–1 Simulation study design 87

5–2 Characteristics of the SPOT participants 92

5–3 Naïve regression and SIMEX-CC to assess the relationship between phylogenetic cluster size and number of sex partners 94

5–4 Simulation results for normally distributed outcomes when conditioning variable Z is common ($P(Z = 1) = 0.5$) or rare ($P(Z = 1) = 0.05$) 100

5–5 Simulation results for Poisson distributed outcomes when the conditioning variable Z is common ($P(Z = 1) = 0.5$) or rare ($P(Z = 1) = 0.05$) 101

6–1 Characteristics of the ARGUS and SPOT participants 115

6–2 Results from weighted and unweighted naïve and SIMEX-CC methods 118

6–3 SPOT participant characteristics 120

6–4 Results from the naïve and SIMEX-CC when measurement error is assumed to be distributed as Poisson(5) 121

6–5 Results from the naïve and SIMEX-CC methods after imputing missing data 122

6–6 Results from the naïve and SIMEX-CC methods, predicting sampling weights from a linear model instead of multinomial regression model 123

LIST OF FIGURES

Figure page

4–1 A generic plot explaining the SIMEX method 43

4–2 Bias and MSE of the parameter estimator as a function of measurement error variance 47

4–3 Bias and MSE of the parameter estimator for two different measurement error distributions 48

4–4 SIMEX estimate and 95% confidence interval for the effect of interest 52

5–1 Comparison of measurement error correction methods in terms of the bias and rMSE 90

5–2 Bias in the naïve estimates as the proportion of the covariate subject to measurement error increases 99

LIST OF ABBREVIATIONS

AIDS	Acquired Immune Deficiency Syndrome
ART	Antiretroviral Therapy
CI	Confidence Interval
CP	Coverage Probability
DHS	Demographic and Health Survey
DNA	Deoxyribonucleic Acid
GLMM	Generalized Linear Mixed Model
HIV	Human Immunodeficiency Virus
IQ	Intelligence Quotient
MCMC	Markov Chain Monte Carlo
MI	Multiple Imputation
MIME	Multiple Imputation for Measurement Error
MSE	Mean Squared Error
MSM	Men who have Sex with Men
NZM-SIMEX	Non-zero Mean Simulation-Extrapolation
OLS	Ordinary Least Squares
RC	Regression Calibration
rMSE	Square Root of Mean Squared Error
RNA	Ribonucleic Acid
SE	Standard Error
SIMEX	Simulation-Extrapolation
SIMEX-CC	SIMEX Conditional on Covariates
SIMEX-NL	SIMEX with Non-linear Extrapolation

SIMEX-Q	SIMEX with Quadratic Extrapolation
TLS	Time-Location Sampling
VDT	Venue-Day-Time

Chapter 1
Introduction

Human immunodeficiency virus (HIV) is an incurable infection which, if left untreated, can lead to acquired immune immunodeficiency syndrome (AIDS), and a destruction of the body's immune defense system against all invading pathogens [1]. HIV is transmitted primarily through unprotected sexual intercourse, contaminated blood transfusions and hypodermic needles, as well as from mother to child during pregnancy, delivery, or breastfeeding [2]. HIV/AIDS was first documented in 1981 among a cluster of injection drug users and homosexual men [3–5]. While the illness is now observed in both homosexual and heterosexual men and women, men who have sex with men (MSM) continue to be the most affected group in Canada [6]. Of all new HIV infections in Canada, 57% are MSM and they are 131 times more likely to get HIV than men who do not have sex with men [7]. While there is no cure or vaccine for HIV, current therapies are highly effective and have dramatically reduced mortality due to HIV. Nevertheless, HIV places an immense burden on individuals and on society; annual costs of new HIV infections in the United States were estimated at 36.4 billion (2002) [8].

Early in the HIV epidemic, research focused on identifying the cause of AIDS and developing treatments as well as strategies to prevent new infections. Treatments have improved now to such a degree that HIV in the Western world is often viewed as a chronic condition requiring long-term management, similar to diabetes or hypertension. Nevertheless, the infectious nature of the illness makes it unique from other chronic illnesses, and potentially more easily preventable. In the last decade, with the advance of Ribonucleic Acid (RNA) and Deoxyribonucleic Acid (DNA) sequencing technology, there is considerable research activity centred on using molecular phylogenetics to understand the social and behavioural drivers of HIV transmission. By comparing viral genomic sequences, relatedness of viruses can be

determined, and from that, clusters of individuals with similar viruses may be inferred [9–17]. While transmission between specific individuals in these clusters cannot be determined, the clustering provides evidence of common contacts among the members of the cluster (primarily through sexual contact or shared needles). This has become the focus of much of the HIV research activity in Montreal.

SPOT is a study of MSM which offers rapid, free, and anonymous HIV testing and administers an anonymous questionnaire that elicits information on socio-demographic characteristics, HIV testing behaviour, sex life, attitude towards HIV, and socio-sexual profile. In addition to providing questionnaire data, all HIV-positive participants' blood undergoes HIV RNA sequencing so that phylogenetic clustering could be used to determine the size of the cluster to which the HIV RNA sequence belongs [13, 18, 19]. This thesis is based on the SPOT data supplemented with RNA and DNA sequencing information from the Quebec Genotyping Program Cohort and the Primary Infection Cohort to determine the size of the phylogenetic cluster to which HIV-positive individuals belong. Phylogenetic cluster size is defined as the number of individuals falling into the same HIV phylogenetic grouping. For example, if the HIV sequence of seven individuals fall into the same cluster, each will be belong to a cluster size seven. On the other hand, if there is an individual whose HIV genome sequence does not cluster with the HIV genome of anyone else in the Quebec Genotyping Program registry of sequences, this individual is said to belong to a cluster of size one.

The SPOT study does not include individuals who are HIV-positive but are unaware of their status (i.e. have never been tested) or those who have not had their HIV genotyped within the province of Quebec (e.g. those who are aware of their HIV status but may be have recently moved from another province or country). Consequently, measurement error occurs in defining the phylogenetic cluster size. This measurement error is characterized by a systematic *undercounting* of the true cluster size due to the absence of the individuals who have not been tested. Thus, to make valid statistical inference about correlates of sexual network size, this measurement error must be taken into consideration.

There are several methods, such as regression calibration, multiple imputation and simulation-extrapolation (SIMEX) that have been proposed to deal with measurement error. Regression calibration is the most popular approach in practice because it is simple to apply and generally performs well [20]. Multiple imputation is also an interesting approach to adjustment for measurement errors, where measurement errors are treated as a missing data problem (see, for example, Rubin [21], Cole et al. [22]; Padilla et al. [23]; Messer and Natarajan [20]). Most of the methods, including regression calibration and multiple imputation, however, require validation or replicate data for some fraction of the observed sample. In the context of undercounting of phylogenetic cluster size (due to unobserved/untested individuals) in the SPOT data, obtaining a validation sample is both ethically and practically unfeasible as it would require the testing of all residents of the province of Quebec.

The SIMEX method of Cook and Stefanski [68] is a simulation based technique for estimating and reducing bias due to additive measurement error. It does not require validation data, but does require that the measurement error distribution is known or can be well-estimated. SIMEX estimates are obtained by adding additional measurement error to the mismeasured data in a resampling-like stage, computing estimates from the deliberately contaminated data, establishing a trend between these estimates and the variance of the added measurement errors, and extrapolating this method back to the case of no measurement error. SIMEX is becoming popular as a general and widely applicable functional method because it does not require any assumptions about the distribution of the unobserved true covariate [24]. However, to date, SIMEX has been limited to mean zero random errors, and hence needs to be extended to alternative error distributions (in particular, one that has non zero mean) to be used in the context of undercounted measures in the cluster size of the SPOT data.

MSM populations have, historically, sometimes been viewed as "hidden" due to discrimination against homosexuality. It can therefore be challenging to obtain a random sample from this population, or indeed any population that has been stigmatized. Specifically, in

the context of HIV, it is also very difficult to obtain a large sample where both epidemiological and HIV RNA sequencing data are available. Traditional sampling methods may not be suitable to obtain a random sample from MSM population. Special techniques such as snow-ball sampling [25], targeted sampling [26], time-location sampling [27] and venue-based sampling [28, 29] are often used to sample from hidden/rare populations. Venue-based sampling is the most commonly used method to recruit the MSM [30]. It is a probabilistic method to sample members of a given population at particular times in fixed venues (e.g., clubs, bars and gyms) [29]. The SPOT study recruits MSM individuals without following any structured sampling plan. Therefore, SPOT's sampling design is not probabilistic, and as such the generalizability of the findings from the data analysis are uncertain. To improve the generalizability of the findings, the sampling scheme in the SPOT should be adjusted.

In this thesis I analyze SPOT data to study the correlates of the phylogenetic cluster size by addressing both challenges: measurement error in cluster size and a non-probabilistic sampling scheme. The proposed approach to address these issues consists of extending the SIMEX method to correct for measurement error that is not mean zero and may depend on other covariates in the data and using sampling weights to adjust for the non-probabilistic sampling scheme, where weights are calculated using an external source of information.

I address the issues in three parts that are briefly described below. In the first manuscript, I extend the SIMEX method to accommodate errors with non-zero means, and refer to it as non-zero mean SIMEX (NZM-SIMEX). I provide theoretical justification for the generalization to the non-zero mean measurement error case by proving the consistency of the estimators in a linear regression setting, and demonstrating its performance in more general settings by computer experiment. Performance of the NZM-SIMEX is empirically compared to the naïve method via a simulation study under ideal and non-ideal conditions. I apply the NZM-SIMEX method to the SPOT data for HIV-positive participants to determine behavioural correlates of cluster size.

The second manuscript focuses on further extending the SIMEX to the settings where the measurement error in a covariate of interest depends on the value of another correctly measured variable. For the SPOT data analysis, the NZM-SIMEX is applicable only to the data from HIV-positive MSM for whom there is measurement error in the cluster size. For HIV-negative MSM, the cluster size is always zero and there is no measurement error. Thus, the measurement error in the cluster size for MSM depends on a correctly measured variable: HIV status. To be able to include the data from both HIV-positive and HIV-negative MSM in the analysis, I extend and validate the SIMEX method to accommodate measurement error distributions that (i) depend on other covariate and (ii) need not have mean zero. I refer this to as SIMEX conditional on covariates (SIMEX-CC) method. I then compare the performance of the SIMEX-CC to the other methods such as regression calibration, multiple imputation for measurement error and naïve methods. I apply the proposed SIMEX-CC method to the full SPOT data that include both HIV-positive and HIV-negative participants.

Finally, I propose solutions to simultaneously adjust for two challenges: measurement error in cluster size, and non-probability sampling scheme. To adjust for the sampling design, I use weighted regression model, where sampling weights for SPOT participants were calculated using data from another study of MSM in Montreal, whose sampling design was known. To correct for measurement error, I use the SIMEX-CC method.

In this thesis, I extend SIMEX to account for measurement error that is structured in this sense that it may have a mean or variance that depends on another variable. Further, I provided a demonstration of how external data can be leveraged to improve the generalizability of a study where participant recruitment was via a convenience sample. These statistical developments were motivated by the SPOT study in Montreal. However, the methods have much wider applicability. For instance, there are many laboratory assays that are subject to measurement error. Further, many populations are difficult to sample. This has driven developments in sampling and estimation such as respondent-driven sampling – however even these methods can be biased. Even in a probabilistic sampling design, studies may suffer from

lack of generalizability due to non-response. The use of data from a study with probabilistic participant sampling can be used to improve generalizability of results in these settings.

Chapter 2
Literature Review

2.1 Introduction

This chapter provides a review of the epidemiological literature on HIV, and the statistical literature on measurement errors and sampling from hard-to-reach populations. Starting with current HIV research that focuses on the phylogenetic clustering, I then provide an overview of important concepts of measurement error and the consequences of measurement error in both exposure and outcome variables. I next discuss different methods for correcting measurement errors. Finally, several sampling techniques for recruiting hard-to-reach populations are considered.

2.2 HIV and AIDS

HIV causes a devastating infection that leads to AIDS, if left untreated [31]. HIV destroys the body's specific defense system against all infectious agents by infecting CD4 immune cells. HIV is transmitted primarily through unprotected sexual intercourse (including anal and even oral sex, although with very low probability), infected blood transfusions and contaminated hypodermic needles, and can also be transmitted from infected mother to child through pregnancy, delivery, or breastfeeding [2]. There is no cure for or vaccine against HIV and without treatment, the risk of death for people living with HIV is very high. However, antiretroviral treatment (ART) can slow or halt the replication of the virus and prevent infections and cancers that often develop in people with HIV. Where modern ART is available, HIV is often seen as a manageable, chronic condition – albeit one that is costly and incurable.

There are two variants of the HIV virus, HIV-1 and HIV-2. From genetic research, it is believed that both HIV-1 and HIV-2 are originated in non-human primates in West-central Africa and were transferred to humans during the early twentieth century [32]. HIV-1 appears to have originated in southern Cameroon through the evolution of SIV(cpz), a simian immunodeficiency virus (SIV) that infects wild chimpanzees [33, 34]. The closest neighbour of HIV-2 is SIV(smm), a virus of the sooty mangabey, an old world monkey living in West Africa, from southern Senegal to western Côte d'Ivoire [35].

AIDS was first clinically observed by the Centers for Disease Control and Prevention (CDC), USA, in 1981 among a cluster of injection drug users and homosexual men [3–5]. Now, HIV is seen in both homosexual and heterosexual men and women, and HIV infection is considered pandemic by the World Health Organization [36]. As of 2015, approximately 36.7 million people have contracted HIV globally [37].

2.3 Phylogenetic Clustering: Offering New Insights into Viral Epidemics

Initially AIDS research was focused on identifying its cause, correlates of infection such as high-risk sexual behavior, developing treatments, and medical and social strategies to prevent new infections. In the last decade, with the advance of RNA and DNA sequencing technology, there is considerable research activity centered on using molecular phylogenetics. By comparing viral genomic sequences, relatedness of viruses can be determined, and from that, clusters may be inferred. This information can be used to study sexual networks on a local level, and to trace the evolution of the infection at a wider level [32].

There is a growing body of work looking at the use of DNA phylogenetics in different areas of research. Phylogenetics have been used in diverse scientific endeavours, from investigating gene duplications shared by animals, fungi and plants [38] to investigating the spread of Hepatitis C virus among European injection drug users with HIV [39]. Other studies that use phylogenetic clustering include gene encoding (see, e.g., [40–44]), and the effect of environmental severity on phylogenetic clustering (e.g., [45]).

In the context of HIV, phylogenetics have been used, for example, to discover the animal origin of AIDS [32]; construct the transmission history of a known HIV-1 [46]; identify the direction of transmission of HIV in criminal cases [47,48]; and track clusters [49,50]. Chalmet et al. [49] investigated whether phylogenetic data could supplement epidemiological data and allow a more detailed description of local HIV epidemics. Using a combination of data from phylogenetic analysis of HIV sequences, patient demographics, infection route, clinical information and laboratory results, they found distinct differences between HIV-1 subtype B and non-B infections. The authors concluded that phylogenetic analysis did complement the epidemiological data and added value to the understanding of local HIV epidemics.

The use of phylogenetic clustering is clearly on the rise in HIV research. While the validity of the clustering algorithms are not in dispute, the utility of cluster size itself as a variable is not yet well understood. A key concern is that the primary variable (cluster size itself, or its categorized version: unique/small/large cluster) is mis-measured. The measurement error is not random, but almost surely underestimated due the absence of some HIV-positive individuals. A key objective of this research will focus on understanding and correcting biases in results that arise from this systematic error of the cluster size.

2.4 HIV Research in Montreal

Montreal, Canada, is one of the key centres for HIV research in Canada, and is home to three major studies focused on men who have sex with men (MSM): ARGUS, SPOT, and OMEGA. There are many challenges in studying individuals infected by a virus that has significant cultural implications, including social stigma, or in recruiting from a population of individuals who may not be eager to identify as part of that population. Each of these studies has complementary strengths. Below, I provide details on SPOT and ARGUS, the two sources of data used in the thesis chapters that follow.

2.4.1 The SPOT Study

In response to the alarming number of new HIV cases in Quebec among MSM and a relatively low rate of awareness of HIV status, community workers, medical professionals, and researches worked together to set up an innovative project called SPOT, an multidisciplinary intervention research project in Montreal (www.spottestmontreal.com). The Montreal SPOT point of care testing site was opened in Montreal's Gay Village neighborhood in 2009. It promoted and recruited participants through advertisements in gay magazines and web sites as well as through outreach activities. The site offered rapid, free and anonymous testing to the MSM community. Rapid testing was offered at flexible hours (day, evening, and weekend) at the SPOT site with testing, questionnaire completion, and counselling performed by nurse or other trained member of the SPOT team. The questionnaire elicited information on socio-demographic characteristics, HIV testing behavior, sex life, attitude towards HIV, socio-sexual profile, etc. It has been found that the most common way for participants to have heard about SPOT is from their friends [51]. The most commonly reported reasons for being tested at SPOT are consistent with the combination of benefits that SPOT offers: short waiting times, convenient hours and location, and the availability of rapid testing that is free of charge and anonymous [52–54].

In addition to providing a large number of rapid tests to the MSM community, SPOT data have produced valuable research insights [52–55], using the socio-demographic and behavioral data to identify the motivation and barriers to seeking HIV testing among MSM. For example, not having a doctor has been identified as the most important barrier that prevented MSM born outside of Canada from being tested in the past [53].

The lab of Dr. Bluma Brenner at the McGill University AIDS Centre performs nucleic acid amplification testing on dry blood spots of all samples from HIV-positive individuals in SPOT [13]. HIV RNA cluster size is determined using phylogenetic linkage (sequence interrelationships) which is established using neighbour-joining trees and maximum likelihood

methods, using BioEdit and MEGA2 integrated software [9]. Transmission cluster membership is based on the robust criterion of high bootstrap values ($> 98\%$), short genetic distances (< 0.01), and congruent polymorphisms and mutational motifs. Specifically, whenever a new HIV-positive individual is identified in SPOT, members of the Brenner lab run that sequence on a consensus tree that has one sequence from each small cluster and each large cluster. If the SPOT sample does not associate with a cluster, the lab searches their database for any virus that shares polymorphisms to assign new clusters which get added to the consensus tree. Since primary infection is the driving force of the epidemic, the size of existing clusters may change rapidly in time.

Some studies have made use of the phylogenetic data that has been collected on SPOT participants' viral RNA. Brenner et al. [56] described the phylogenetic expansion of the MSM epidemic in Montreal during last 10 years. It was confirmed by the phylogenetic clustering analyses that primary and recent stage infection plays an important role in transmission dynamics. Brenner et al. [57] investigated the underlying factors affecting the temporal growth of HIV epidemic among MSM by combining viral phylogenetic and behavioural risk data. The study found no behavioural correlates of cluster size, though there was an association between participant's age and cluster size. The study did not, however, account for potentially important sources of bias in the data, including measurement error and non-random sampling from the target population.

With a view to investigate the significance of cluster size to the spread of the HIV epidemic among MSM, the Brenner lab, in previous work, stratified clusters into three sizes: (i) "unique" infections; (ii) "small" clusters of two to four infected individuals; and (iii) "large" clusters of five or more infected individuals [9]. The episodic durations of clustered and non-clustered clustered transmissions are the basis for these selected cutoffs in the [9], but more recently, "small" cluster have been redefined to contain two to nine individuals, and "large" to contain 10 or more. Given the changing understanding of what constitutes a large cluster, there is interest in assessing correlates of cluster size as measured on its

natural (count) scale, and ensuring appropriate corrections are made for inaccuracies in the 'measurement' of the cluster size.

2.4.2 The ARGUS Study

ARGUS (www.argusquebec.ca) was a survey of MSM in Quebec that collected information on HIV status, sexually transmitted infections, viral hepatitis and associated risk behaviours, aimed at combining data on infection surveillance and behavioural monitoring. ARGUS was executed under the direction of the Direction de santé publique de l'Agence de la santé et des services sociaux de Montréal, the Public Health Agency of Canada and the Institut national de santé publique du Québec and by a team of representatives from the community, university and public health. ARGUS is one of the only studies of MSM to employ a probabilistic recruitment scheme. Individuals were enlisted from a wide range of locations such as saunas, bars, coffee shops, and sports and recreational groups where gay men interact with each other. Individuals were recruited by the interviewers following a sampling method that is adapted to the location visited. The recruited individuals are considered to be a representative sample of all the MSM individuals who interact in the locations where recruitment took place. ARGUS was repeated in waves, tracking the changes in risk behaviours and prevalence of infections over time. ARGUS recruits all MSM aged 18 and over irrespective of their HIV status.

The ARGUS questionnaire focused on participants' socio-demographic characteristics, health, drugs and alcohol use, the structure of their social network, and their attitudes towards HIV. From the 42 locations in Quebec (37 in Montreal, 4 in Quebec and 1 in Laval), 1873 individuals were enlisted between May 2008 and March 2009.

2.5 Measurement Error

In modelling the association between a response and covariates, it is typical to assume implicitly or explicitly that the response variable and the covariates are measured without errors. However, this ideal situation is not always met in practice for several reasons including

non-response errors, reporting errors, and computing errors. No matter what the reasons are, measurement error is a difficult problem. Errors of measurement in covariate(s) cause various problems including: biased parameter estimation in regression models; loss of power in detecting association among variables; and concealing of features of the data [58]. The impact of measurement error in a response is typically loss of efficiency. Thus to make valid statistical inference, measurement error in variables must be taken into consideration. In this section, I will review important concepts and consequences of measurement error.

Let Y denote a response variable and V is a correctly measured covariate. Moreover, we have another covariate U whose imprecise measure is available to us which is X. Here, U is often called the error-prone predictor or latent predictor whereas X is called the surrogate variable [58]. Our intention is to relate the response Y with the true predictors U and V. If X is being used instead of U for modelling purpose, then this is often called a naïve approach. Adopting a naïve approach typically leads to biased parameter estimates and hence inferences can be misleading.

For analyzing data in the presence of measurement error, it is crucial to understand the measurement error process so as to decide on the most appropriate form of analytic correction. There are two general types of error models: differential and non-differential error. Within either type of error, further classifications and distinctions can be made. For example, whether differential or not, error may be additive or multiplicative. Within the class of additive, non-differential error, we may consider additional categories of error, such as the classical additive measurement error model and the Berkson additive measurement model. In this thesis, I will focus on non-differential, classical additive errors.

2.5.1 Differential Measurement Error

Differential measurement error occurs when the magnitude and/or direction of the error is different for individuals who have and have not experienced the (binary) outcome. For example, if individuals are being asked about their smoking habits after being diagnosed with

lung cancer, it seems plausible that their answers may be affected by this knowledge. The resulting error is then called differential measurement error.

Another example could be related to assessing food exposures in a case-control study of Chron's disease (or stomach cancer or some other gastric or colonic condition), where the cases may well have different recall than the non-cases, making the error differential.

2.5.2 Non-differential Measurement Error

Non-differential measurement error is an error that is unrelated to the outcome status; the magnitude and direction are equal for individuals who have the outcome compared to those without the outcome.

2.5.2.1 Classical Additive Measurement Error

In the classical additive measurement error model, the conditional distribution of X given U, in its simplest form, is as follows:

$$X_i = U_i + \delta_i, \tag{2.1}$$

where δ_i are independent and identically distributed (i.i.d.) with mean 0 and variance σ_δ^2, and is independent of U_i. This model is suitable when it is desirable to determine U_i directly, but one is unable to do so due to several errors in measurement. For example, consider a study that investigates the effect of mean exposure to microwave radiation emitted from cell phone towers on birth defects of babies born to women living near cell phone tower(s) during a 15 year window. The birth defect is detected by diagnostic tests. Also, each woman has one measurement of radiation level, taken at randomly selected time during the 15 year exposure window. However, if we could make several replicate measurements (at randomly selected different times), the mean radiation would be a better indicator of exposure during the 15 year window. This measurement error in the radiation exposure is therefore classical.

2.5.2.2 Berkson Measurement Error

In the Berkson measurement error model, the conditional distribution of U given X takes the form:

$$U_i = X_i + \delta_i, \tag{2.2}$$

where the δ_i are defined as before, and are independent of X_i. This form of error often arises in laboratory studies, where X can be measured directly but the 'uptake' of X cannot. For instance, consider a study of multivitamin to maintain muscle strength, where an accurately measured dose X of multivitamin was taken by an individual. However, the actual amount of multivitamin absorbed by the body is the exposure of interest (U), which differs from the amount taken (X), e.g., due to possible variation between individuals' body ability to absorb.

2.5.3 Effects of Measurement Error

Differential measurement error is particularly challenging, as its effects can be considerable and may lead to unpredictable biases.

In contrast, non-differential measurement error in the exposure is predictable: it leads to bias the exposure-outcome association towards the null (attenuation bias) [58], and a reduction in power. Non-differential classical error in (continuous) outcome variable does not systematically bias an association estimator but does increase its standard error. The consequences of non-differential measurement errors in the independent variable (covariate) and dependent variable (outcome) appear below in further detail.

2.5.3.1 Measurement Error in a Predictor Variable

Let us consider the simple linear regression model

$$Y_i = \beta_0 + \beta_1 U_i + \epsilon_i, \tag{2.3}$$

where the model error term $\epsilon_i \overset{indep}{\sim} N(0, \sigma_\epsilon^2)$, and assume a classic additive measurement error model where we have access to the observable variable X such that

$$X_i = U_i + \delta_i, \tag{2.4}$$

where $\delta_i \overset{indep}{\sim} N(0, \sigma_\delta^2)$ and is independent of U_i and ϵ_i. Therefore, instead of estimating (2.3) we estimate the model

$$\begin{aligned} Y_i &= \beta_0 + \beta_1(X_i - \delta_i) + \epsilon_i \\ &= \beta_0 + \beta_1 X_i + \epsilon_i^*, \end{aligned} \tag{2.5}$$

where $\epsilon_i^* = \epsilon_i - \beta_1\delta_i$. While it is still the case that $E(\epsilon_i^*) = 0, E(\epsilon_i^*, \epsilon_j^*) = 0, \forall\ i \neq j$ and $E(\epsilon_i^*\epsilon_i) = 0$, it is not true that $Cov(X_i, \epsilon_i^*) = 0$. Rather, we have that $Cov(X_i, \epsilon_i^*) = -\beta_1\sigma_\delta^2$. Thus , the covariate and error term in (2.5) are correlated, violating one of the fundamental assumptions of the classical linear regression model. Consequently, ordinary least squares (OLS) estimators thus obtained are biased, even asymptotically.

For the OLS estimator $\hat{\beta}_1$ from model (2.5), it can be shown that

$$\hat{\beta}_1 \xrightarrow{P} \beta_1 \left[\frac{\sigma_U^2}{\sigma_U^2 + \sigma_\delta^2}\right] = \beta_1\gamma,$$

where $\gamma = \frac{\sigma_U^2}{\sigma_U^2+\sigma_\delta^2}$ which is expected to be less than 1, hence the bias towards zero. This type of bias is commonly referred to as attenuation towards null. The attenuating factor, γ, is called the reliability ratio and its inverse is called the linear correction for attenuation.

Fricsh (1934) [59] derived the bounds on the regression coefficients considering linear regression in two directions: regressing Y on X and X on Y. For the naïve regression of Y on X, the OLS estimator $\hat{\beta}_1$ has a probability limit given by equation (2.6). To perform a reverse regression of X on Y, model (2.5) can be rewritten as

$$X_i = \alpha_0 + \alpha_1 Y_i + \epsilon_i^{**}, \tag{2.6}$$

where $\alpha_0 = -\beta_0/\beta_1$, $\alpha_1 = 1/\beta_1$ and $\epsilon_i^{**} = (\epsilon_i - \beta_1\delta_i)/\beta_1$. Applying OLS to (2.6), we obtain

$$\hat{\alpha}_1 = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}$$

so that

$$\hat{\beta}_1^{Rev} = \frac{1}{\hat{\alpha}_1},$$

where $\hat{\beta}_1^{Rev}$ is the new estimator of β_1 based on the slope coefficient from a regression of X on Y. The probability limit of $\hat{\beta}_1^{Rev}$ is

$$\hat{\beta}_1^{Rev} \xrightarrow{P} \beta_1 + \frac{\sigma_\epsilon^2}{\beta_1\sigma_U^2}. \tag{2.7}$$

From (2.6) and (2.7), it is clear that when $\beta_1 > 0$

$$\text{plim}\hat{\beta}_1 < \beta_1 < \text{plim}\hat{\beta}_1^{Rev}$$

and inequalities reverse when $\beta_1 < 0$.

Thus, measurement error in a single predictor variable can be bounded and, as seen in the section 2.6, numerous analytic approaches have been developed. Measurement error in multiple predictors, or in settings that are more complex than a straightforward linear regression, may have unpredictable effects. For example, Regier et al. [60] show that classical additive errors in confounders for inverse weighted estimators have completely unpredictable effects – attenuation, reversal of the effect (protective instead of risk-inducing or vice versa), augmentation of the effect, or no bias at all. Similarly, in non-linear regression models (logistic models), when several risk factors are subject to measurement error, the odds ratio estimates corresponding to any of these factors could be biased toward or away from the null value [61].

2.5.3.2 Measurement Error in the Outcome Variable

Let us again consider the simple linear regression model

$$Y_i = \beta_0 + \beta_1 U_i + \epsilon_i, \tag{2.8}$$

where Y_i is the true response that is not directly measurable. Instead, we may observe Y_i^* such that

$$Y_i^* = Y_i + \omega_i,$$

where ω_i denote measurement errors in Y_i with $\text{Var}(\omega_i) = \sigma_\omega^2$. Therefore, rather than estimating (2.8), we estimate

$$\begin{aligned} Y_i^* &= \beta_0 + \beta_1 U_i + \epsilon_i + \omega_i \\ &= \beta_0 + \beta_1 U_i + v_i, \end{aligned} \tag{2.9}$$

where $v_i = \epsilon_i + \omega_i$ is a composite error term which contains the traditional random error term as well as the measurement error term. To avoid complexity, we assume

$$\begin{aligned} E(\epsilon_i) = E(\delta_i) &= 0,\ Cov(X_i, \epsilon_i) = 0;\ \text{a typical assumption of linear regression,} \\ Cov(X_i, \delta_i) &= 0;\ \text{implying the measurement errors in } Y_i \text{ are uncorrelated with } X_i, \\ \text{and } Cov(\epsilon_i, \omega_i) &= 0;\ \text{implying the model error and the measurement error are uncorrelated.} \end{aligned}$$

Under these assumptions, applying OLS to model (2.8) or (2.9) would yield an unbiased estimator of β_1. That is, the measurement errors in Y_i do not affect the consistency of the OLS estimator. However, the variance of $\hat{\beta}_1$ will be affected. Using the typical formulas, we

obtain

$$\begin{aligned} Var(\hat{\beta}_1) &= \frac{\sigma_\epsilon^2}{\sum_{i=1}^{n}(X_i - \bar{X})^2} \quad \text{from model (2.8)} \\ Var(\hat{\beta}_1) &= \frac{\sigma_\epsilon^2 + \sigma_\omega^2}{\sum_{i=1}^{n}(X_i - \bar{X})^2} \quad \text{from model (2.9).} \end{aligned}$$

Certainly, the variance of the estimator from model (2.9) is larger than that of the estimator from model (2.8). Thus, even though the measurement errors in the dependent variable provide an unbiased estimator of the model parameter, the estimated variance would be larger than that of having no such measurement errors which would consequently affect power (negatively).

When the outcome variable is discrete, measurement error is often referred to as misclassification. Using misclassified responses in regression model can lead to inconsistent coefficient estimates when typical estimation techniques (for example, logit or probit) are used [62].

2.6 Measurement Error Correction Methods

This section describes the sources of data required for correcting measurement errors and several commonly used measurement errors correction techniques.

2.6.1 Sources of Data

A measurement error analysis requires information about U given (X, V) or about X given (U, V). These data sources can be broadly classified into two categories: internal and external. Internal data are usually subsets of the primary data with additional information whereas external data comes from independent studies of individuals not included in the primary dataset. Within each of these two categories, there are three types of data, namely, (i) validation data, (ii) replication data, and (iii) instrumental data. In validation data, measurements are available on both X and U. Replicated, or repeated, measurements on X

are available in replication data, whereas instrumental data contains measurements on X, together with another variable T which is highly correlated with U.

An internal validation dataset is often considered as most desirable, because all recognized analytical techniques can be applied to such data. Further, an internal validation dataset allows direct investigation of the error structure and tests of crucial model assumptions for measurement error, leading to relatively greater precision of estimation and inference [58].

Sometimes it is implausible to obtain exact measurement on U, as for example, when the measurement of interest is the average of long-term systolic blood pressure. In such situations, when there is a good reason to believe that mean of the replicated measurements is a superior estimate of U than a single observation, replicate measurements are made. With the classical measurement error model, replicated data also can be used to estimate the variance of measurement error.

The instrumental variable T may or may not be an unbiased estimator for U (i.e., $E(T) = U$). If T is internal, it is not necessary for it to be unbiased in order to be useful: it can be included in a traditional instrumental variables analysis. However, if T is external, it is typically useful only if it is unbiased for U. In such case T can be used in the regression calibration (RC) analysis [58].

2.6.2 Functional and Structural Approaches to Measurement Error

Depending on how X is related to U, measurement error models (correction approaches) can broadly be classified into two major classes: models that are *functional* or *structural*. In functional modeling, no assumption is made about the true covariate U; the values may be either fixed constants or random variables. In contrast, structural modeling considers U to be a random, and a parametric distribution of U is assumed.

In functional modeling, if U is regarded as random, only minimal, or no assumptions are made about the distribution of the unobserved U and, as such, resulting estimators and inference may be more robust. Some popular functional modeling approaches include

regression calibration and simulation-extrapolation (which is the focus of the methodological developments in this thesis).

Results from structural modeling depend on the assumed distributional form of U, leading to the potential for bias and lack of robustness (see, e.g., Fuller [63], Carroll et al. [58]). Likelihood-based methods, and Bayesian modelling are examples of structural modeling approach.

There is no decisive preference between functional or structural modeling approaches. Some researchers favour functional model arguing that one should consider as few model assumptions as possible. Other researchers prefer structural modeling on the grounds that one should try as best as possible to model every feature of the data to perform appropriate statistical analysis. Below, details of several approaches from each approach are provided.

2.6.3 Regression Calibration

Regression calibration (RC) [58] is a simple and widely used approach for adjusting measurement error in regression analysis [20, 61, 64–66]. The basic idea is to replace U by the regression of U on X. After this approximation, a standard regression analysis is performed. Regression calibration is effective and applicable to any regression model as long as the approximation is reasonable [58]. The approximation can be poor for highly non-linear models [58]. For a simple linear regression of the form

$$Y_i = \beta_0 + \beta_1 U_i + \epsilon_i,$$

the regression calibration method can be better explained as follows:

1. At the first step U_i are regressed on X_i and parameter estimates are obtained. This can be done using 'validation data' or 'unbiased instrumental variable' [61] or 'replicate data' [67].
2. At the second step the resulting estimates are used to predict the unobserved U_i for the original sample which we call $\hat{U}_i$. The standard regression of Y_i on $\hat{U}_i$ (for all

individuals, even those included in the validation sample or replicate data) is then run to obtain the parameter estimates $(\hat{\beta}_0, \hat{\beta}_1)$.

3. Finally, the standard errors of $(\hat{\beta}_0, \hat{\beta}_1)$ are adjusted (to account for the estimation of U) by adopting either the sandwich or bootstrap method (see, [67] for details).

Note that the RC method is based on the assumptions that the errors are non-differential with small variance. Further, the error model is required to be linear and nearly homoscedastic. Violation of these assumptions may lead to ineffective bias reduction, especially in non-linear outcome models [58].

2.6.4 Simulation-Extrapolation (SIMEX)

The simulation-extrapolation method developed by Cook and Stefanski [68] is a simulation based-technique for estimating and reducing bias due to additive measurement error. The method was further extended by Carrol et al. [69] and Stefanski and Cook [70]. SIMEX is a two-step estimation procedure consisting of a simulation step and an extrapolation step. Estimates are obtained by adding additional measurement error (in known increments) to the mis-measured data in a resampling-like stage, computing estimates from the contaminated data, establishing a trend between these estimates and the variance of the added measurement errors, and extrapolating this trend back to the case of no measurement error.

The main idea is to use the information from an incremental addition of measurement error to the mis-measured data using computer-simulated random errors. Adding extra measurement error to the data by simulation allows one to understand how the estimation bias is affected by the increase of the measurement error variance. This is the so-called simulation step. In the extrapolation step, the obtained parameter estimates are modelled as a function of the magnitude of the variance of the measurement error and extrapolated to the case of no measurement error. The algorithm is detailed in Chapter 4.

The SIMEX procedure for non-differential, mean-zero classical additive measurement error has been implemented in most popular statistical software (e.g., the `simex` package in `R`, and the simex function in Stata).

Simulation Step

Suppose U_i, $i = 1, \ldots, n$, is the true explanatory variable whose imperfect measurement is available which is denoted by X_i. Let us define X_i as

$$X_i = U_i + \delta_i,$$

where δ_i is an independent normal random variable with mean zero and variance σ_δ^2, and is independent of U_i and Y_i. In the simulation step, additional measurement error is added to the imperfectly measured covariate X_i, and B new covariates $X_{i,b}(\lambda_k)$ are generated using the rule:

$$X_{i,b}(\lambda_k) = X_i + \sqrt{\lambda_k}\, \delta_{ib},$$

where $b = 1, \ldots, B$; $k = 1, \ldots, K$ and $i = 1, \ldots, n$. Also, $\lambda_k \geq 0$ are assumed parameters which control the variance of the measurement error, and $\{\delta_{i,b}\}_{b=1}^{B}$ are independent computer simulated normal random numbers from $N(0, \sigma_\delta^2)$. Carroll et al. [58] recommended to choose λ_k as $0 = \lambda_0 \leq \lambda_1 \leq \cdots \leq \lambda_K = 2$. Note that the simulation step creates B additional datasets (replication samples to remove simulation variability) with the same dependent variable Y_i and the explanatory variable $X_{i,b}(\lambda_k)$ for each λ_k. The variance of $X_{i,b}(\lambda_k)$ is

$$\begin{aligned} \mathbb{V}\left[X_{i,b}(\lambda_k)\right] &= \mathbb{V}\left[X_i + \sqrt{\lambda_k}\, \delta_{ib}\right] \\ &= \mathbb{V}\left[U_i + \delta_i + \sqrt{\lambda_k}\, \delta_{ib}\right] \\ &= \sigma_U^2 + (1 + \lambda_k)\sigma_\delta^2 \end{aligned}$$

which increases with the control parameter λ_k. For each λ_k, let $\hat{\beta}_b(\lambda_k)$ denote the vector of naive estimates obtained by regressing Y on $X_{i,b}(\lambda_k)$. Using B estimates for each λ_k, an average estimate can be obtained as

$$\hat{\beta}(\lambda_k) = \frac{1}{B} \sum_{b=1}^{B} \hat{\beta}_b(\lambda_k).$$

Extrapolation Step

In the extrapolation step each component of vector $\hat{\beta}(\lambda_k)$ are plotted against λ_k for $\lambda_k \geq 0$, and regression techniques are used to fit an extrapolant function. The SIMEX estimator is obtained as the extrapolation of $\hat{\beta}(\lambda_k)$ at $\lambda_k = -1$, which is the ideal case of no measurement error.

The Literature on SIMEX

Since its original development by Cook and Stefanski [68] more than two decades ago, SIMEX has seen numerous extensions [68, 69, 71–79]. For example, Wang et al. [80] extended SIMEX to the correlated data setting for use in the generalized linear mixed models framework, considering normal additive measurement error in a single covariate. The resulting models were called generalized linear mixed measurement error models. Following this, Lin and Carroll [77] developed a score test for testing if variance components across clusters are zero. Yi et al. [81] employed SIMEX for marginal analysis of longitudinal data with covariate measurement error and missing data. Hu et al. [82] used SIMEX to develop a nonparametric procedure for analyzing survival and longitudinal outcomes measured with error.

In the last 15 years, a number of papers have appeared that use SIMEX method to survival analysis with covariates subject to measurement error. Mallick et al. [83] used SIMEX to adjust for measurement error in the exposure included in the Cox proportional hazard model and compared its performance to regression calibration. They used a version of RC where the observed exposure was replaced by its expected value based on the measurement error distribution. Overall, RC was found to perform better than SIMEX when measurement error distribution was correctly specified, although RC was not robust against misspecification of the error distribution. Greene and Cai [84] showed the asymptotic normality and consistency of the SIMEX estimator for models with multivariate failure time data and measurement error. Hu and Lin [85] proposed a modified score equation and established the asymptotic properties of the estimators for multivariate failure time data. Li and Lin [79] used SIMEX in frailty models with variables measured with error. He et al. [86] explored the SIMEX method under the accelerated failure time model. He et al. [86] discussed accelerated

failure time models with error-prone covariates and studied the bias induced by the naïve approach of ignoring measurement error in covariates. He et al. [87] explored the SIMEX method for survival data in the proportional odds model setting.

SIMEX has also been used for a variety of other settings. For example, it has been extended to correct for bias correction when discrete data are misclassified, known as the misclassified SIMEX (MC-SIMEX) [88]. Küchenhoff et al. [89] developed the asymptotic variance for the MC-SIMEX. A double SIMEX approach for bivariate random effects meta analysis of diagnostic accuracy studies where diagnostic accuracy measures were subject to measurement error was proposed by Guolo [90]. Other research areas where SIMEX has been developed to correct for measurement error include microarray data [91,92], semi-parametric modeling [75], and studying space using artificial intelligence techniques [93].

2.6.5 Multiple Imputation

Multiple imputation (MI) [94] is a three-step technique which consists of imputation, analysis and combination. Assuming the data are missing at random (MAR), the first step of MI imputes the missing values based on the predictors which are assumed to be associated with error prone variables. Each missing value is replaced with several ($m > 1$) plausible values to take into account the uncertainty around the actual value to be imputed. There are several imputation techniques such as joint modelling (typically assuming joint normality of the data), or conditional modelling using a combination of approaches which may include predictive mean matching [95] and regression techniques. At second MI step, all m imputed data sets are analysed using the usual complete case method. Step 3 combines all the results from step 2 by applying Rubins rules [21].

2.6.6 Likelihood Method

The likelihood method for correcting measurement error in exposure requires specification of a parametric model for each component of the data, i.e. for the full joint likelihood of the outcome, the exposure (both true and its mis-measured counterpart, and any covariates.

This involves in specifying: (i) the likelihood model as if true exposure (U) were observed, (ii) the error model (e.g, classical or Berkson), (iii) distribution for the unobserved U given the other covariates V (if the classical error model is chosen). Subsequently, the likelihood function is constructed by integrating the product of these densities over the latent true exposure (U) [65]. Model parameters are then estimated by maximizing the likelihood function adopting either analytical approximations or numerical methods [65].

To specify the distribution of true exposure, the likelihood method requires a validation study with several replicates of the unbiased measurements or strong and untestable (data-free) assumptions about the true covariate distribution. Further, the method is based on the non-differential measurement error assumption [96]. If all components are correctly specified, the likelihood method can be more efficient than other simpler approaches such as regression calibration. Nevertheless, it is rarely used in practice because of its computational complexity and difficulties in checking the parametric assumptions. Moreover, the robustness of the likelihood method to the modelling assumptions is often poor and difficult to assess [58].

2.6.7 Bayesian Method

Bayesian methods have long been used for correcting measurement error in the covariates (see, e.g., [97, 98]). To correct for measurement error in an exposure, the Bayesian approach involves several essential steps. First, as with likelihood method, a parametric model is specified for every component of the data, namely, the likelihood model treating U as observed, the error model, and the model for U given V. Second, the Bayesian approach treats U as missing data and imputes it several times by drawing samples from the posterior distribution of U given other covariates. Thus, the likelihood function of all the data is constructed. Third, all parameters are treated as random and appropriate prior distributions are assigned to them. Finally, Bayesian quantities, that is, posterior summaries of the association parameters are computed with the mean (or mode) of the posterior distributions serving as point estimates.

Bayesian analyses can be implemented via a flexible sampling-based Markov chain Monte Carlo (MCMC) algorithm [99] or with a computationally efficient non-sampling based integrated nested Laplace approximation [100, 101]. Although Bayesian MCMC method is flexible, it is computationally intensive. Moreover, as with the likelihood method, the Bayesian approach requires validation data or knowledge of the full joint data density. This method is appropriate for non-differential measurement errors.

2.6.8 Method of Moments

Method of moments is a commonly used method for eliminating bias. Considering the simple linear regression model

$$Y_i = \beta_0 + \beta_1 U_i + \epsilon_i,$$

where $U_i \overset{indep}{\sim} N(\mu_U, \sigma_U^2)$ and $\epsilon_i \overset{indep}{\sim} N(0, \sigma_\epsilon^2)$. With classical additive error model $X_i = U_i + \delta_i$; it can be shown that plim$\hat{\beta}_1 = \beta_1\gamma$ (see, Section (2.5.3.1) for details). If γ is known, then an unbiased estimate of β_1 can be obtained by simply dividing the OLS slope by γ. However, γ is usually not known in practice and we of course need to estimate it. If $\hat{\sigma}_\delta^2$ is an estimate of the variance of measurement error and $\hat{\sigma}_U^2$ is the sample variance of U, a consistent estimate of γ is $\hat{\gamma} = (\hat{\sigma}_U^2 - \hat{\sigma}_\delta^2)/\sigma_U^2$. Therefore, the estimate of β_1 is $\hat{\beta}_1/\hat{\gamma}$. The sampling distribution of this estimator is highly skewed for small samples, and as such, a modified version of this estimator is recommended by Fuller [63]. The method of moments estimator is not limited to the simple linear regression model but can also be constructed for general linear model [58].

2.7 Sampling Techniques for Recruiting Hard-to-Reach Populations

A number of sampling techniques have been proposed to gather information on hard-to-reach populations, i.e. groups of people who may not wish to self-identify publicly, often for fear of stigmatization. High-risk populations for HIV may include of sex workers, injection drug users, men who have sex with men, and specific mobile or migrating groups.

Such populations are often "hidden" or hard-to-reach. To obtain a sample of a suitable size for research from such a population, conventional approaches such as a household survey or national censuses are typically not appropriate as they often cannot obtain a sufficiently large enough sample, or because individuals may be unwilling to report high risk behaviours in such settings. For meaningful surveillance or surveying of such populations, a number of sampling strategies have been developed that are feasible and capable of producing unbiased estimates (or at least more realistic, less biased estimates) than traditional probabilistic survey methods; these include snowball sampling, targeted sampling, respondent-driven sampling, time-location sampling, and venue-based sampling. All but the last of these are based on the principle that members of the hard-to-reach population know one another, and are thus connected by a network.

2.7.1 Snowball Sampling

Snowball sampling [25] has long been used by researchers for recruiting hidden populations for surveillance. To take a snowball sample, the researcher must identify a few members of the target population (termed the "seeds"), then ask each of those individuals to identify other members of that population ideally naming all the members that they know. Each new participant in the sample is asked to identify additional members, until the desired sample size is attained. Sampling bias is one of the major concern for this technique: the obtained sample may not be representative of the target population as the composition of the sample is greatly influenced by the initial seeds and by the "connectedness" of the social network of the population. Moreover, the sample selection may favour the most cooperative or open subjects (a trait which may be related to risk-taking or other behaviours relevant to the question of interest) as well as those belonging to larger personal networks.

2.7.2 Targeted Sampling

Targeted sampling extends the ideas of snowball sampling to overcome some of its limitations [27]. In this approach, an initial ethnographic assessment is performed with a view to

identifying possible sub-groups or networks within the population. The identified subgroups are then regarded as strata, and individuals are chosen from each stratum using systematic sampling (if possible) [26]. The success of this sampling however depends heavily on validity and completeness of the ethonographic assessment.

2.7.3 Respondent-Driven Sampling

Like snowball and targeted sampling, respondent-driven sampling (RDS) [102, 103] also relies on a chain referral sampling approach, but the selection process is implemented in such a way that allows calculating selection probabilities, and hence it can be seen as a probability sampling technique. Specifically, in RDS, the selection process starts with identifying initial participants (again termed seeds) who are enlisted as both participants and recruiters. They are provided an explanation of the study and fixed number of coupons (e.g. three or even ten) that are to be given to recruited peers who are eligible for the study. Each new participant (respondent) receives similar number of coupons, as do each their recruits, until the desired sample size is attained. Further, the coupon approach allows the researchers to determine the network structure in the population, and as participants recruit their peers directly (but do not reveal peer information to researchers), the "masking effect", whereby participants may wish to protect their peers anonymity in a snowball sample, is reduced. Nevertheless, challenges remain as the connectedness of the network and the response rate both influence bias so that, as in snowball sampling, the 'cooperative' individuals may be more likely to be part of a large personal network.

RDS differs from snowball sampling in that the seeds are limited to recruit only as many participants as the number of coupons they receive, which in turn reduces the influence of initial seeds on the final sample composition. Restricting the number of recruits may encourage longer recruitment chains, thus increasing the 'reach' of the sample into potentially more hidden pockets of the population. Another distinguishing feature of RDS from snowball sampling is that at each cycle the relationship between recruiters and the recruits is documented. This allows researchers to assess any recruitment biases and to account for these

in the analysis. For instance, homophily (the tendency of individuals in a network to be similar) is a key parameter in RDS estimation in techniques that can be incorporated into the analyses.

2.7.4 Venue-Based Sampling

Venue-based sampling is one of the frequently used sampling techniques in recruiting hard-to-reach population such as MSM. This probabilistic sampling scheme is used to recruit MSM at particular times in set venues, for example, gay bars, parks, gyms, clubs [29]. The sampling frame includes a list of all potential venue-day-time (VDT) units where the MSM typically gathers. The example of a VDT unit could be a particular time of 3 hours on a Sunday in a specific venue. For locating the members of the target population, a range of VDT units is identified by interviewing the key informer, service providers, and members of the target population. The data collection team then visits the different venues, checks the presence of the specific individuals, prepare the list of possible VDT units, and estimate the population size for every VDT unit. Once the sampling frame is constructed the sample is then chosen in two stages. In the first stage venues are selected as primary sampling units using simple or stratified sampling with probability proportional to the estimated size. In the second stage, a sample of participants from the selected venues is drawn using systematic sampling. Under the venue-based sampling, informal venues, such as private houses could be included in the sampling frame that allows to reach least noticeable units of the target population, or those who typically less frequent in public places [29]. Venue-based sampling has several advantages, for example (i) it allows the calculation of the selection probability for each individual in the sample; (ii) unlike convenience sampling, it greatly diminishes the arbitrary selection of venues, subjects and provides a replicable sample selection method; and (iii) it does not require the comprehensive list of target population so long as all members of the population can be reached at fixed cites at different times. However, it requires intensive fieldwork for visiting and mapping venues, day-time units. Additionally, potential

bias arising from missing non-venue-based members of the target population/ MSM may restrict the generalizibility of the results.

Time-location sampling (TLS) is an extension of the venue-based sampling [104]. It relies on the assumptions that, while the sampling frame of the *population* of interest cannot be constructed or directly enumerated, *specific areas* (locations/events) where the desired population can be found are known and these can be enumerated. Thus, the sampling frame of the events can be constructed, and then a random sample of time(s) at these locations is selected (note that a particular location may be sampled at more than one time). If locations vary in the frequency with which they have the necessary number of attendees, some care is required in constructing the sampling frame or calendar of events, such as first choosing sampling periods for locations with the fewest available periods. At each sampled event, a sample of individuals at that location is chosen, randomly where possible. The investigators must then calculate (or estimate) the sampling fraction by recording the total number of persons at the location at that time who meet, or appear to meet, the eligibility criterion for the study [102].

A TLS analysis proceeds using weighting to account for the sampling approach. The sampling weight is defined as the reciprocal of the selection probability. Therefore, subjects who are more likely to be selected in the sample receive less weight. Conversely, those who are less likely to be included in the sample receive more weight.

Ideally, a selected sample should be representative of the population. That is, the sample should mimic the population characteristics with respect to all variables measured in the survey. Unfortunately, this may not be the case due to the several reasons. One of the causes is non-response that may lead to some groups to be under- or over-represented. Another problem is self-selecting sampling (typically in the online survey). This problem of cooperativeness is not specific to TLS or indeed any of the aforementioned sampling approaches, but rather can affect any survey in which participants must actively consent to participate. When information on the representatitiveness of the sample is available (thanks

to some external information on the non-participants), a commonly used correction method is a further weighting adjustment to address selection bias [105]. It assigns a weight to each member in the sample. Individuals in the under-represented groups are assigned a weight that is greater than 1, whereas subjects in the over-represented groups are assigned a weight smaller than 1. The sampling weights are then used to run weighted regression models, or to compute weighted means, totals and percentages, rather than reporting unweighted quantities.

2.8 Summary

In this chapter, I have provided a very brief overview of the epidemiology of HIV, and given a more comprehensive overview of measurement error and recruitment (sampling) strategies for hard-to-reach populations, with particularly focus on SIMEX and venue-based sampling, which will be key to the developments in the coming chapters.

Chapter 3
Objectives

The overarching aim of this thesis is to develop and validate methodological tools that can be used for studying correlates of HIV phylogenetic cluster size in MSM by combining phylogenetic and epidemiological data. Specifically, to propose new methods and address two major challenges while dealing with the SPOT data: (i) systematic undercounting in the cluster size, and (ii) the use of a non-probability sampling mechanism, which is common in studies of hard-to-reach populations such as high-risk HIV populations and MSM.

There are many methods for dealing with measurement error including regression calibration, multiple imputation, and simulation-extrapolation. While most of these methods require validation data or replicate data for some fraction of the observed sample, SIMEX does not require such validation data. In the context of my motivating example, obtaining a validation sample is both ethically and practically infeasible. Therefore, while SIMEX is an appropriate choice for the SPOT data analysis, SIMEX is limited to mean zero random errors, and hence further extensions are required to apply the approach in the settings where error distribution has non-zero mean in order to be used in the context of undercounted cluster size of SPOT data.

Further, the generalizability of results from a study with non-probability sampling scheme may be improved by fitting a weighted adjusted model. For SPOT study, there is no internal information that can help calculating sampling weights and hence an external source of information is needed that can be used for estimating or predicting sampling weights.

Therefore, the specific objectives of my doctoral thesis are to:

1. Extend the simulation-extrapolation method to non-mean zero measurement error.

2(a). Extend the SIMEX method to the settings where the measurement error distribution depends on a correctly measured covariate which may have non-zero mean.

2(b). Compare the performance of this extended SIMEX to other commonly used methods such as regression calibration and multiple imputation.

3. Demonstrate an analysis which simultaneously implements adjustment for a non-probabilistic sampling mechanism and measurement error in covariates.

4. Study the correlates of phylogenetic cluster size of MSM using the SPOT study data.

Chapter 4
Manuscript I: The Non-Zero Mean SIMEX: Improving Estimation in the Face of Measurement Error

Preamble

This is the first manuscript in a series of three that collectively addresses the overall thesis objective of addressing methodological issues that arise when studying correlates of phylogenetic cluster size in MSM. The research is based on data from SPOT, which is a study of MSM in Montreal that offers free HIV testing and collects data on socio-demographic and behavioural characteristics along with HIV phylogenetic cluster size, a measure which is subject to systematically undercounting. That is, phylogenetic cluster size is measured with error that does not have mean zero.

In this manuscript, I extend and validate the SIMEX to the non-zero mean measurement errors that mimic the undercounted cluster size setting in the SPOT data. I investigate large sample properties of the extended SIMEX estimators and compare its performance to the naïve method that ignores the measurement error. The methods are then applied to the SPOT data to reveal the association of HIV phylogenetic cluster size with demographic and sexual behavioural characteristics of MSM.

This article was published in *Observational Studies* in 2015. The references for this article have been combined with the overall thesis bibliography.

Manuscript I: The Non-Zero Mean SIMEX: Improving Estimation in the Face of Measurement Error

Nabila Parveen[1], Erica E. M. Moodie[1], and Bluma Brenner[2]

[1]McGill University, Department of Epidemiology, Biostatistics and Occupational Health

[2]Lady Davis Research Institute, Montreal, Quebec, Canada

Abstract

The simulation-extrapolation method developed by Cook and Stefanski (1995) is a simulation based technique for estimating and reducing bias due to additive measurement error armed only with knowledge of the variance of the measurement error distribution. However there are many instances in which validation data are not available, and measurement error is known not to have mean zero. For example, in assessing phylogenetic cluster size of HIV viruses, cluster size is systematically underestimated since clustering can only be performed on the viruses of those individuals who have presented for testing. In this setting, it is not possible to obtain validation data; however, using knowledge gleaned from the literature, the distribution of the errors may be estimated. In this work, we extend the simulation- extrapolation procedure to accommodate errors with non-zero means, motivated by an interest in determining behavioural correlates of HIV phylogenetic cluster size. We provide theoretical justification for the generalization to the non-zero mean measurement error case, proving its consistency and demonstrating its performance via simulation. We then apply the result to a data from the province of Quebec in Canada to show that findings from a naïve analysis are robust to a substantial range of possible measurement error distributions.

Keywords: SIMEX; non-zero mean measurement error; HIV.

4.1 Introduction

Since the discovery of the human immunodeficiency virus (HIV) in 1981, HIV has caused nearly 36 million deaths (as of 2012) [1]. While there is no cure or vaccine for HIV, current therapies are highly effective and have dramatically reduced mortality due to HIV. Nevertheless, HIV places an immense burden on individuals and societies, with the annual costs (medical and lost productivity) of new HIV infections in the United States estimated at $16 billion in 2010 [106]. There is considerable research activity on HIV in Montreal, Canada. One such study is SPOT [107], which offers rapid, free and anonymous testing to the community of men who have sex with men (MSM), primarily targeting men who frequent gay social venues. Individuals who are tested at SPOT provide questionnaire data, and for all individuals found to be HIV-positive, their blood undergoes HIV sequencing. The HIV sequencing information is supplemented with HIV sequencing information from the Quebec genotyping program [1] to determine the size of the sexual network to which the individual belongs, i.e. the number of other HIV-positive individuals in the province of Quebec whose HIV sequence fall into the cluster in a phylogenetic analysis. Researchers wish to combine the phylogenetic and epidemiological data to learn about correlates of large phylogenetic clusters [13,18,19]. Transmission cluster size (or simply cluster size) is defined as the number of individuals falling into the same HIV phylogenetic grouping. For example, if the HIV sequence of six individuals fall into the same cluster, each will be said to belong to a cluster of size six; if there is an individual whose HIV genome sequence does not cluster with the HIV genome of anyone else in the Quebec genotyping program registry of sequences, this individual is said to belong to a cluster of size one. However, the data available do not include individuals who are HIV-positive but are unaware of their status (i.e. have never been tested) nor those who have not had their HIV genotyped (viral load less than 400 copies per ml) [13]; there may also be a small number who have been tested outside of the province of Quebec and not yet been seen by a physician in the province. Consequently, measurement error occurs in defining the cluster size. This measurement error is characterized by a systematic

undercounting of the true cluster size due to the absence of the individuals who have not been tested. Thus, to make correct statistical inference about correlates of sexual network size, this measurement error must be taken into consideration.

There are several approaches to handle measurement error: e.g., method of moments, regression calibration [73, 108], multiple imputation [22, 94], and simulation-extrapolation (SIMEX) [68]; most require validation data, which is infeasible to collect in the case of phylogenetic or transmission cluster size. Unlike regression calibration and multiple imputation, SIMEX does not require validation data. The approach does, however, require that the measurement error distribution is known or can be well-estimated. In some instances, such as when data arise from a well-understood laboratory assay, the error distribution may be known exactly. In other instances, the distribution may be estimated from validation data if available, or posited based on information available in the literature, or simply assumed (and varied) as in a sensitivity analysis. In the few existing applications of SIMEX in the epidemiological literature, the error distribution has been determined or estimated using a combination of expert judgement and data from the literature [79, 87, 109–114].

The simulation-extrapolation method developed by Cook and Stefanski [68–70] is a simulation based technique for estimating and reducing bias due to additive measurement error. The SIMEX procedure does not require validation data, but does require the distribution of the measurement error to be posited, which may be possible using known properties of a measurement instrument such as a laboratory assay, or from existing literature. SIMEX is a two-step estimation procedure in which additional measurement error is added (in known increments) to the mis-measured data in a resampling-like stage, and a trend between the resulting estimates and the variance of the added measurement errors is established. To date, SIMEX has been limited to mean zero random errors, and will therefore need to be extended to alternative error distributions to be used in the context of under-counted measures. We shall extend the method to accommodate errors with non-zero means, so as to apply it to the

SPOT data to determine behavioural correlates of cluster size. In Section 4.2, we develop the theory, then demonstrate its performance in simulations in Section 4.3. Next, we apply the method to SPOT. Section 4.5 discusses the findings.

4.2 The Simulation-Extrapolation (SIMEX) Method

In SIMEX, estimation proceeds in two steps: a simulation step and an extrapolation step. Estimates are obtained by *increasing* the measurement error in the mis-measured data in a resampling-like stage, computing estimates from the contaminated data, establishing a trend between these estimates and the variance of the added measurement errors, and extrapolating this trend back to the case of no measurement error. The main idea is to use the information from an incremental addition of measurement error to the mis-measured data using computer-simulated random errors. Adding extra measurement error to the data by simulation allows the researcher to learn about how the estimator's bias is affected by the increase of the measurement error variance. This is the so-called simulation step. In the extrapolation step, the obtained parameter estimates are modelled as a function of the magnitude of the variance of the measurement error and extrapolated to the case of no measurement error. We begin by briefly describing the simulation-extrapolation procedure for zero mean measurement error and then present in detail the extension to non-zero mean measurement error, which we call the non-zero mean SIMEX (NZM-SIMEX), then proceed to derive its large sample properties.

A Short Description of SIMEX: Suppose U_i, $i = 1, ..., n$, is the unobserved true explanatory variable and an error-prone version X_i is available, where $X_i = U_i + \delta_i$, for $\delta_i \sim N(0, \sigma_\delta^2)$ and it is independent of U_i and Y_i. In the simulation step of SIMEX procedure, artificial measurement error is added to X_i, and B new covariates $X_{i,b}(\lambda_k)$ are generated via $X_{i,b}(\lambda_k) = X_i + \sqrt{\lambda_k}\delta_{ib}$, where $b = 1, ..., B$; $k = 1, ..., K$ and $i = 1, ..., n$ for values of λ_k are chosen by the analyst and $\{\delta_{i,b}\}_{b=1}^{B}$ are independent computer simulated normal random numbers from $N(0, \sigma_\delta^2)$. It can be shown that the variance of $X_{i,b}(\lambda_k)$ is $\sigma_U^2 + (1+\lambda_k)\sigma_\delta^2$, which

increases with λ_k. For each λ_k, let $\hat{\beta}_b(\lambda_k)$ denote the vector of naïve estimates obtained by regressing Y on $X_{i,b}(\lambda_k)$. Using B estimates for each λ_k, an average estimate can be obtained as $B^{-1}\sum_{b=1}^{B}\hat{\beta}_b(\lambda_k)$. By regressing $\hat{\beta}_b(\lambda_k)$ on λ_k, and extrapolating back to $\lambda_k = -1$, we find the estimate $\hat{\beta}(-1)$ corresponding to error $\sigma_U^2 + (1+\lambda_k)\sigma_\delta^2 = \sigma_U^2$, i.e., to the error free setting. A prototypical example (based on simulated data) on the estimates $\hat{\beta}(\lambda_k)$ and the extrapolating function that describes the regression of $\hat{\beta}(\lambda_k)$ on λ_k is given in Figure 4–1 for illustration.

Simulation Step

Let us consider the simple linear regression model

$$Y_i = \beta_0 + \beta_1 U_i + \epsilon_i,$$

where the true predictor U_i follows a distribution with finite variance σ_U^2 and $\mathbb{E}[\epsilon_i] = 0$. Suppose X_i is an imperfect measurement of U_i which is defined as

$$X_i = U_i - \delta_i^*,$$

where δ_i^* follows a distribution with $\mathbb{E}[\delta_i^*] = \mu_{\delta^*}$ and $Var[\delta_i^*] = \sigma_{\delta^*}^2$. Also, δ_i^* is independent of Y_i and U_i. For example, in the SPOT data, where U_i is the true value of the count variable 'cluster size', it may be reasonable to assume $\delta_i^* \sim Poisson(\mu)$, so that $\mathbb{E}[\delta_i^*] = Var[\delta_i^*] = \mu$. In other instances we may wish to consider $\delta_i^* = |\delta_i|$, where $\delta_i \sim N(0, \sigma_\delta^2)$, so that δ_i^* follows a folded Normal distribution with $\mathbb{E}[\delta_i^*] = \sigma_\delta\sqrt{\frac{2}{\pi}}$ and $Var[\delta_i^*] = \sigma_\delta^2(1-\frac{2}{\pi})$. In the simulation step, additional, simulated measurement error is added to the imperfectly measured covariate X_i, and B new covariates $X_{i,b}(\lambda_k)$ are generated using the rule:

$$\begin{aligned} X_{i,b}(\lambda_k) &= X_i - \sqrt{\lambda_k}\delta_{ib}^* + (1+\sqrt{\lambda_k})\mathbb{E}(\delta_{ib}^*) \\ &= X_i - \sqrt{\lambda_k}\delta_{ib}^* + (1+\sqrt{\lambda_k})\mu_{\delta^*}, \end{aligned} \tag{4.1}$$

where $b = 1, \ldots, B$; $k = 1, \ldots, K$ and $i = 1, \ldots, n$. The parameters $\lambda_k \geq 0$ control the variance of the measurement error, and are chosen by the analyst, while $\left\{\delta_{i,b}^*\right\}_{b=1}^{B}$ are artificially introduced random numbers from the distribution of δ_i^*. Note that this is *not* identical to the

simulation step in the traditional (mean zero error) SIMEX, but rather an additional term, $(1+\sqrt{\lambda_k})\mu_{\delta^*}$, has been included in the generation of $X_{i,b}(\lambda_k)$ to account for the non-zero mean of the errors. Carroll et al. [58] recommended taking λ_k as $0 = \lambda_0 < \lambda_1 < \cdots < \lambda_K = 2$. Note that using (4.1) ensures that

$$\mathbb{E}[X_{i,b}(\lambda_k)] = \mathbb{E}(U_i).$$

The simulation step creates B additional datasets (replication samples to reduce simulation variability) with the same dependent variable Y_i and covariate $X_{i,b}(\lambda_k)$ for each λ_k. The variance of $X_{i,b}(\lambda_k)$ is

$$\begin{aligned}
\mathbb{V}\left[X_{i,b}(\lambda_k)\right] &= \mathbb{V}\left[X_i - \sqrt{\lambda_k}\delta^*_{ib} + (1+\sqrt{\lambda_k})\mu_{\delta^*}\right] \\
&= \mathbb{V}\left[U_i - \delta^*_i - \sqrt{\lambda_k}\delta^*_{ib} + (1+\sqrt{\lambda_k})\mu_{\delta^*}\right] \\
&= \sigma^2_U + (1+\lambda_k)\sigma^2_{\delta^*}
\end{aligned}$$

which increases with the control parameter λ_k. For each λ_k, let $\hat{\beta}_b(\lambda_k)$ denote the vector of naïve estimates obtained by regressing Y on $X_{i,b}(\lambda_k)$. Using B estimates for each λ_k, an average estimate can be obtained as

$$\hat{\beta}^{NZM}(\lambda_k) = \frac{1}{B}\sum \hat{\beta}^{NZM}_b(\lambda_k). \tag{4.2}$$

Extrapolation Step

In the extrapolation step, each component of the vector $\hat{\beta}(\lambda_k)$ is plotted against λ_k for $\lambda_k \geq 0$, and regression techniques are used to fit an extrapolant function. In particular, $\hat{\beta}^{NZM}(\lambda_k)$ is typically regressed on λ_k assuming either a quadratic or a non-linear relationship (e.g., a lowess smoother). The NZM-SIMEX estimator, denoted $\hat{\beta}^{NZM}$, is obtained as the extrapolation of $\hat{\beta}(\lambda_k)$ at $\lambda_k = -1$, which is the ideal case in which there is no measurement error. See Figure 4–1 for a prototypical figure showing a plot of $\hat{\beta}(\lambda_k)$ against λ_k and the resulting NZM-SIMEX estimate.

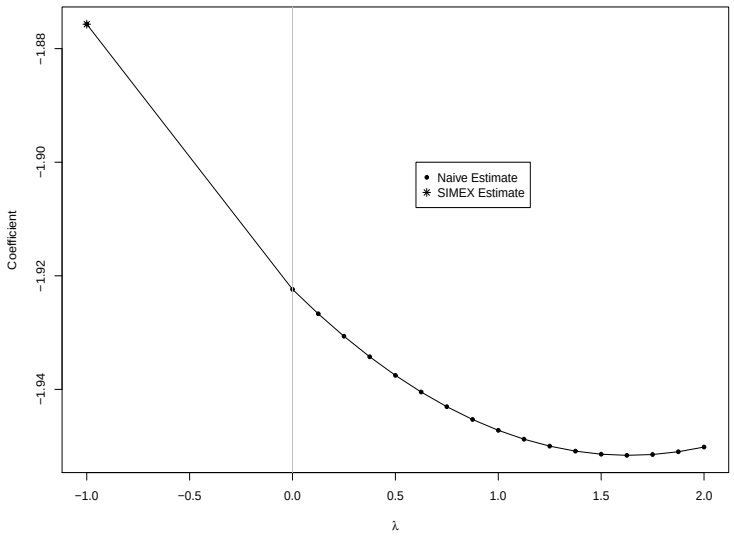


Figure 4–1: A generic plot of the effect of measurement error of size $(1+\lambda)\sigma_\delta^2$ on the parameter estimates. The SIMEX estimate is an extrapolation to $\lambda = -1$ whereas the naïve estimate occurs at $\lambda = 0$.

Below, we state two key properties of the NZM-SIMEX estimator, $\hat{\beta}^{NZM}$; proofs in the linear regression setting are provided in the Appendix A. As in the zero-mean error distribution setting [68], results hold for more general regression problems, including the fitting of generalized linear models [79], non-linear regression models [69], quantile regression models [113], accelerated failure time models [87], and even generalized linear mixed models [80], but cannot be shown in closed form; results demonstrating the feasibility of the SIMEX in these setting has relied on simulations. As in the previous literature, we provide theorems for the linear regression setting, and demonstrate the performance of the method in the generalized linear regression setting by simulation but not analytically. Both theorems rely on the assumption that the variance of the measurement error is known and finite. The proofs rely extrapolating to the no-error setting; while we can show this explicitly (i.e. in a closed form solution) in a linear regression setting, the extrapolation does not rely on the distribution of Y.

Theorem 1:

The SIMEX estimator for non-zero mean measurement error, $\hat{\beta}^{NZM}$, converges in probability to β.

Theorem 2:

$\hat{\beta}^{NZM}$ is a non-linear function of λ_k.

4.3 Simulation Study

A simulation study was carried out to empirically evaluate the performance of the NZM-SIMEX procedure under ideal and non-ideal conditions for a variety of outcome and covariate distributions at different sample sizes. In particular, we consider both the case where the error distribution is known exactly, and cases where it is not (e.g. it is known that the error follows a Poisson distribution, but an incorrect mean is assumed). A large range of settings were considered, including but not limited to the Poisson-distributed error setting which will be used in the empirical analysis of Section 4.4, to showcase the versatility of the methodology across a variety of possible scenarios.

4.3.1 Design of the Simulation Study

As the derivation of the NZM-SIMEX is general, we aimed to assess its performance under a variety of conditions specified by the outcome and error distributions. Parameters were chosen to follow those used by Cook and Stefanski (1995). In all instances, we report the bias, standard error (SE) and mean squared error (MSE) of the naïve and NZM-SIMEX estimators based on 1000 simulations. The sample sizes considered were $n = 100, 500$ and 1000. In our motivating data that has been analysed in Section 4.4.2, we have sample size $n = 33$. Therefore, we considered some simulation situations for $n = 33$.

Three outcome distributions were considered: normal, Poisson and Bernoulli distributions. For normally distributed outcomes, data were generated from the model

$$E(Y|U,V) \quad = \quad \beta_0 + \beta_U U + \beta_V V + \beta_{UV} UV.$$

For the Poisson distributed outcomes, data were generated from a log linear regression model

$$log[E(Y|U,V)] \quad = \quad \beta_0 + \beta_U U + \beta_V V + \beta_{UV} UV.$$

For the binary response, data were generated from a logistic regression model

$$logit[P(Y=1|U,V)] \quad = \quad \beta_0 + \beta_U U + \beta_V V + \beta_{UV} UV.$$

Details of the simulation settings for (U,V), $\boldsymbol{\beta} = (\beta_0, \beta_U, \beta_V, \beta_{UV})'$, δ and δ^* are given in Table 4–2 of Appendix B, with Scenarios 1-10 covering Normally-distributed outcomes; Scenarios 11-12 the Poisson-distributed outcomes, and Scenario 13 the binary outcome.

For NZM-SIMEX procedure, we considered $\lambda_k \in \{0, \frac{1}{8}, \frac{2}{8}, \ldots, \frac{15}{8}, \frac{16}{8}\}$, $b = 200$, and

$$X_b = X - \sqrt{\lambda_k}\delta_b^* + (1 + \sqrt{\lambda_k})E(\delta_b^*),$$

where for the normally distributed outcome only, $\delta_b^* = |\delta_b|$.

4.3.2 Results of the Simulation Study

The simulation results are shown in Figure 4–2, Figure 4–3 and Tables 4–3 to 4–6 in Appendix B. It is evident from these results that the NZM-SIMEX procedure leads to a considerable reduction of the bias compared to the naïve estimator.

When the error distribution is correctly specified by the analyst in the NZM-SIMEX method, the bias of the NZM-SIMEX estimator is much less than the naïve estimator. Biases depend on the magnitude of measurement error, whatever the distribution of the measurement error (Tables 4–3, 4–5 and 4–6). However, we also see that the bias reduction in the NZM-SIMEX estimators is less pronounced with increasing degrees of measurement error.

Irrespective of the parametric distribution of the errors (folded normal or Poisson), when parameters of the measurement error distribution are incorrectly specified, it is observed from Table 4–4 that the NZM-SIMEX estimator performs sub-optimally. However, while the NZM-SIMEX estimator using an incorrect measurement error distribution to generate the simulated errors performs worse than the NZM-SIMEX using the correct measurement error

distribution, performance remains superior to that of the naïve estimator. Under-estimation the variability of the measurement error leads to greater bias in the NZM-SIMEX than over-estimation. It is also apparent from the results that, with the very few exceptions, the non-linear fit in NZM-SIMEX procedure yields less biased estimates than quadratic fit.

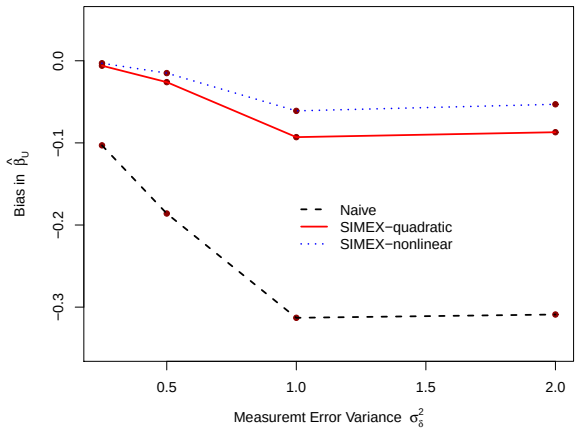
For discrete Poisson and binary distributed outcomes, it is observed from Table 4–5 and 4–6 that for the correctly specified error distribution, the NZM-SIMEX yields a less biased estimator than the naïve approach. In all cases, performance of NZM-SIMEX improves as the sample size increases. Thus, when the distribution of the errors is known, NZM-SIMEX performs well in recovering the true value of the parameter of interest. When the error distribution is mis-specified, the NZM-SIMEX procedure exhibits some bias, but nevertheless significantly outperforms the naïve estimator.

4.4 Analysis of the SPOT Data: Correlating Behaviour and Cluster Size

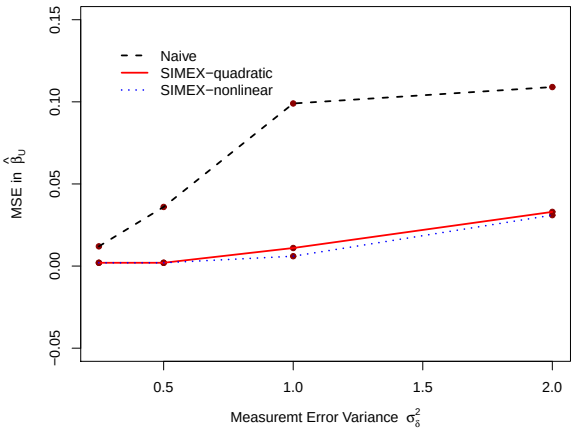
We now turn back to the motivating question in the analysis of the SPOT data. As noted above, neither SPOT nor the Quebec HIV genotyping program includes HIV-positive people who are unaware of their HIV status. We may also fail to capture individuals who underwent testing outside the province of Quebec. This induces measurement error in defining cluster size. In particular, it causes an underestimation of the true cluster size so that, clearly, measurement error in cluster size is not mean zero.

We used data from the SPOT study up until April 2012. At that time, SPOT had tested 1803 MSM, 34 of whom were found to be HIV positive. For all participants, questionnaire data includes several measurements on socio-demographic characteristics, HIV testing behaviour, sexual practices including risk behaviour, history of sexually transmitted infections, and attitudes toward HIV. In this analysis, we focus on the HIV-positive individuals and consider whether any of the following variables are correlates of cluster size: age, whether or not a condom was used at last sexual intercourse, number of sex partners, and whether

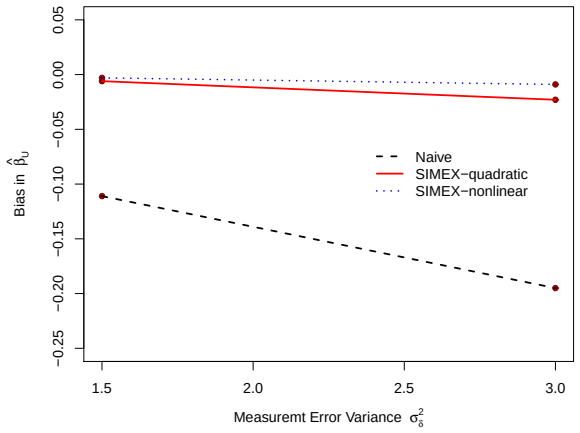
or not an HIV test was taken in the last 24 months. Except for cluster size, one individual's questionnaire data was incomplete; we omit this individual from the analysis, instead analyzing the 33 men with complete data.



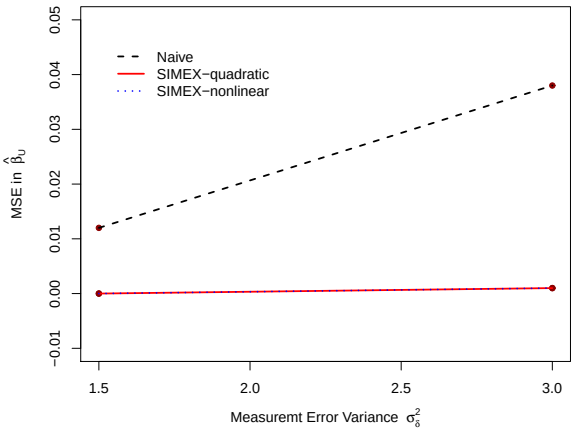
(a) Error Distribution: Folded Normal



(b) Error Distribution: Folded Normal



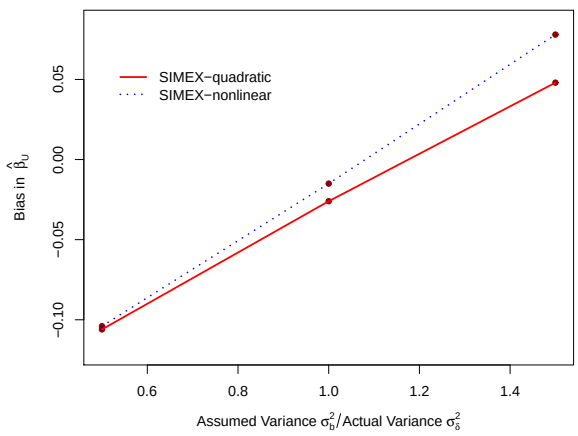
(c) Error Distribution: Poisson



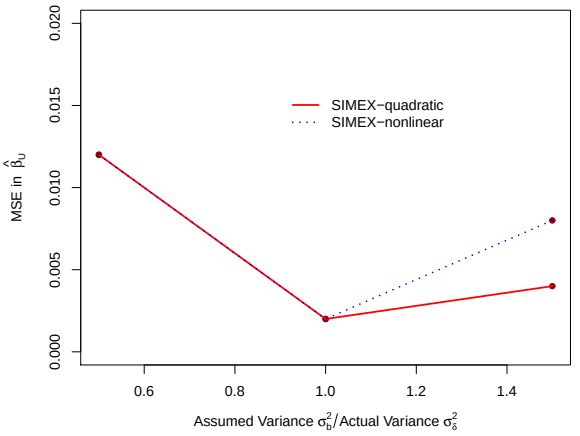
(d) Error Distribution: Poisson

Figure 4–2: Bias and MSE of the parameter estimator associated with the error prone variable for two different measurement error distributions.

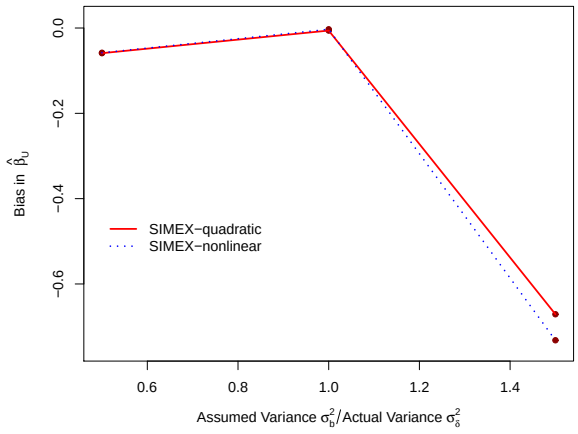
With the goal of identifying the relationship between cluster size and age, number of sex partners, not using a condom at last sexual intercourse, HIV testing status during last 24 months and number of one night partners we adopted seven distinct regression models.



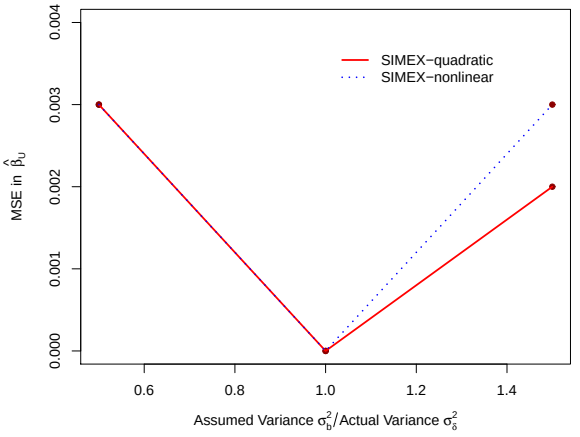
(a) Error Distribution: Folded Normal



(b) Error Distribution: Folded Normal



(c) Error Distribution: Poisson



(d) Error Distribution: Poisson

Figure 4–3: Bias and MSE of the parameter estimator associated with the error prone variable for two different measurement error distributions.

For each variable, both NZM-SIMEX (using quadratic and non-linear extrapolation) and a naïve model were used to obtain estimates. We fit two linear regression models of age on cluster size and number of sex partners on cluster size. We fit two logistic regression models, where in the first model not using a condom at the last sexual intercourse was considered as response variable and in the second model HIV testing status (during last 24 months) was taken as the outcome. Also considering number of sex partners and number of one night partners as count variables, we fit two log-linear models: number of sex partners on cluster size, and number of one night partners on cluster size. Furthermore, considering number of one night partners as a categorical variable (Category 1: < 2 partners, Category 2: $2 - 4$ partners, and Category 3: ≥ 5 partners), we fit a multinomial regression model considering Category 1 as the reference group. In all models, cluster size was the only covariate.

4.4.1 Measurement Error Cluster Size

Cluster size is an error-prone covariate; it is cardinal, and hence we assumed the error followed a Poisson distribution. Unfortunately, for data such as SPOT, there is no means of obtaining validation data to inform the distribution of the error short of testing all residents of the province of Quebec, which is both unethical and infeasible. Thus, to specify the mean of this Poisson distribution, we were required to estimate the cluster size distribution for those HIV-positive individuals who were not in the Quebec genotyping program because they had not received an HIV test or had been tested outside of Quebec. We now describe the process by which we estimated the distribution of the error in cluster size.

The adult (age > 15) population of Quebec in 2012 was $6,802,700$ [115] with HIV incidence rate 7 per $100,000$ [116]. Thus, the total number of *newly* HIV-positive individuals in Quebec can be estimated as $6,802,700 \times 0.00007 \approx 476$. In Canada, approximately 25% of people who are living with HIV do not know that they are infected [117]. Therefore, the estimated number of people who are HIV-positive but not in the Quebec genotyping cohort in Quebec can be estimated as $(476/0.75) - 476 \approx 159$. These 159 subjects are not included in determining the cluster size. Experts believe that among the MSM community, 15% do

not know their status [118], so our estimate of 25% may be conservative. Also, because clusters typically persist [119] for one or at most two years (i.e. after 12-24 months, few or no new infections are observed with a viral sequence that is genetically very similar), we use the annual HIV incidence rate rather than the prevalence rate to estimate the number of HIV-positive individuals in Quebec who are "missing" from our clustering cohort.

We then looked at the cluster size distribution of the 34 HIV-positive individuals from the SPOT data (see, Table 4–1) to estimate the cluster size for 159 unseen HIV-positive individuals. In the SPOT study 36% are linked to clusters that are at least of size 2-9, 29% are linked to clusters of size 1 and 35% are linked to clusters of size ≥ 10. Brown et al. [10] estimated cluster size for MSM from HIV sequences in the United Kingdom. They reported 29% belonged to cluster size 1, 41% are linked to $2-9$ individuals and 29% are part of cluster size of more than 10 people. [11] studied the short term dynamics of the episode among MSM in the United Kingdom. In their analysis they found that 15% belonged to cluster size 1, 60% are linked to $2-9$ and 25% belonged to cluster size ≥ 10. Based on these studies and the SPOT cluster size distribution, we propose a Poisson distribution for the error whose mean, on average, is big enough, to give us a distribution of cluster sizes that is similar to the percentages listed above (i.e. 25-30 % of people in clusters ≥ 10, 40 % in clusters of $2-9$ people). A reasonable Poisson distribution to achieve this would be Poisson(3). Poisson distributions with mean 1, 5, and 10 were also considered to evaluate the sensitivity of the results to the observed measurement error distribution.

4.4.2 Results

Table 4–1 shows the summaries of selected characteristics for 33 HIV-positive MSM. The mean age of the HIV-positive MSM in SPOT is 33. The average number of sex partners is 5.8. About 85% of individuals reported not using a condom on their last sexual intercourse and the majority (88.2%) reported having been tested for HIV in the last two years. Moreover, most (about 62%) belonged to clusters of size 3.

Results from all analyses, whether fit ignoring measurement error or accounting for the error using the NZM-SIMEX, were not significantly different from 0. The lack of significant findings does not appear to be driven by the small sample size leading to highly variable estimators: the estimates themselves were near the null values. For example, log-linear models examining the association between cluster size and number of sex partners (one night or total), point estimates indicate that a one-person increase in the cluster size is associated with a 0.3 - 0.5% increase in the number of sex partners. Considering that the average number of one night partners reported in the SPOT sample is (approximately) 4, one would need to compare groups of men whose cluster size differed by at least 40 people for the expected number of one night partners to increase by one individual to 5.

See Tables 4–7 to Table 4–10 in Appendix C for full results. A graphical representation of the SIMEX estimate has also been presented in Figure 4–4. To obtain standard errors (and p-values for the tests of association) for the NZM-SIMEX estimates, we used a bootstrap procedure with 1000 resamples. That is, both the naïve and NZM-SIMEX (both quadratic and non-linear) approaches yielded the same conclusions (cluster effect is not significant); the estimated parameter were different, but in most cases, not dramatically so. Moreover, different error distributions in all the models produce the similar results ensuring that results are robust to the assumption regarding the mean of measurement error distribution. We therefore conclude that the point estimates appear to be robust to the presence of measurement error. We observe that cluster size is not statistically significantly associated with the demographic and behavioural covariates of interest, suggesting that these individual level characteristics are unlikely to be helpful in identifying – and potentially breaking the cycle of HIV transmission within – large clusters.

4.4.3 Limitations and Discussion of the Analysis

This ongoing study primarily targets participants who frequent gay social venues and therefore may not be representative of the Montreal MSM population. Therefore, the results

Table 4–1: Characteristics of 33 HIV-positive MSM. For quantitative variables, mean (SD) are provided; for factor variables, counts (percentage) are reported

Characteristic	Summary Measure
Age	33 (9.5)
No.of Sex Partners	5.8 (4.7)
No condom use	29 (85.3%)
(in last sexual intercourse)	
HIV tested	30 (88.2%)
(during last 24 months)	
Cluster Size	
1	10 (29.4%)
$2-3$	3 (8.8%)
>3	21 (61.7%)
Number of one night partners	4.27 (4.7)
<2	14 (42.4%)
$2-4$	4 (12.1%)
>4	15 (45.5%)

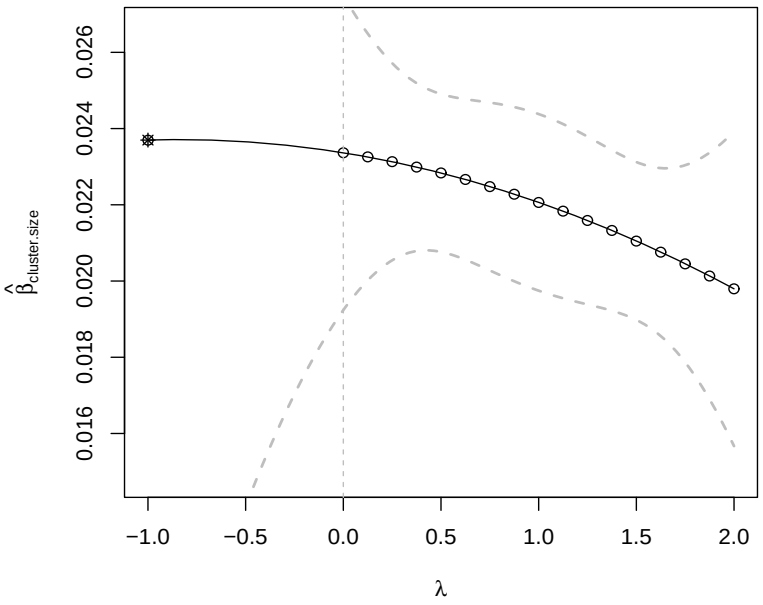


Figure 4–4: SIMEX estimate (using quadratic extrapolation) at $\lambda = -1$ from the SPOT analysis relating number of sex partner to cluster size. The naïve estimate occurs at $\lambda = 0$. The 95% pointwise confidence intervals are indicated by dotted (-) lines.

from this study may not be generalized to all MSM. More importantly, our conclusions are likely affected by limited power.

It is reasonable to speculate that the data in SPOT may be correlated: it is plausible that the individuals in the study may know one another, and have similar demographic or behavioural characteristics. While the available data provide no means of assessing any correlation beyond the phylogenetic clustering, and approximately half of the individuals in the SPOT study do not share HIV phylogenetic clusters with other SPOT participants, a simple approach did not reveal significant within-cluster pairwise correlation. For example, fitting (naïve) models of the association between each of age and number of one night partners as a function of cluster size via generalized estimating equations positing an exchangeable working covariance reveals a non-significant estimate of the within-cluster correlation of approximately -0.2. The very small size of the SPOT sample creates two challenges in this regard: lack of significance in the correlation could be driven by lack of power. On the other hand, a larger sample permit the inclusion of more covariates in the mean model, thus affording better assessment of the residual within-cluster correlation. While membership in the same HIV phylogenetic cluster can suggest direct sexual partnership, it is by no means strong evidence of it. Routinely collected sequencing data is not well suited to investigating transmission sources, as an individual whose HIV has not been sequenced may be a common source of infection or missing link in a transmission chain between two individuals in the same cluster with genetically similar viruses, thus creating challenges in identifying the likelihood that two individuals are indeed clustered in some sense beyond that suggested by the phylogeny of the virus which infects them [120]. Estimators based on analysis that acknowledge the impact of clustering in data tend to be more efficient for factors that vary within cluster, thus it is possible that our analyses missed a significant finding through statistical inefficiency. Given the very small point estimates, however, it seems implausible that any relationship that would be pertinent to public health planning or policy exists in the relationships examined here.

In estimating the error variance of cluster size from the existing literature, it should be noted that measurement error in cluster size was not taken into account in the cited studies [10, 11]. It is possible that our estimates of the error variance are thus too low; for this reason, we considered a range of plausible error distributions, however these did not serve to change the conclusions of our analyses.

All samples in the SPOT study were sequenced on the same platform: ABIPrism 3130xl genetic Analyser; this platform was also used in the Quebec genotyping program for the majority of the cohort's history from 2002 onwards, however the TrueGene/Bayer HIV platform was used from April 2004 to August 2006. Genome sequence interrelationships were determined using maximum likelihood phylogenies estimated using BioEdit and MEGA2 integrated software and PAUP (version 4, Sinauer Associates). Clusters were then assigned based on high bootstrap values (>98%), short genetic lengths (<1%), and congruent polymorphisms and mutational motifs [12]. To assess stability of the estimated cluster membership, phylogeny and estimate genetic distance was also estimated using a Bayesian approach via the BEAST (version 1.6.1) software; cluster size and membership estimated this way were similar to the maximum likelihood phylogenies. They were not, however, identical. Thus, both the sequencing platform and the clustering approach are additional sources of error introduced to the variable 'observed cluster size'. The distributional parameters of this error (mean, variance) are unknown, and were not taken into account in our analyses. As noted above, conclusions were unchanged under a range of plausible error distributions, suggesting that taking into account additional sources of error is unlikely to alter the conclusions of the analyses.

Finally, we wish to make two points regarding the interpretation of the our analyses. First, we remind the reader that a cluster is not representative of a sexual or social network. Rather, these are clusterings of the HIV genome taken from an individuals' serum sample at a fixed point in time (fixed for each individual, but varying across individuals). Individuals are then said to cluster if the sequenced HIV genomes are determined to be 'close', in terms

of phylogenetic distance. Second, we note that the analyses were undertaken only in an attempt to uncover whether there exists a significant correlation between various individual characteristics and cluster size. Cluster size evolves over time in a highly dynamic fashion, and thus the cluster size used in the analysis may not be reflective of the size of the cluster at the time when an individual was infected with HIV. We do not attempt to attribute any causal interpretation to the associations under investigation. The plausible directionality of the relationship is that individual characteristics could lead to bigger or faster-growing clusters, however as correlation (our estimand of interest) is a symmetric measure, we may 'reverse' the regressor and regressand without compromising its estimation.

4.5 Discussion

The simulation-extrapolation procedure is useful and easily implemented technique to deal with measurement error (Cook and Stefanski 1995), however its development was until now limited to mean zero random errors. In this work, we have extended SIMEX to the case where errors can have non-zero mean errors that can follow any known parametric distribution. This was developed with the goal of analyzing data of HIV infected MSM from the SPOT study, where measurement error occurs in defining the transmission cluster size because of not including people who were unaware of their status or who had been tested outside the province of Quebec.

In this work, we focused on the relatively simple setting of additive error that is independent of both measured covariates and the true, unobserved value of the mismeasured covariate. There are many settings in which this may not be realistic. For example, in the case of the variable cluster size, it is plausible to posit that the error is related to the size of the cluster so that bigger clusters have a greater error variance than smaller clusters. This situation is considerably more challenging, since the error variance is then completely unobservable. We are currently working on extending the NZM-SIMEX to the setting where the error variance depends on observed covariates; extensions to the latent variable setting will follow.

Through a number of simulation studies we evaluated the performance of NZM-SIMEX under ideal and non-ideal conditions for a variety of outcomes and covariate distributions at different sample sizes. Simulation studies showed that NZM-SIMEX performed reasonably well in reducing biases as compared to naïve approach in all cases. The method performs well in recovering the true value of the parameter when the distribution of the measurement errors are known, and offers improvements (reduced bias) over the naïve estimator even when the distribution of errors is known only approximately.

We then applied the method to the SPOT study data, in a first attempt to elucidate correlates of HIV phylogenetic cluster size. However this method is applicable in a number of other settings. For example, in studying the association between mother's age and child mortality using data from Demographic and Health Survey (DHS) of Bangladesh, researchers are faced with the challenge that women in the DHS frequently understate their age. The NZM-SIMEX could be applied to model the relationship between child mortality and mother's age, estimating the distribution of the reporting error through hospital records or other official registries. In other populations, the impact of illicit drug use on a variety of health and quality of life outcomes is of interest. Illicit drug use may be under-reported, and the magnitude of the error could be assessed via hair or urine samples.

The main limitation of the NZM-SIMEX is that it requires knowledge of the measurement error distribution. In case of mis-specified (or, if validation data were available, poorly estimated) error distribution, it may be safer to overestimate variability of measurement error. In such cases, the NZM-SIMEX estimators perform significantly better than the naïve estimators. Thus, to reduce the measurement error bias in a variety of problems, NZM-SIMEX may be considered as a useful and easily implementable approach.

Acknowledgements

This work was supported by Dr. Moodie's Operating Grant from Canadian Institutes of Health Research (CIHR); she is also supported by a Chercheur-Boursier junior 2 career award from the Fonds de recherche du Québec-Santé (FRQ-S).

The authors wish to thank Dr. Michel Roger, and are grateful to the SPOT study group and its participants.

4.6 Appendix for Manuscript I

4.6.1 Appendix A: Proofs

4.6.1.1 Proof of Theorem 1

Proof. Let us again consider the following simple linear regression model

$$Y_i = \beta_0 + \beta_1 U_i + \epsilon_i, \tag{4.3}$$

where true predictor U_i follows $\mathrm{N}(\mu_U, \sigma_U^2)$ and ϵ_i has mean 0. Suppose X_i is an imperfect measurement of U_i which is defined as

$$X_i = U_i - \delta_i^*, \tag{4.4}$$

where δ_i^* follows a distribution with mean μ_{δ^*} and variance $\sigma_{\delta^*}^2$, independent of U_i and Y_i. Note that under this measurement error specification, $P(X_i < U_i)$ may be at or near 1, depending on the distribution of U_i and δ_i^*.

As noted above, B new covariates $X_{i,b}(\lambda_k)$ are generated according to equation 4.1 so that the total measurement error variance is then the variance of $X_{i,b}(\lambda_k)$, i.e. $\sigma_{\delta^*}^2(1+\lambda_k)$. For the b^{th} data set, regressing Y on $X_b(\lambda_k)$ gives the vector of naïve estimates $\hat{\beta}_b^{NZM}(\lambda_k) = (\hat{\beta}_{0,b}(\lambda_k),\ \hat{\beta}_{1,b}(\lambda_k))'$ of $\beta_b(\lambda_k)$ found via ordinary least squares (OLS), with the average estimate at each λ_k computed according to equation 4.2.

To study the asymptotic mean of the average estimate of slope and intercept, we substitute (4.4) into (4.3), which gives

$$\begin{aligned} Y_i &= \beta_0 + \beta_1(X_i + \delta_i^*) + \epsilon_i \\ &= \beta_0 + \beta_1[X_{i,b}(\lambda_k) + \sqrt{\lambda_k}\delta_{ib}^* - (1+\sqrt{\lambda_k})\mu_{\delta^*} + \delta_i] + \epsilon_i \\ &= \beta_0 + \beta_1 X_{i,b}(\lambda_k) + \epsilon_i^*, \end{aligned}$$

where $\epsilon_i^* = \beta_1\{\sqrt{\lambda_k}\delta_{ib}^* - (1+\sqrt{\lambda_k})\mu_{\delta^*} + \delta_i\} + \epsilon_i$. For the b^{th} data set, the naïve estimate of the slope β_1 can be obtained by OLS, which yields

$$\begin{aligned}\hat{\beta}_{1b}^{NZM}(\lambda_k) &= \frac{\sum_{i=1}^{n}(X_{i,b}-\bar{X}_b)(Y_i-\bar{Y})}{\sum_{i=1}^{n}(X_{i,b}-\bar{X}_b)^2} \\ &= \frac{S_{XY}-\sqrt{\lambda_k}S_{Y\delta_b^*}}{S_{XX}+\lambda_k S_{\delta_b^*\delta_b^*}-2\sqrt{\lambda_k}S_{X\delta_b^*}}. \end{aligned} \tag{4.5}$$

The naïve estimate of the intercept is

$$\hat{\beta}_{0b}^{NZM}(\lambda_k) = \bar{Y} - \hat{\beta}_{1b}(\lambda_k)\bar{U}. \tag{4.6}$$

At each λ_k, the expected value of the estimator is

$$\hat{\beta}_1^{\ NZM}(\lambda_k) = E\left[\hat{\beta}_{1,b}^{NZM}(\lambda_k)|\{Y_i,X_i\}_{i=1}^n\right]$$

and

$$\hat{\beta}_0^{\ NZM}(\lambda_k) = E\left[\bar{Y} - \hat{\beta}_{1b}^{NZM}(\lambda_k)(\bar{X}+\sqrt{\lambda_k}\bar{\delta^*})|\{Y_i,X_i\}_{i=1}^n\right],$$

where the expectation is in terms of the distribution of $\{\delta_{i,b}\}$ only.

It then follows that

$$E\left[\hat{\beta}_1^{\ NZM}(\lambda_k)\right] = E\left[\hat{\beta}_{1,b}^{NZM}(\lambda_k)\right]$$

and

$$E\left[\hat{\beta}_0^{\ NZM}(\lambda_k)\right] = E\left[\hat{\beta}_{0,b}^{NZM}(\lambda_k)\right].$$

Using the fact that

$$
\begin{aligned}
S_{XY} &\xrightarrow{P} \sigma_{XY}, \\
S_{XX} &\xrightarrow{P} \sigma_{XX}, \\
S_{Y\delta_b} &\xrightarrow{P} \sigma_{Y\delta_b^*}, \\
S_{\delta_b\delta_b} &\xrightarrow{P} \sigma_{\delta_b^*\delta_b^*} \\
\text{and} \qquad S_{X\delta_b^*} &\xrightarrow{P} \sigma_{X\delta_b^*},
\end{aligned}
$$

we obtain

$$
\hat{\beta}_{1,b}^{NZM}(\lambda_k) \xrightarrow{P} \frac{\sigma_{XY} - \sqrt{\lambda_k}\sigma_{Y\delta_b}^*}{\sigma_{XX} + \lambda_k\sigma_{\delta_b^*\delta_b^*} - 2\sqrt{\lambda_k}\sigma_{X\delta_b^*}}
$$

and hence

$$
\hat{\beta}_{1}^{NZM}(\lambda_k) \xrightarrow{P} \frac{\sigma_{XY} - \sqrt{\lambda}\sigma_{Y\delta_b^*}}{\sigma_{XX} + \lambda_k\sigma_{\delta_b^*\delta_b^*} - 2\sqrt{\lambda_k}\sigma_{X\delta_b^*}}.
$$

Here,

$$
\begin{aligned}
\sigma_{XY} &= Cov(X,Y) = Cov(U,Y), \\
\sigma_{Y\delta_b^*} &= Cov(Y,\delta_b^*) = 0, \\
\sigma_{XX} &= Var(X) = Var(U+\delta^*) = \sigma_U^2 + \sigma_{\delta^*}^2, \\
\sigma_{\delta_{b^*}\delta_{b^*}} &= Var(\delta_b^*) = \sigma_{\delta^*}^2 \\
\text{and} \qquad \sigma_{X\delta_b^*} &= Cov(X,\delta_b^*) = Cov(U+\delta^*,\delta_b^*) = 0.
\end{aligned}
$$

By substitution into (4.5), we obtain

$$
\begin{aligned}
\hat{\beta}_{1}^{NZM}(\lambda_k) &\xrightarrow{P} \frac{Cov(U,Y)}{\sigma_U^2 + (1+\lambda_k)\sigma_{\delta^*}^2} \\
&= \frac{Cov(U,Y)}{Var(U)} \frac{Var(U)}{\sigma_U^2 + (1+\lambda_k)\sigma_{\delta^*}^2} \\
&= \beta_1 \left[\frac{\sigma_U^2}{\sigma_U^2 + (1+\lambda_k)\sigma_{\delta^*}^2} \right].
\end{aligned}
$$

Hence,

$$\lim_{\lambda_k \to -1} \text{plim} \hat{\beta_1^{NZM}}(\lambda_k) \quad = \quad \beta_1.$$

Similarly, considering (4.6), it can be shown that

$$\lim_{\lambda_k \to -1} \text{plim} \hat{\beta}_0^{NZM}(\lambda_k) \quad = \quad \beta_0.$$

□

In the SIMEX extrapolation step, each component of the vector $\hat{\beta}(\lambda_k)$ is modelled as a function of λ_k for $\lambda_k \geq 0$. For example, for the slope parameter, this modelling can be considered as a non-linear regression problem, with dependent variable $\hat{\beta}_1^{NZM}(\lambda_k)$ and independent variable λ_k having a mean function of the form

$$g(\lambda_k) = \beta_1 \left[\frac{\sigma_U^2}{\sigma_U^2 + (1+\lambda_k)\sigma_{\delta*}^2} \right].$$

The parameter of interest, β_1, can be obtained from $g(\lambda_k)$ by extrapolation to $\lambda_k = -1$, yielding SIMEX estimate of β. We now demonstrate that the dependence of $\hat{\beta}^{NZM}$ on λ_k is a complex, non-linear form.

4.6.1.2 Proof of Theorem 2

Proof. For the purposes of the proof, we will consider the slightly more complex and more realistic setting of multiple linear regression:

$$\begin{aligned} Y_i &= \beta_0 + \beta_Z Z_i + \beta_U U_i + \epsilon_i \\ &= \beta_V^t V_i + \beta_U U_i + \epsilon_i, \end{aligned} \tag{4.7}$$

where now $\beta_V = (\beta_0, \beta_Z)$, $V_i = (1, Z_i)$, and ϵ_i has mean 0. Here Y, V and U denote the response variable, and two covariates measured without error, respectively. As before, instead of the true predictor, U_i, an imperfect measurement X_i is available.

In the multiple linear regression setting, for the b^{th} data set, the regression model (4.7) can be expressed as

$$\begin{aligned} Y_i &= \beta_V^t V_i + \beta_U X_{bi} + \epsilon_i \\ &= \beta_V^t V_i + \beta_U \{X_i - \sqrt{\lambda_k}\delta_{bi}^* + (1+\sqrt{\lambda_k})\mu_{\delta^*}\} + \epsilon_i \\ &= \beta_V^t V_i + \beta_U \{X_i - \sqrt{\lambda_k}\delta_{bi}^* + a\} + \epsilon_i, \\ &= \left(V_i, \quad X_i - \sqrt{\lambda_k}\delta_{bi}^* + a\right) \begin{pmatrix} \beta_V \\ \beta_U \end{pmatrix} + \epsilon_i, \end{aligned} \tag{4.8}$$

where $a = (1+\sqrt{\lambda_k})\mu_{\delta^*}$. Using OLS to estimate the parameter in (4.8), we obtain

$$\hat{\beta}_b^{NZM}(\lambda_k) = \left[\begin{pmatrix} \boldsymbol{A} & \boldsymbol{B^{*T}} \\ \boldsymbol{B^*} & \boldsymbol{C^*} \end{pmatrix} \right]^{-1} \begin{pmatrix} k_1 \\ k_2^* \end{pmatrix},$$

where

$$\begin{aligned} A &= \sum V_i^{'} V_i, \\ B^* &= \sum V_i^{'} X_i - \sqrt{\lambda_k} \sum V_i^{'} \delta_{bi}^* + a \sum V_i^{'}, \\ C^* &= \lambda_k \sum \delta_{bi}^{*2} + na^2 - 2\sqrt{\lambda_k} \sum X_i^{'} \delta_{bi} + 2a \sum X_i \\ &- 2a\sqrt{\lambda_k} \sum \delta_{bi}, \\ K_1 &= \sum V_i^{'} Y_i, \\ K_2^* &= \sum X_i^{'} Y_i - \sqrt{\lambda_k} \delta_{bi}^* Y_i + a \sum Y_i. \end{aligned}$$

Equation (4.8) can be expressed as

$$\begin{pmatrix} \boldsymbol{A} & \boldsymbol{B^{*T}} \\ \boldsymbol{B^*} & \boldsymbol{C^*} \end{pmatrix} \begin{pmatrix} \hat{\beta}_V(\lambda_k) \\ \hat{\beta}_U(\lambda_k) \end{pmatrix} = \begin{pmatrix} k_1 \\ k_2^* \end{pmatrix}$$

$$\text{or, } \boldsymbol{A}\hat{\beta}_V(\lambda_k) + \boldsymbol{B^{*T}}\hat{\beta}_U(\lambda_k) = k_1$$

$$\boldsymbol{B^*}\hat{\beta}_V(\lambda_k) + \boldsymbol{C^*}\hat{\beta}_U(\lambda_k) = k_2^*.$$

Solving this system of equations, we obtain the following parameters estimates:

$$\hat{\beta}_V^{NZM}(\lambda_k) = \boldsymbol{A}^{-1}k_1 - \frac{\boldsymbol{A}^{-1}\boldsymbol{B^*}k_2^* - \boldsymbol{A}^{-1}\boldsymbol{B^*}\boldsymbol{B^{*'}}\boldsymbol{A}^{-1}k_1}{\boldsymbol{C^*} - \boldsymbol{B^{*'}}\boldsymbol{A}^{-1}\boldsymbol{B^*}}$$

and

$$\hat{\beta}_U^{NZM}(\lambda_k) = \frac{k_2^* - \boldsymbol{B^{*'}}\boldsymbol{A}^{-1}k_1}{\boldsymbol{C^*} - \boldsymbol{B^{*'}}\boldsymbol{A}^{-1}\boldsymbol{B^*}} = \frac{g_1(\sqrt{\lambda_k}) - g_2(\sqrt{\lambda_k})}{g_3(\lambda_k) - g_4(\sqrt{\lambda_k})}.$$

Thus, we see that the components of $\hat{\beta}^{NZM}(\lambda_k)$ are non-linear functions of λ_k. □

The complex dependence of the NZM-SIMEX estimator on λ_k suggests that the estimator may be sensitive to the choice of extrapolating function. We explore this in a comprehensive series of simulations in the section that follows.

4.6.2 Appendix B: Details of Simulation Study

4.6.2.1 Design of the Simulation Study

Table 4–2: Simulation scenarios

Scenario	Distribution of (U,V)	True δ^*	Y	Assumed δ_b^*
1	$N\left\{\begin{pmatrix}0\\0\end{pmatrix},\begin{pmatrix}1 & 0.447\\0.447 & 1\end{pmatrix}\right\}$	$\delta^* = \lvert\delta\rvert$ and $\delta \sim N(0, 0.25)$	$N(\eta_1^a, 1)$	$\delta_b^* = \lvert\delta_b\rvert$ and $\delta_b \sim N(0, 0.25)$
2	$N\left\{\begin{pmatrix}0\\0\end{pmatrix},\begin{pmatrix}1 & 0.447\\0.447 & 1\end{pmatrix}\right\}$	$\delta^* = \lvert\delta\rvert$ and $\delta \sim N(0, 0.5)$	$N(\eta_1^a, 1)$	$\delta_b^* = \lvert\delta_b\rvert$ and $\delta_b \sim N(0, 0.5)$
3	$N\left\{\begin{pmatrix}0\\0\end{pmatrix},\begin{pmatrix}1 & 0.447\\0.447 & 1\end{pmatrix}\right\}$	$\delta^* = \lvert\delta\rvert$ and $\delta \sim N(0, 1)$	$N(\eta_1^a, 1)$	$\delta_b^* = \lvert\delta_b\rvert$ and $\delta_b \sim N(0, 1)$
4	$N\left\{\begin{pmatrix}0\\0\end{pmatrix},\begin{pmatrix}1 & 0.447\\0.447 & 1\end{pmatrix}\right\}$	$\delta^* = \lvert\delta\rvert$ and $\delta \sim N(0, 2)$	$N(\eta_1^a, 1)$	$\delta_b^* = \lvert\delta_b\rvert$ and $\delta_b \sim N(0, 2)$
5	$U \sim P(12), V \sim N(0, 1)$	$\delta^* \sim P(1.5)$	$N(\eta_2^b, 1)$	$\delta_b^* \sim P(1.5)$
6	$U \sim P(12), V \sim N(0, 1)$	$\delta^* \sim P(3)$	$N(\eta_2^b, 1)$	$\delta_b^* \sim P(3)$
7	$N\left\{\begin{pmatrix}0\\0\end{pmatrix},\begin{pmatrix}1 & 0.447\\0.447 & 1\end{pmatrix}\right\}$	$\delta \sim N(0, 0.5)$	$N(\eta_1^a, 1)$	$\delta_b \sim N(0, 0.25)$
8	$N\left\{\begin{pmatrix}0\\0\end{pmatrix},\begin{pmatrix}1 & 0.447\\0.447 & 1\end{pmatrix}\right\}$	$\delta \sim N(0, 0.5)$	$N(\eta_1^a, 1)$	$\delta_b \sim N(0, 0.75)$
9	$U \sim P(12), V \sim N(0, 1)$	$\delta^* \sim P(1.5)$	$N(\eta_2^b, 1)$	$\delta_b^* \sim P(0.75)$
10	$U \sim P(12), V \sim N(0, 1)$	$\delta^* \sim P(1.5)$	$N(\eta_2^b, 1)$	$\delta_b^* \sim P(2.25)$
11	$U \sim P(12), V \sim N(0, 1)$	$\delta^* \sim P(1.5)$	$P(\exp(\eta_3^c))$	$\delta_b^* \sim P(1.5)$
12	$U \sim P(12), V \sim N(0, 1)$	$\delta^* \sim P(3)$	$P(\exp(\eta_3^c))$	$\delta_b^* \sim P(3)$
13	$U \sim P(12), V \sim N(0, 1)$	$\delta^* \sim P(3)$	$Bernoulli(p^d, 1)$	$\delta_b^* \sim P(3)$

$\eta_1^a = -2 + 1 * U + 0.25 * V + 0.25 * UV$

$\eta_2^b = 1 + 1 * U + 1 * V + 0.5 * UV$

$\eta_3^c = 0.25 + 0.5 * U + 0.05 * V + 0.05 * UV$

$p^d = \frac{\exp(\eta_4)}{1+\exp(\eta_4)}$, where $\eta_4 = -2 + 0.25 * U - 1 * V + 0.25 * UV$

4.6.2.2 Simulation Results

Table 4–3: Simulation results for a continuous outcome and a correctly specified error distribution. SIMEX-Q is the NZM SIMEX with a quadratic fit in the extrapolation step; SIMEX-NL is the NZM SIMEX with a non-linear fit in the extrapolation step.

	Naïve				SIMEX-Q				SIMEX-NL			
	$\hat{\beta}^{OLS}$	Bias	SE	MSE	$\hat{\beta}^{NZM}$	Bias	SE	MSE	$\hat{\beta}^{NZM}$	Bias	SE	MSE
Scenario 1: $Y \sim N(\eta_1, 1)$ and $\delta \sim N(0, 0.25)$. True values of the parameters are $\beta_0 = -2$, $\beta_U = 1$, $\beta_V = 0.25$ and $\beta_{UV} = 0.25$												
$n = 100$												
β_0	-1.631	0.369	0.123	0.151	-1.997	0.003	0.119	0.014	-1.998	0.002	0.119	0.014
β_U	0.900	-0.099	0.121	0.025	0.998	-0.002	0.136	0.018	1.000	0.000	0.136	0.019
β_V	0.389	0.139	0.124	0.035	0.255	0.005	0.121	0.015	0.254	0.004	0.122	0.015
β_{UV}	0.225	-0.025	0.101	0.011	0.243	-0.008	0.110	0.012	0.242	-0.008	0.111	0.012
$n = 500$												
β_0	-1.634	0.366	0.054	0.137	-2.000	-0.000	0.053	0.003	-2.001	-0.001	0.053	0.003
β_U	0.899	-0.101	0.048	0.013	0.996	-0.004	0.054	0.003	0.998	-0.002	0.054	0.003
β_V	0.388	0.138	0.055	0.022	0.252	0.002	0.055	0.003	0.251	0.001	0.055	0.003
β_{UV}	0.232	-0.018	0.041	0.002	0.249	-0.001	0.045	0.002	0.249	-0.001	0.045	0.002
$n = 1000$												
β_0	-1.634	0.366	0.037	0.135	-1.999	0.000	0.037	0.001	-1.999	0.000	0.037	0.001
β_U	0.897	-0.103	0.034	0.012	0.994	-0.006	0.038	0.002	0.997	-0.003	0.039	0.002
β_V	0.389	0.139	0.039	0.021	0.253	0.003	0.038	0.001	0.257	0.002	0.038	0.001
β_{UV}	0.232	-0.018	0.031	0.001	0.249	-0.001	0.033	0.001	0.249	-0.001	0.033	0.001
Scenario 2: $Y \sim N(\eta_1, 1)$ and $\delta \sim N(0, 0.5)$. True values of the parameters are $\beta_0 = -2$, $\beta_U = 1$, $\beta_V = 0.25$ and $\beta_{UV} = 0.25$												
$n = 100$												
β_0	-1.523	0.477	0.132	0.245	-1.997	0.003	0.125	0.016	-1.998	0.002	0.125	0.016
β_U	0.817	-0.183	0.119	0.048	0.979	-0.021	0.147	0.022	0.990	-0.009	0.149	0.023
β_V	0.454	0.204	0.133	0.059	0.264	0.014	0.129	0.017	0.259	0.009	0.131	0.017
β_{UV}	0.209	-0.041	0.101	0.012	0.239	-0.011	0.119	0.014	0.239	-0.010	0.121	0.015
$n = 500$												
β_0	-1.526	0.474	0.058	0.228	-1.999	0.000	0.056	0.003	2.001	-0.001	0.056	0.003
β_U	0.816	-0.184	0.048	0.036	0.976	-0.024	0.058	0.004	0.987	-0.013	0.059	0.004
β_V	0.455	0.205	0.059	0.045	0.261	0.011	0.058	0.004	0.256	0.006	0.058	0.003
β_{UV}	0.217	-0.033	0.042	0.003	0.247	-0.003	0.048	0.002	0.248	-0.002	0.049	0.002
$n = 1000$												
β_0	-1.526	0.474	0.039	0.226	-1.999	0.001	0.039	0.001	-1.999	0.001	0.039	0.002
β_U	0.814	-0.186	0.034	0.036	0.974	-0.026	0.041	0.002	0.985	-0.015	0.042	0.002
β_V	0.456	0.206	0.041	0.044	0.262	0.012	0.039	0.002	0.257	0.007	0.040	0.002
β_{UV}	0.217	-0.033	0.031	0.002	0.247	-0.003	0.036	0.001	0.248	-0.002	0.036	0.001

continued

Table 4–3: (cont.) Simulation results for a continuous outcome and a correctly specified measurement error distribution. SIMEX-Q is the NZM SIMEX with a quadratic fit in the extrapolation step; SIMEX-NL is the NZM SIMEX with a non-linear fit in the extrapolation step.

	Naïve				SIMEX-Q				SIMEX-NL			
	$\hat{\beta}^{OLS}$	Bias	SE	MSE	$\hat{\beta}^{NZM}$	Bias	SE	MSE	$\hat{\beta}^{NZM}$	Bias	SE	MSE
Scenario 3: $Y \sim N(\eta_1, 1)$ and $\delta \sim N(0, 1)$. True values of the parameters are $\beta_0 = -2$, $\beta_U = 1$, $\beta_V = 0.25$ and $\beta_{UV} = 0.25$												
$n = 100$												
β_0	-1.424	0.576	0.146	0.353	-1.994	0.006	0.134	0.018	-1.997	0.003	0.136	0.019
β_U	0.690	-0.309	0.115	0.109	0.914	-0.087	0.158	0.033	0.947	-0.053	0.168	0.031
β_V	0.539	0.289	0.147	0.105	0.293	0.043	0.139	0.021	0.278	0.028	0.144	0.022
β_{UV}	0.183	-0.067	0.101	0.015	0.229	-0.021	0.132	0.018	0.232	-0.018	0.137	0.019
$n = 500$												
β_0	-1.425	0.575	0.063	0.334	-1.996	0.004	0.060	0.0034	-1.999	0.001	0.061	0.004
β_U	0.689	-0.311	0.047	0.099	0.911	-0.089	0.064	0.012	0.943	-0.0567	0.067	0.008
β_V	0.542	0.292	0.064	0.089	0.289	0.039	0.063	0.006	0.275	0.025	0.064	0.005
β_{UV}	0.191	-0.059	0.042	0.005	0.239	-0.011	0.053	0.011	0.243	-0.007	0.055	0.003
$n = 1000$												
β_0	-1.426	0.574	0.043	0.331	-1.995	0.005	0.042	0.002	-1.997	0.003	0.042	0.002
β_U	0.687	-0.313	0.033	0.099	0.907	-0.093	0.044	0.011	0.939	-0.061	0.047	0.006
β_V	0.544	0.294	0.045	0.089	0.292	0.042	0.043	0.004	0.277	0.027	0.044	0.003
β_{UV}	0.192	-0.058	0.030	0.004	0.292	0.042	0.043	0.004	0.243	-0.007	0.039	0.003
Scenario 4: $Y \sim N(\eta_1, 1)$ and $\delta \sim N(0, 2)$. True values of the parameters are $\beta_0 = -2$, $\beta_U = 1$, $\beta_V = 0.25$ and $\beta_{UV} = 0.25$.												
$n = 100$												
β_0	-1.367	0.633	0.163	0.428	-1.986	0.014	0.145	0.021	-1.993	0.008	0.151	0.023
β_U	0.527	-0.473	0.105	0.235	0.769	-0.230	0.161	0.079	0.833	-0.167	0.181	0.061
β_V	0.631	0.381	0.167	0.173	0.357	0.107	0.150	0.034	0.328	0.078	0.159	0.031
β_{UV}	0.147	-0.103	0.097	0.020	0.203	-0.047	0.142	0.022	0.212	-0.038	0.155	0.026
$n = 500$												
β_0	-1.366	0.634	0.0699	0.407	-1.986	0.014	0.065	0.004	-1.992	0.008	0.067	0.005
β_U	0.526	-0.474	0.044	0.227	0.768	-0.232	0.066	0.058	0.829	-0.170	0.074	0.035
β_V	0.637	0.387	0.072	0.155	0.353	0.103	0.067	0.015	0.325	0.075	0.071	0.011
β_{UV}	0.155	-0.095	0.040	0.010	0.215	-0.035	0.057	0.005	0.226	-0.024	0.062	0.004
$n = 1000$												
β_0	-1.427	0.5763	0.1458	0.353	-1.994	0.006	0.134	0.018	-1.997	0.003	0.136	0.019
β_U	0.690	-0.309	0.115	0.109	0.914	-0.087	0.158	0.033	0.947	-0.053	0.168	0.031
β_V	0.539	0.289	0.147	0.105	0.293	0.043	0.139	0.021	0.278	0.028	0.144	0.022
β_{UV}	0.183	-0.067	0.101	0.015	0.229	-0.021	0.132	0.018	0.232	-0.018	0.137	0.019

continued

Table 4–3: (cont.) Simulation results for a continuous outcome and a correctly specified measurement error distribution. SIMEX-Q is the NZM SIMEX with a quadratic fit in the extrapolation step; SIMEX-NL is the NZM SIMEX with a non-linear fit in the extrapolation step.

	Naïve				SIMEX-Q				SIMEX-NL			
	$\hat{\beta}^{OLS}$	Bias	SE	MSE	$\hat{\beta}^{NZM}$	Bias	SE	MSE	$\hat{\beta}^{NZM}$	Bias	SE	MSE
Scenario 5: $Y \sim N(\eta_2, 1)$ and $\delta^* \sim P(1.5)$. True values of the parameters are $\beta_0 = 1$, $\beta_U = 1$, $\beta_V = 1$ and $\beta_{UV} = 0.5$.												
$n = 100$												
β_0	3.672	2.672	0.510	7.402	1.076	0.076	0.696	0.489	1.035	0.035	0.712	0.509
β_U	0.889	-0.111	0.045	0.014	0.994	-0.006	0.055	0.003	0.998	-0.002	0.056	0.003
β_V	2.300	1.300	0.628	2.085	0.995	-0.005	0.856	0.733	0.982	-0.018	0.872	0.761
β_{UV}	0.447	-0.053	0.054	0.006	0.500	0.000	0.066	0.004	0.501	0.001	0.067	0.005
$n = 500$												
β_0	3.667	2.667	0.238	7.169	1.076	0.076	0.308	0.101	1.038	0.038	0.310	0.098
β_U	0.889	-0.111	0.021	0.013	0.994	-0.006	0.024	0.001	0.997	-0.003	0.024	0.001
β_V	2.337	1.337	0.273	1.862	1.049	0.048	0.366	0.136	1.029	0.029	0.373	0.139
β_{UV}	0.444	-0.056	0.024	0.004	0.496	-0.004	0.029	0.002	0.487	-0.002	0.029	0.001
$n = 1000$												
β_0	3.658	2.658	0.166	7.093	1.067	0.067	0.220	0.053	1.028	0.028	0.224	0.051
β_U	0.889	-0.111	0.015	0.012	0.994	-0.006	0.017	0.000	0.998	-0.003	0.017	0.000
β_V	2.333	1.334	0.183	1.812	1.038	0.038	0.244	0.061	1.017	0.017	0.249	0.063
β_{UV}	0.444	-0.056	0.016	0.003	0.497	-0.003	0.019	0.000	0.499	-0.002	0.019	0.000
Scenario 6: $Y \sim N(\eta_2, 1)$ and $\delta^* \sim P(3)$. True values of the parameters are $\beta_0 = 1$, $\beta_U = 1$, $\beta_V = 1$ and $\beta_{UV} = 0.5$.												
$n = 33$												
β_0	5.767	4.767	1.088	23.914	1.324	0.324	1.850	3.531	1.144	0.144	1.971	3.908
β_U	0.801	-0.198	0.108	0.051	0.972	-0.027	0.145	0.021	0.987	-0.012	0.155	0.024
β_V	3.307	2.307	1.246	6.879	1.030	0.030	2.165	4.689	0.923	-0.076	2.341	5.490
β_{UV}	0.406	-0.093	0.122	0.023	0.495	-0.004	0.168	0.028	0.503	0.003	0.183	0.033
$n = 100$												
β_0	5.764	4.764	0.518	22.968	1.287	0.287	0.880	0.858	1.119	0.119	0.934	0.887
β_U	0.804	-0.197	0.052	0.041	0.976	-0.024	0.069	0.005	0.989	-0.010	0.074	0.006
β_V	3.344	2.344	0.643	5.908	1.088	0.088	1.108	1.236	1.015	0.015	1.167	1.362
β_{UV}	0.405	-0.095	0.063	0.013	0.492	-0.008	0.086	0.007	0.498	-0.002	0.091	0.008
$n = 500$												
β_0	5.762	4.762	0.241	22.732	1.286	0.286	0.389	0.233	1.125	0.125	0.402	0.178
β_U	0.804	-0.196	0.024	0.039	0.976	-0.024	0.030	0.002	0.989	-0.010	0.031	0.001
β_V	3.383	2.383	0.282	5.756	1.156	0.156	0.477	0.252	1.076	0.076	0.503	0.259
β_{UV}	0.402	-0.098	0.028	0.010	0.487	-0.013	0.037	0.002	0.494	-0.007	0.039	0.002
$n = 1000$												
β_0	5.749	4.749	0.168	22.579	1.269	0.269	0.276	0.149	1.106	0.106	0.288	0.094
β_U	0.805	-0.195	0.017	0.038	0.977	-0.023	0.022	0.001	0.990	-0.009	0.023	0.001
β_V	3.379	2.379	0.193	5.701	1.143	0.143	0.321	0.123	1.061	0.061	0.334	0.116
β_{UV}	0.402	-0.098	0.019	0.010	0.488	-0.012	0.025	0.001	0.495	-0.006	0.026	0.001

Table 4–4: Simulation results for a continuous outcome and a mis-specified measurement error distribution. SIMEX-Q is the NZM SIMEX with a quadratic fit in the extrapolation step; SIMEX-NL is the NZM SIMEX with a non-linear fit in the extrapolation step.

	Naïve				SIMEX-Q				SIMEX-NL			
	$\hat{\beta}^{OLS}$	Bias	SE	MSE	$\hat{\beta}^{NZM}$	Bias	SE	MSE	$\hat{\beta}^{NZM}$	Bias	SE	MSE
Scenario 7: $Y \sim N(\eta_1, 1)$, $\delta \sim N(0, 0.5)$ and $\delta_b \sim N(0, 0.25)$. True values of the parameters are $\beta_0 = -2$, $\beta_U = 1$, $\beta_V = 0.25$ and $\beta_{UV} = 0.25$.												
$n = 100$												
β_0	-1.523	0.476	0.132	0.245	-1.842	0.157	0.121	0.039	-1.842	0.157	0.121	0.039
β_U	0.817	-0.182	0.119	0.047	0.897	-0.102	0.132	0.028	0.899	-0.100	0.133	0.027
β_V	0.454	0.204	0.132	0.059	0.337	0.087	0.124	0.023	0.336	0.086	0.125	0.023
β_{UV}	0.209	-0.040	0.101	0.011	0.224	-0.025	0.109	0.012	0.223	-0.026	0.109	0.012
$n = 500$												
β_0	-1.525	0.474	0.057	0.228	-1.844	0.155	0.054	0.027	-1.844	0.155	0.054	0.027
β_U	0.816	-0.183	0.047	0.036	0.895	-0.104	0.052	0.013	0.897	-0.102	0.052	0.013
β_V	0.454	0.204	0.058	0.045	0.334	0.084	0.056	0.010	0.334	0.084	0.056	0.010
β_{UV}	0.216	-0.033	0.041	0.002	0.231	-0.018	0.044	0.002	0.231	-0.018	0.044	0.002
$n = 1000$												
β_0	-1.526	0.473	0.039	0.226	-1.844	0.155	0.037	0.025	-1.844	0.155	0.037	0.025
β_U	0.814	-0.185	0.033	0.035	0.893	-0.106	0.037	0.012	0.895	-0.104	0.037	0.012
β_V	0.456	0.206	0.041	0.044	0.336	0.086	0.038	0.008	0.335	0.085	0.038	0.008
β_{UV}	0.216	-0.033	0.030	0.002	0.231	-0.018	0.032	0.001	0.231	-0.018	0.032	0.001
Scenario 8: $Y \sim N(\eta_1, 1)$, $\delta \sim N(0, 0.5)$ and $\delta_b \sim N(0, 0.75)$. True values of the parameters are $\beta_0 = -2$, $\beta_U = 1$, $\beta_V = 0.25$ and $\beta_{UV} = 0.25$.												
$n = 100$												
β_0					-2.136	-0.136	0.131	0.036	-2.143	-0.143	0.133	0.038
β_U					1.054	0.054	0.160	0.028	1.084	0.084	0.168	0.035
β_V					0.198	-0.051	0.135	0.021	0.184	-0.065	0.139	0.023
β_{UV}					0.254	0.004	0.128	0.016	0.256	0.006	0.133	0.017
$n = 500$												
β_0					-2.139	-0.139	0.059	0.023	-2.145	-0.145	0.059	0.024
β_U					1.051	0.051	0.063	0.006	1.081	0.081	0.066	0.010
β_V					0.193	-0.056	0.061	0.006	0.179	-0.070	0.062	0.008
β_{UV}					0.262	0.012	0.051	0.002	0.265	0.015	0.053	0.003
$n = 1000$												
β_0					-2.138	-0.138	0.040	0.020	-2.143	-0.143	0.041	0.022
β_U					1.048	0.048	0.044	0.004	1.078	0.078	0.046	0.008
β_V					0.194	-0.055	0.041	0.004	0.181	-0.068	0.042	0.006
β_{UV}					0.261	0.011	0.038	0.001	0.264	0.014	0.039	0.001

continued

Table 4–4: (cont.) Simulation results: Simulation results for a continuous outcome and a misspecified measurement error distribution. SIMEX-Q is the NZM SIMEX with a quadratic fit in the extrapolation step; SIMEX-NL is the NZM SIMEX with a non-linear fit in the extrapolation step.

	Naïve				SIMEX-Q				SIMEX-NL			
	$\hat{\beta}^{OLS}$	Bias	SE	MSE	$\hat{\beta}^{NZM}$	Bias	SE	MSE	$\hat{\beta}^{NZM}$	Bias	SE	MSE
Scenario 9: $Y \sim N(\eta_2, 1)$, $\delta^* \sim P(1.5)$ and $\delta_b^* \sim P(0.75)$. True values of the parameters are $\beta_0 = 1$, $\beta_U = 1$, $\beta_V = 1$ and $\beta_{UV} = 0.5$.												
$n = 100$												
β_0	3.672	2.672	0.510	7.401	2.423	1.423	0.590	2.374	2.417	1.417	0.596	2.363
β_U	0.888	-0.111	0.045	0.014	0.940	-0.059	0.049	0.006	0.941	-0.058	0.049	0.006
β_V	2.300	1.300	0.628	2.085	1.670	0.670	0.729	0.981	1.673	0.673	0.729	0.985
β_{UV}	0.447	-0.052	0.054	0.006	0.473	-0.026	0.059	0.004	0.473	-0.026	0.059	0.004
$n = 500$												
β_0	3.666	2.666	0.237	7.168	2.420	1.420	0.268	2.089	2.415	1.415	0.268	2.076
β_U	0.889	-0.110	0.020	0.012	0.940	-0.059	0.021	0.003	0.941	-0.058	0.021	0.004
β_V	2.336	1.336	0.273	1.861	1.717	0.717	0.313	0.612	1.713	0.713	0.316	0.609
β_{UV}	0.444	-0.055	0.023	0.003	0.469	-0.030	0.025	0.001	0.470	-0.029	0.026	0.001
$n = 1000$												
β_0	3.658	2.658	0.166	7.093	2.411	1.411	0.189	2.028	2.405	1.405	0.190	2.010
β_U	0.889	-0.110	0.014	0.012	0.940	-0.059	0.015	0.003	0.941	-0.058	0.015	0.003
β_V	2.333	1.333	0.183	1.811	1.710	0.710	0.209	0.548	1.705	0.705	0.211	0.542
β_{UV}	0.444	-0.055	0.015	0.003	0.470	-0.029	0.016	0.001	0.470	-0.029	0.017	0.001
Scenario 10: $Y \sim N(\eta_2, 1)$, $\delta^* \sim P(1.5)$ and $\delta_b^* \sim P(2.25)$. True values of the parameters are $\beta_0 = 1$, $\beta_U = 1$, $\beta_V = 1$ and $\beta_{UV} = 0.5$.												
$n = 100$												
β_0					-0.343	-1.343	0.817	2.472	-0.464	-1.464	0.859	2.884
β_U					1.047	0.047	0.061	0.005	1.056	0.056	0.064	0.007
β_V					0.283	-0.716	1.001	1.516	0.232	-0.767	1.048	1.687
β_{UV}					0.526	0.026	0.073	0.006	0.530	0.030	0.077	0.006
$n = 500$												
β_0					-0.341	-1.341	0.354	1.925	-0.458	-1.458	0.364	2.260
β_U					1.046	0.046	0.025	0.002	1.055	0.055	0.026	0.003
β_V					0.343	-0.656	0.424	0.611	0.285	-0.714	0.444	0.708
β_{UV}					0.522	0.022	0.031	0.001	0.526	0.026	0.033	0.001
$n = 1000$												
β_0					-0.351	-1.351	0.255	1.892	-0.470	-1.470	0.265	2.232
β_U					1.047	0.047	0.018	0.002	1.056	0.056	0.019	0.003
β_V					0.328	-0.671	0.284	0.531	0.267	-0.732	0.297	0.625
β_{UV}					0.523	0.023	0.020	0.001	0.528	0.021	0.028	0.001

Table 4–5: Simulation results for a Poisson outcome and a correctly specified measurement error distribution. SIMEX-Q is the NZM SIMEX with a quadratic fit in the extrapolation step; SIMEX-NL is the NZM SIMEX with a non-linear fit in the extrapolation step.

	Naïve				SIMEX-Q				SIMEX-NL			
	$\hat{\beta}^{OLS}$	Bias	SE	MSE	$\hat{\beta}^{NZM}$	Bias	SE	MSE	$\hat{\beta}^{NZM}$	Bias	SE	MSE
Scenario 11: $Y \sim P(\exp(\eta_3))$ and $\delta^* \sim P(1.5)$. True values of the parameters are $\beta_0 = 0.25$, $\beta_U = 0.5$, $\beta_V = 0.05$ and $\beta_{UV} = 0.05$.												
$n = 100$												
β_0	1.258	1.008	0.136	1.036	0.302	0.052	0.247	0.064	0.231	-0.018	0.275	0.076
β_U	0.379	-0.120	0.045	0.016	0.483	-0.016	0.056	0.003	0.501	0.001	0.063	0.004
β_V	0.163	0.113	0.136	0.031	0.060	0.010	0.271	0.073	0.054	0.004	0.306	0.094
β_{UV}	0.036	-0.013	0.047	0.002	0.047	-0.002	0.065	0.004	0.047	-0.002	0.074	0.005
$n = 500$												
β_0	1.240	0.990	0.071	0.985	0.268	0.018	0.125	0.016	0.194	-0.055	0.138	0.022
β_U	0.390	-0.109	0.024	0.012	0.491	-0.008	0.028	0.000	0.510	0.010	0.031	0.001
β_V	0.150	0.100	0.072	0.015	0.040	-0.009	0.138	0.019	0.030	-0.019	0.155	0.024
β_{UV}	0.042	-0.007	0.025	0.000	0.052	0.002	0.032	0.001	0.054	0.004	0.036	0.001
$n = 1000$												
β_0	1.241	0.991	0.050	0.985	0.270	0.020	0.084	0.007	0.198	-0.051	0.094	0.011
β_U	0.390	-0.109	0.017	0.012	0.491	-0.008	0.019	0.000	0.510	0.010	0.021	0.000
β_V	0.153	0.103	0.053	0.013	0.051	0.001	0.101	0.010	0.042	-0.007	0.115	0.013
β_{UV}	0.041	-0.008	0.019	0.000	0.049	-0.000	0.023	0.000	0.051	0.001	0.027	0.000
Scenario 12: $Y \sim P(\exp(\eta_3))$ and $\delta^* \sim P(3)$. True values of the parameters are $\beta_0 = 0.25$, $\beta_U = 0.5$, $\beta_V = 0.05$ and $\beta_{UV} = 0.05$.												
$n = 33$												
β_0	2.695	2.445	1.232	7.502	0.415	0.165	2.369	5.642	0.304	0.054	2.970	8.827
β_U	0.411	-0.088	0.100	0.017	0.487	-0.012	0.154	0.024	0.495	-0.004	0.193	0.037
β_V	0.320	0.270	1.238	1.607	0.059	0.009	1.540	1.454	0.075	0.025	1.3011	1.898
β_{UV}	0.039	-0.010	0.104	0.010	0.048	-0.001	0.169	0.028	0.047	-0.002	0.219	0.048
$n = 100$												
β_0	1.888	1.638	0.105	2.694	0.490	0.240	0.320	0.160	0.351	0.101	0.370	0.147
β_U	0.301	-0.198	0.053	0.042	0.430	-0.069	0.071	0.009	0.466	-0.033	0.082	0.007
β_V	0.229	0.179	0.106	0.043	0.090	0.040	0.351	0.124	0.078	0.028	0.425	0.181
β_{UV}	0.027	-0.022	0.054	0.003	0.040	-0.009	0.082	0.006	0.042	-0.007	0.100	0.010
$n = 500$												
β_0	1.889	1.639	0.051	2.691	0.433	0.183	0.173	0.063	0.288	0.038	0.195	0.039
β_U	0.313	-0.186	0.029	0.035	0.445	-0.054	0.038	0.004	0.483	-0.016	0.043	0.002
β_V	0.226	0.176	0.054	0.034	0.053	0.003	0.177	0.031	0.033	-0.016	0.208	0.043
β_{UV}	0.035	-0.014	0.028	0.001	0.049	-0.001	0.040	0.001	0.053	0.003	0.047	0.002
$n = 1000$												
β_0	1.892	1.642	0.035	2.699	0.435	0.185	0.113	0.047	0.292	0.042	0.123	0.017
β_U	0.313	-0.186	0.020	0.035	0.446	-0.053	0.025	0.003	0.482	-0.017	0.027	0.001
β_V	0.226	0.176	0.039	0.032	0.064	0.014	0.128	0.016	0.049	-0.001	0.148	0.022
β_{UV}	0.034	-0.015	0.022	0.001	0.046	-0.003	0.029	0.001	0.049	-0.001	0.034	0.001

Table 4–6: Simulation results for a Bernoulli outcome and a correctly specified measurement error distribution. SIMEX-Q is the NZM SIMEX with a quadratic fit in the extrapolation step; SIMEX-NL is the NZM SIMEX with a non-linear fit in the extrapolation step.

	Naïve				SIMEX-Q				SIMEX-NL			
	$\hat{\beta}^{OLS}$	Bias	SE	MSE	$\hat{\beta}^{NZM}$	Bias	SE	MSE	$\hat{\beta}^{NZM}$	Bias	SE	MSE
Scenario 13: $Y \sim Bernoulli(p)$ and $\delta^* \sim P(3)$. True values of the parameters are $\beta_0 = -2$, $\beta_U = 0.25$, $\beta_V = -1$ and $\beta_{UV} = 0.25$.												
$n = 100$												
β_0	-0.836	1.164	0.769	1.946	-2.033	-0.033	1.379	1.905	-2.083	-0.083	1.439	2.079
β_U	0.204	-0.045	0.097	0.011	0.259	0.009	0.131	0.017	0.263	0.013	0.137	0.019
β_V	0.236	1.236	0.929	2.391	-0.953	0.048	1.657	2.748	-0.991	0.009	1.724	2.974
β_{UV}	0.204	-0.046	0.122	0.017	0.260	0.010	0.165	0.027	0.264	0.014	0.171	0.029
$n = 500$												
β_0	-0.767	1.233	0.279	1.599	-1.872	0.128	0.483	0.249	-1.916	0.084	0.499	0.256
β_U	0.189	-0.061	0.034	0.005	0.239	-0.011	0.045	0.002	0.243	-0.007	0.046	0.002
β_V	0.224	1.224	0.349	1.621	-0.864	0.136	0.607	0.386	-0.904	0.096	0.625	0.400
β_{UV}	0.187	-0.063	0.045	0.006	0.238	-0.012	0.059	0.004	0.242	-0.008	0.060	0.004
$n = 1000$												
β_0	-0.777	1.223	0.195	1.533	-1.886	0.114	0.337	0.126	-1.930	0.069	0.345	0.124
β_U	0.189	-0.060	0.024	0.004	0.239	-0.010	0.031	0.001	0.244	-0.006	0.032	0.001
β_V	0.228	1.228	0.253	1.571	-0.858	0.142	0.439	0.213	-0.898	0.102	0.448	0.211
β_{UV}	0.186	-0.064	0.031	0.005	0.238	-0.012	0.042	0.002	0.242	-0.008	0.043	0.002

4.6.3 Appendix C: Additional Results

Table 4–7: Results from simple linear regression model. SIMEX-Q is the NZM SIMEX with a quadratic fit in the extrapolation step; SIMEX-NL is the NZM SIMEX with a non-linear fit in the extrapolation step. In the first panel, β_1 indicates the expected difference in age between two groups of men whose cluster size differs by one individual, whereas in the second panel, β_1 expected difference in the number of sex partners associated with a one-person difference in cluster size.

		Naïve			SIMEX-Q			SIMEX-NL		
Parameter	μ^*	$\hat{\beta}^{OLS}$	SE	p-value	$\hat{\beta}^{NZM}$	SE	p-value	$\hat{\beta}^{NZM}$	SE	p-value
Model: Relating age to cluster size										
β_0	3	34.283	2.207	0.000	34.607	2.801	0.000	34.733	2.801	0.000
β_1	3	-0.123	0.141	0.388	-0.119	0.149	0.421	-0.131	0.149	0.381
β_0	1				34.441	2.633	0.000	34.374	2.635	0.000
β_1	1				-0.126	0.153	0.409	-0.120	0.154	0.434
β_0	5				34.979	3.142	0.000	34.905	3.143	0.000
β_1	5				-0.129	0.158	0.415	-0.123	0.158	0.438
β_0	10				5.279	1.645	0.001	5.239	1.659	0.002
β_1	10				0.025	0.077	0.746	0.027	0.078	0.729
Model: Relating number of sex partners to cluster size										
β_0	3	5.552	1.119	0.000	5.524	1.064	0.000	5.505	1.064	0.000
β_1	3	0.023	0.071	0.753	0.023	0.065	0.720	0.025	0.065	0.702
β_0	1				5.449	1.164	0.000	5.466	1.167	0.000
β_1	1				0.025	0.067	0.705	0.024	0.067	0.724
β_0	5				5.349	1.321	0.000	5.415	1.320	0.000
β_1	5				0.028	0.069	0.678	0.024	0.069	0.725
β_0	10				1.673	0.299	0.000	1.673	0.299	0.000
β_1	10				0.004	0.013	0.755	0.004	0.013	0.759

* mean of the measurement error distribution, Poisson(μ)

Table 4–8: Results from simple logistic regression model. SIMEX-Q is the NZM SIMEX with a quadratic fit in the extrapolation step; SIMEX-NL is the NZM SIMEX with a non-linear fit in the extrapolation step. In the top and bottom panels, β_1 represents the difference in the log odds ratio for, respectively, the use of a condom at the last sexual intercourse and having had an HIV last in the last 24 months associated with a one-person difference in cluster size

		Naïve			SIMEX-Q			SIMEX-NL		
Parameter	μ^*	$\hat{\beta}^{OLS}$	SE	p-value	$\hat{\beta}^{NZM}$	SE	p-value	$\hat{\beta}^{NZM}$	SE	p-value
Model: Relating condom use to cluster size										
β_0	3	2.619	0.772	0.001	2.198	1.104	0.012	2.197	1.105	0.012
β_1	3	-0.047	0.036	0.195	-0.013	0.083	0.567	-0.013	0.083	0.559
β_0	1				2.151	0.946	0.005	2.151	0.945	0.005
β_1	1				-0.010	0.083	0.569	-0.010	0.083	0.572
β_0	5				2.229	1.289	0.025	2.229	1.291	0.027
β_1	5				-0.013	0.085	0.564	-0.013	0.085	0.581
β_0	10				2.289	1.739	0.076	2.288	1.746	0.069
β_1	10				-0.014	0.088	0.591	-0.013	0.088	0.565
Model: Relating HIV tests in the last 24 months to cluster size										
β_0	3	1.686	0.732	0.021	1.226	0.654	0.015	1.228	0.655	0.018
β_1	3	0.038	0.067	0.573	0.044	0.054	0.515	0.044	0.054	0.482
β_0	1				1.311	0.564	0.004	1.311	0.564	0.003
β_1	1				0.046	0.057	0.499	0.046	0.057	0.535
β_0	5				1.126	0.781	0.055	1.130	0.781	0.071
β_1	5				0.045	0.058	0.527	0.044	0.058	0.448
β_0	10				0.955	1.082	0.255	0.964	1.095	0.219
β_1	10				0.042	0.058	0.486	0.041	0.059	0.559

* mean of the measurement error distribution, Poisson(μ)

Table 4–9: Results from log-linear model of number of sex partners on cluster size. SIMEX-Q is the NZM SIMEX with a quadratic fit in the extrapolation step; SIMEX-NL is the NZM SIMEX with a non-linear fit in the extrapolation step. β_1 indicates the expected difference the number of sex partners (top panel) and one night sex partners (bottom panel), on the log scale, between two groups of men whose cluster size differs by one individual.

		Naïve			SIMEX-Q			SIMEX-NL		
Parameter	μ^*	$\hat{\beta}^{OLS}$	SE	p-value	$\hat{\beta}^{NZM}$	SE	p-value	$\hat{\beta}^{NZM}$	SE	p-value
Model: Relating number of sex partners on cluster size										
β_0	3	1.716	0.097	0.000	1.709	0.202	0.000	1.700	0.201	0.000
β_1	3	0.004	0.006	0.526	0.003	0.011	0.752	0.004	0.011	0.702
β_0	1				1.714	0.184	0.000	1.709	0.184	0.000
β_1	1				0.004	0.011	0.737	0.004	0.011	0.711
β_0	5				1.692	0.224	0.000	1.705	0.224	0.000
β_1	5				0.004	0.011	0.707	0.003	0.011	0.762
β_0	10				1.659	0.275	0.000	1.659	0.278	0.000
β_1	10				0.005	0.012	0.704	0.005	0.013	0.706
Model: Relating number of one night partners on cluster size										
β_0	3	1.411	0.112	0.000	1.399	0.296	0.000	1.389	0.296	0.000
β_1	3	0.004	0.006	0.568	0.004	0.015	0.797	0.004	0.015	0.761
β_0	1				1.407	0.267	0.000	1.403	0.267	0.000
β_1	1				0.004	0.014	0.787	0.004	0.014	0.767
β_0	5				1.390	0.314	0.000	1.391	0.313	0.000
β_1	5				0.004	0.016	0.801	0.003	0.016	0.803
β_0	10				1.354	0.408	0.001	1.345	0.414	0.001
β_1	10				0.005	0.018	0.789	0.005	0.018	0.778

* mean of the measurement error distribution, Poisson(μ)

Table 4–10: Results from multinomial model of number of one night partners on cluster size. SIMEX-Q is the NZM SIMEX with a quadratic fit in the extrapolation step; SIMEX-NL is the NZM SIMEX with a non-linear fit in the extrapolation step. $\beta_{1(2-4)}$ indicates the expected difference in the log odds of having 2-4 one night partners between two groups of men whose cluster size differs by one individual; $\beta_{1(5+)}$ is the expected difference in the log odds of having at least 5 one night partners between two groups of men whose cluster size differs by one individual.

		Naïve			SIMEX-Q			SIMEX-NL		
Parameter	μ^*	$\hat{\beta}^{OLS}$	SE	p-value	$\hat{\beta}^{NZM}$	SE	p-value	$\hat{\beta}^{NZM}$	SE	p-value
Model: Relating number of one night partners on cluster size										
$\beta_{0(2-4)}$	3	-1.605	0.780	0.039	-1.727	15.143	0.909	-1.740	31.497	0.955
$\beta_{1(2-4)}$	3	0.036	0.051	0.474	0.037	0.573	0.948	0.038	1.179	0.974
$\beta_{0(5+)}$	3	-0.302	0.522	0.562	-0.427	0.760	0.574	-0.441	0.762	0.562
$\beta_{1(5+)}$	3	0.038	0.039	0.331	0.039	0.058	0.499	0.039	0.057	0.488
$\beta_{0(2-4)}$	1				-1.660	12.812	0.897	-1.630	22.317	0.942
$\beta_{1(2-4)}$	1				0.038	0.524	0.942	0.036	0.903	0.968
$\beta_{0(5+)}$	1				-0.347	0.660	0.599	-0.354	0.660	0.591
$\beta_{1(5+)}$	1				0.038	0.055	0.483	0.039	0.055	0.477
$\beta_{0(2-4)}$	5				-1.792	16.775	0.915	-1.817	38.775	0.963
$\beta_{1(2-4)}$	5				0.036	0.591	0.950	0.039	1.351	0.977
$\beta_{0(5+)}$	5				-0.499	0.879	0.569	-0.501	0.881	0.569
$\beta_{1(5+)}$	5				0.038	0.060	0.524	0.038	0.060	0.518
$\beta_{0(2-4)}$	10				-2.012	14.199	0.887	-2.01	34.997	0.954
$\beta_{1(2-4)}$	10				0.039	0.428	0.927	0.038	1.045	0.970
$\beta_{0(5+)}$	10				-0.767	1.182	0.516	-0.746	1.214	0.538
$\beta_{1(5+)}$	10				0.042	0.064	0.511	0.041	0.067	0.544

* mean of the measurement error distribution, Poisson(μ)

Chapter 5
Manuscript II: Correcting Covariate-Dependent Measurement Error with Non-Zero Mean

Preamble

The first manuscript extended the SIMEX method to accommodate errors with non-zero means. While NZM-SIMEX performs very well in reducing the measurement error bias, the analysis of the SPOT performed in the first manuscript suffered from low power due to analyzing data only from the small number of HIV-positive MSM. For HIV-negative MSM, the cluster size is zero and there is no measurement error herein (as their blood does not undergo HIV sequencing). Thus, the measurement error in phylogenetic cluster size clearly depends on the HIV status.

Therefore, to include both HIV-positive and HIV-negative MSM in the analysis, in the second manuscript, I further extend the NZM-SIMEX to the cases where measurement error in a covariate of interest depends on the value of another correctly specified covariate, and call this SIMEX conditional on covariates. I prove the validity of SIMEX-CC approach theoretically and compare it to two other measurement error correction techniques in simulation studies. The SIMEX-CC is then applied to the all MSM in the SPOT data to further study the relationship between phylogenetic cluster size and demographic and behavioural characteristics of MSM.

This article was published in *Statistics in Medicine* in 2017. The references for this article have been merged with the overall thesis bibliography.

Manuscript II: Correcting Covariate-Dependent Measurement Error with Non-Zero Mean

Nabila Parveen[1], Erica E. M. Moodie[1], and Bluma Brenner[2]

[1] McGill University, Department of Epidemiology, Biostatistics and Occupational Health

[2] Lady Davis Research Institute, Montreal, Quebec, Canada

Abstract

There are many settings in which the distribution of error in a mis-measured covariate varies with the value of another covariate. Take, for example, the case of HIV phylogenetic cluster size, large values of which are an indication of rapid HIV transmission. Researchers wish to find behavioral correlates of HIV phylogenetic cluster size, however the distribution of its measurement error depends on the correctly measured variable, HIV status, and does not have a mean of zero. Further, it is not feasible to obtain validation data or repeated measurements. We propose an extension of simulation-extrapolation, an estimation technique for bias reduction in the presence of measurement error that does not require validation data and can accommodate errors whose distribution depends on other, error-free covariates. The proposed extension performs well in simulation, typically exhibiting less bias and variability than either regression calibration or multiple imputation for measurement error. We apply the proposed method to data from the province of Quebec in Canada to examine the association between HIV phylogenetic cluster size and the number of reported sex partners.

Keywords: bias; measurement error; simulation-extrapolation; HIV

5.1 Introduction

It is sometimes the case that the distribution of measurement error in a covariate of interest depends on the value of another variable. For instance, there is evidence that some lab assays exhibit different variability between men and women [121]. Our work is motivated by a setting in which the measurement error distribution depends on an error-free covariate and, in one subpopulation defined by that error-free covariate, the distribution of the measurement error does not have mean zero.

With the increasing availability of genetic sequencing, HIV researchers have made significant progress in discovering clusters (networks) that are defined by the phylogenetic similarity of the HIV RNA of infected members of a population. The SPOT study (`http://www.spotmontreal.com/?lang=en`), based in Montreal, Canada, offers HIV testing to the community of men who have sex with men, and also collects data on socio-demographic and behavioral characteristics through questionnaires. Attention has now turned to the epidemiological data; the hope is that the data may reveal correlates of large cluster size [9, 13, 18]. Large clusters are indicative of rapid HIV transmission; understanding their correlates may help to construct targeted interventions to interrupt the HIV transmissions. Many SPOT participants are HIV-negative; these participants' cluster size is said to be 0, and this is measured without error. However among the HIV-positive participants, the phylogenetic cluster size is systematically undercounted since only those individuals who have been tested within the province of Quebec are used to determine (measure) cluster size.

Several methods have been proposed to deal with measurement error, including regression calibration [108, 122], multiple imputation [21, 22], and simulation-extrapolation (SIMEX) [68]. All but the last of these require either a validation sample or replicate data for some fraction of the observed sample. In the context of undercounting of phylogenetic cluster size due to unobserved (untested) individuals, obtaining a validation sample is both ethically and practically unfeasible as it would require the testing of all members of the population of interest (in our case, all residents of the province of Quebec). In this paper, we extend

the SIMEX procedure to accommodate measurement error distributions that (i) depend on other covariates and (ii) need not have mean zero. In Section 5.2 we briefly review two common measurement errors procedures regression calibration (RC) and multiple imputation for measurement error (MIME), and introduce our extension to the SIMEX, which we call the SIMEX conditional on covariates (SIMEX-CC). The performance of the three approaches is compared in a simulation study in Section 5.3, and SIMEX-CC is applied to the SPOT data in Section 5.4.

5.2 Measurement Error Correction

In modeling the association between a response and covariates, it is typically assumed that covariates are measured without error. However, this is often not the case in practice, due, for example, to reporting errors, inaccurate recall, or a noisy instrument. Whatever the reason, measurement error in covariates is a potentially troublesome problem [123,124]. Let Y denote the response variable and V and Z are perfectly measured covariates. Moreover, there exists another covariate U whose true value in unavailable; instead, an imprecise measure is available to us, which we shall denote X. Here, U is often called the latent predictor and X is called the surrogate variable. Our intention is to relate the response Y with the true predictors U and V, using realizations of X and V. If X is being used instead of U for modelling purposes, this is often called a naïve method. Adopting this approach typically leads to bias parameter estimates and hence inferences can be misleading.

In this work, we are concerned with differential error in the classical error model. The classical additive measurement error model describes the situation in which $X = U + \delta$, where the stochastic error, δ, is assumed to have zero mean and constant variance, and δ is independent of U, V and Y. Note that a multiplicative version of these two models are also occasionally used. An important characteristic of measurement error is whether it is differential or not. Non-differential measurement error occurs when the distribution of the surrogate variable (X) depends only on the true predictor (U) and not on the response variable, whereas in differential measurement error this condition is not satisfied. In our

setting, the distribution of the errors δ will depend on the value of some other covariate, say Z.

To begin, let us consider a simple linear regression model

$$Y_i = \beta_0 + \beta_1 U_i + \beta_2 V_i + \epsilon_i, \tag{5.1}$$

where U_i is the correctly measured variable whose imperfect measure X_i is available. To begin, we can suppose that there are subjects for whom validation data exists, indexed by $i \in \mathcal{I} \subset \{1, ..., n\}$; thus, for these individuals, we can obtain both U_i and X_i. Our goal is to estimate the regression model parameters, β_0, β_1, and β_2, without bias.

5.2.1 Two Common Approaches: Regression Calibration and Multiple Imputation for Measurement Error

Regression calibration [122] is a simple and widely used two-step method for adjusting for measurement error in regression analyses, but is only applicable when a validation sample or repeated measures in the main study are available. In the first step of RC, a regression of U_i on X_i is performed only on those individuals $i \in \mathcal{I}$. Using estimates from the regression of U_i on X_i, $\hat{U}_i$ is computed for all subjects. Then at the second step, Y_i is regressed on $\hat{U}_i$ to obtain estimates of β_0 and β_1. Note that $\hat{U}_i$ is used for all subjects, even those individuals $i \in \mathcal{I}$ for whom the true value of U is available. Standard errors are typically obtained by bootstrap to account for the estimation of parameters in the first step of the procedure, although analytic variance calculations can also be used.

Multiple imputation was originally proposed to address missing data [21]. It is an attractive technique in which each missing value is replaced by a set of ($m > 1$) plausible values, creating m completed datasets. Each imputed dataset is analyzed and results from these analyses are combined to produce a final result (point estimate). Measurement error can, of course, be viewed as a missing data problem: the true, error-free value of U is missing, and multiple imputation for measurement error (MIME) [20, 22, 23, 125] can be applied. Like RC, validation data are needed to predict U for all $i \notin \mathcal{I}$. Unlike in RC, predicted values

for U are not used for individuals who make up the validation sample. Thus, in MIME, m completed datasets are obtained, each is analysed and the results averaged to obtain a single point estimate. Measures of variability can be computed using Rubin's rules [21] by combining the between- and within-imputation variability or using bootstrap.

RC and MIME provide unbiased estimators provided the validation sample is randomly drawn to ensure that the missing data are missing at random, i.e. the probability of a data point being missing (not in the validation sample) is independent of the (unmeasured) true value of U but may depend on measured characteristics.

5.2.2 Proposed Approach: SIMEX Conditional on Covariates

Consider again the simple linear regression model given in equation (5.1), and suppose that the true predictor U_i has mean μ_U and variance σ_U^2. Suppose further that X_i, the imperfect measurement of U_i, can be written as

$$X_i = U_i - \delta_i,$$

where the conditional distribution of the measurement error δ_i given a correctly measured variable Z_i is $P(\delta_i|Z_i = z_i) = f_z$ with finite mean and variance μ_δ and σ_δ^2, respectively. We have chosen to represent the error in this fashion (subtracting δ_i) to mimic our motivating example, in which $X_i \leq U_i$, however our proposed approach could equally have been developed for error of the form $X_i = U_i + \delta_i$. Further suppose that δ_i is independent of Y_i and U_i. For example, in the SPOT data, U_i is the true value of the count variable *cluster size* and Z_i is binary indicator of HIV status of the SPOT participants, taking value 1 if an individual is HIV-positive. Then we can write $P(\delta_i|Z_i = z) = f_z$ for $z \in \{0, 1\}$. In the SPOT data, all HIV-negative patients have cluster size 0, and there is no measurement error in their cluster size. Therefore, for HIV-negative individuals we have f_0 is the density function putting all mass at the value 0 so that $\delta_i = 0$ if $Z_i = 0$. For HIV infected participants, i.e. those with $Z_i = 1$, however, it may be reasonable to assume that f_1 follows a Poisson distribution, since counts are discrete and the error is such that $X_i \leq U_i$. Note that these examples of f_0 and

f_1 are specific to the SPOT example, but that the theory extends to settings where f_0 is not a point mass but rather some other distribution (e.g. one with the same mean as f_1 but a larger variance).

The standard SIMEX procedure, appropriate when measurement error is independent of covariates and had mean zero, proceeds by generating new datasets in which additional, simulated measurement error is added to the imperfectly measured covariate X_i. This allows the researcher to examine the impact of increasing measurement error on naïve estimates, and to estimate the functional relationship between the estimates and the degree of measurement error (controlled by a parameter λ). The analyst then extrapolates the estimated function back to the unobservable setting in which there is no measurement error. We now detail how to perform SIMEX when measurement error is conditional on error-free covariates, and the mean of conditional error distribution may not be zero.

The SIMEX-CC Algorithm

Simulation Step: In the simulation step of SIMEX-CC, both additional, simulated measurement error and a fixed constant are added to the imperfectly measured covariate X_i, to produce simulated covariates $X_{ib}(\lambda_k)$:

$$X_{ib}(\lambda_k) = X_i - \sqrt{\lambda_k} \times \delta_{ib} + (1 + \sqrt{\lambda_k}) \times \mu_\delta, \tag{5.2}$$

where $\mu_\delta = E[\delta_i | Z_i = z_i]$; $b = 1, \ldots, B$; $k = 1, \ldots, K$ and $i = 1, \ldots, n$. The random variables $\{\delta_{ib}\}_{b=1}^{B}$ are drawn from the distribution of $\delta_i | Z_i = z_i$, while the parameter $\lambda_k \geq 0$ control the variance of measurement error which is added to X_i, typically chosen by the analyst to be in the range $0 = \lambda_0 \leq \lambda_1 \leq \cdots \leq \lambda_K = 2$ [58]. Note that extending the largest value, λ_K, to be greater than 2 may not offer significant improvements in the estimation of the extrapolant function, and that computational resources may be better deployed by increasing the number, K, of variance multipliers chosen since extrapolation occurs in the negative range of the real line.

The simulation step creates B datasets, each consisting of the original dependent variable Y_i, all correctly measured covariates, and the 'new' covariate $X_{ib}(\lambda_k)$ for each λ_k. Note that in a typical SIMEX setting where measurement error has mean zero, the artificial covariates are constructed simply as $X_{ib}(\lambda_k) = X_i - \sqrt{\lambda_k} \times \delta_{ib}$. When measurement error does not have mean zero for at least some values of Z, using equation (5.2) ensures that $\mathbb{E}[X_{ib}(\lambda_k)] = \mathbb{E}[U_i]$. The variance of $X_{ib}(\lambda_k)$ is

$$\begin{aligned} \mathbb{V}\left[X_{ib}(\lambda_k)\right] &= \mathbb{V}\left[X_i - \sqrt{\lambda_k} \times \delta_{ib} + (1 + \sqrt{\lambda_k}) \times \mu_\delta\right] \\ &= \mathbb{V}\left[U_i - \delta_{ib} - \sqrt{\lambda_k} \times \delta_{ib}\right] = \sigma_U^2 + (1 + \lambda_k)^2 \sigma_\delta^2 \end{aligned}$$

which increases with the control parameter λ_k. For each λ_k, let $\hat{\beta}_b(\lambda_k)$ denote the vector of naïve estimates obtained by regressing Y on $X_{ib}(\lambda_k)$. Using B estimates for each λ_k, an average estimate can be obtained as

$$\hat{\beta}(\lambda_k) = \frac{1}{B} \sum_{b=1}^{B} \hat{\beta}_b(\lambda_k). \tag{5.3}$$

Extrapolation Step: In the extrapolation step, the vector $\hat{\beta}(\lambda_k)$ is plotted against λ_k for $\lambda_k \geq 0$, and regression techniques are used to fit an extrapolant function. The SIMEX-CC estimator is obtained as the extrapolation of $\hat{\beta}(\lambda_k)$ to the value $\lambda_k = -1$, a setting which would lead to no measurement error and yield $\mathbb{V}\left[X_{ib}(\lambda_k)\right] = \sigma_U^2$. We denote this estimator, $\hat{\beta}(\lambda = -1)$, by $\hat{\beta}^{S-CC}$.

Note that the procedure has been specifically adapted to the SPOT setting, where because the covariate is known to be undercounted, we have focused on the setting where $X_i = U_i - \delta_i$. The setting where $X_i = U_i + \delta_i$ can easily be accommodated with a small change in the simulation step of the SIMEX-CC by taking the simulated covariates to be

$$X_{ib}(\lambda_k) = X_i - \sqrt{\lambda_k} \times \delta_{ib} + (1 - \sqrt{\lambda_k}) \times \mu_\delta,$$

in place of equation (5.1). It then follows that $E[X_{ib}(\lambda_k)] = E[U_i]$ and $\mathbb{V}\left[X_{ib}(\lambda_k)\right] = \sigma_U^2 + (1 + \lambda_k)^2 \sigma_\delta^2$, as desired.

As for the original SIMEX [68], results hold for more general regression problems such as fitting of non-linear regression models [69], generalized linear models [79], generalized linear mixed models [80], quantile regression models [113], and accelerated failure time models [87], but unbiasedness cannot be shown in closed form as estimators for such models are computed via iterative methods. In these settings, the unbiasedness and feasibility of the SIMEX were demonstrated by simulations. Following these authors, we provide a theorem and proof of unbiasedness in the linear regression setting, and demonstrate the performance of the SIMEX-CC in the generalized linear model setting via simulation rather than analytically. The theorem relies on the assumption that the mean and variance of the measurement error are finite and known. No further distributional assumptions on the mis-measured covariates, on the outcome, or on the measurement error beyond its first moments are required.

Theorem

Unbiasedness of the SIMEX-CC estimator

The estimator obtained via the SIMEX conditional on covariates procedure, $\hat{\beta}^{S-CC}$, converges in probability to the true β regression parameter in a linear regression.

Proof of the theorem in the linear regression setting relies on the observation that

$$\hat{\beta}_1(\lambda_k) \xrightarrow{P} \frac{Cov(U,Y)}{\sigma_U^2 + (1+\lambda_k)\sigma_\delta^2}\beta_1 = \left[\frac{\sigma_U^2}{\sigma_U^2 + (1+\lambda_k)\sigma_\delta^2}\right]\beta_1,$$

which equals β_1 when λ_k is taken to be -1; see Appendix D for details. In more general settings, solutions are not available in closed form and demonstration of the validity of the method relies on simulations as well as the heuristic observation at $\lambda_k = -1$, estimates that would correspond to covariates $X_{ib}(\lambda = -1)$ with $\mathbb{E}[X_{ib}(\lambda = -1)] = \mathbb{E}[U_i]$ and $\mathbb{V}[X_{ib}(\lambda = -1)] = \mathbb{V}[U_i]$. See, for example, [79, 113] and [87].

Standard errors of the SIMEX estimators can be obtained either via a sandwich estimator derivation [68] or using a bootstrap procedure [126]. While the latter is more computationally expensive, it may be more accurate in finite samples than the sandwich calculation which relies on asymptotic derivations. We, therefore, adopt a bootstrap procedure to obtain standard errors for the SIMEX-CC as well as the RC and MIME so that all estimators and inference is considered on a equal footing.

5.3 Simulation Study

In order to evaluate the SIMEX-CC and compare its performance to a naïve approach ignoring measurement error, regression calibration, and multiple imputation for measurement error, we carried out a simulation study.

5.3.1 Design of the Simulation Study

Our simulations were broadly designed to mimic the SPOT setting, with a single binary error-free covariate, Z, as the variable on which the error distribution depends. We considered two outcome distributions (Gaussian and Poisson) and two covariates at different sample sizes. A validation sample was selected to perform RC and MIME; a range of measurement error distributions was assumed for SIMEX-CC.

The "conditioning variable" Z was drawn from a Bernoulli with either probability 0.5 or 0.05, the latter being similar to the observed prevalence of HIV in the SPOT data. For Gaussian distributed outcomes, the outcome data were generated according to the mean model

$$E(Y|U,V) \quad = \quad \beta_0 + \beta_U U + \beta_V V + \beta_{UV} UV,$$

where V is the error free covariate, and the error-prone covariate is $X = U - \delta$ where δ is the measurement error whose distribution depends on Z. The simulation settings for (U,V), $\boldsymbol{\beta} = (\beta_0, \beta_U, \beta_V, \beta_{UV})$ and $\delta|Z$ are given in Table 5–1, Scenarios (a) to (d).

Table 5–1: Simulation study design: assumed distributions for data generation and analysis. Outcomes were generated using $\eta_1 = 1 + 1 \times U + 1 \times V + 0.5 \times UV$ (Scenarios (a)-(b)) or $\eta_2 = \exp(0.25 + 0.1 \times U + 0.1 \times V + 0.01 \times UV)$ (Scenarios (c)-(d)).

Scenario	(U, V)	$P(Z = 1)$	True f_1	Y	Assumed f_1
(a)	$U \sim P(12), V \sim N(0, 1)$	0.5	$\delta \mid Z = 1 \sim P(1.5)$	$N(\eta_1, 1)$	$P(0.75)$, $P(1.5)$, or $P(3)$
(b)	$U \sim P(12), V \sim N(0, 1)$	0.05	$\delta \mid Z = 1 \sim P(1.5)$	$N(\eta_1, 1)$	$P(0.75)$, $P(1.5)$, or $P(3)$
(c)	$U \sim P(25), V \sim N(0, 1)$	0.5	$\delta \mid Z = 1 \sim P(5)$	$P(\eta_2)$	$P(3.5)$, $P(5)$, or $P(7)$
(d)	$U \sim P(25), V \sim N(0, 1)$	0.05	$\delta \mid Z = 1 \sim P(5)$	$P(\eta_2)$	$P(3.5)$, $P(5)$, or $P(7)$

For the Poisson distributed outcomes, data were generated according to a log-linear model

$$\log[E(Y|U, V)] \quad = \quad \beta_0 + \beta_U U + \beta_V V + \beta_{UV} UV.$$

The details of the data-generation for (U, V), $\boldsymbol{\beta}$ and $\delta|Z$ are given in Table 5–1, Scenarios (c) and (d).

5.3.2 Analysis of the Simulated Data

For each setting, 1000 simulated datasets were generated. Each dataset was analyzed with no error correction (naïve), as well as error correction by RC, MIME, and SIMEX-CC. Ten percent [127] of each simulated dataset was randomly chosen to serve as the validation sample for RC and MIME.

For each of the three measurement error corrections, analytic choices were required. We explored different modelling options for both RC and MIME. For MIME, we initially considered multivariate imputation by chained equations, an iterative imputation approach in which a series of conditional regression models are fit to each variable with missing data given the other variables in the dataset. We attempted first to use MIME using the package `mice` in `R` with default settings, however performance was unacceptably poor (results not shown). While bias was much lower than that of the naïve estimator, it was still considerable: in some cases, relative bias exceeded 25%. We hypothesize that this was a result of the use

of predictive mean matching as the imputation method. Thus, we settled on a simple linear regression of U on X for both RC and MIME. For RC, $\hat{U}$ was taken to be the predicted value. For MIME, $\hat{U}$ was taken to be the predicted value plus a random noise draw from a Gaussian distribution with variance given by the residual variance in the regression of U on X. Five imputations were used in MIME; a small number appeared adequate for the simple simulation setting (very few covariates) and thus was chosen to reduce computational burden. Note that in more complex real data settings, a larger number of simulations may be preferred if there is significant between-imputation variability.

For SIMEX-CC, we took $\lambda_k \in \{0, \frac{1}{8}, \frac{2}{8}, \ldots, \frac{15}{8}, \frac{16}{8}\}$, set the number of repetitions to $B = 200$, and took $X_b(\lambda_k)$ as in equation (6.1).

For Gaussian and Poisson distributed outcomes, the assumed distribution of the conditional measurement errors are given in the last column of Table 1. For each data generation scenario, we considered SIMEX-CC with the correct measurement error distribution, as well as a case where the error variance (and mean) in the subgroup for whom $Z = 1$ was either too large or too small. We explored the use of both a quadratic and non-linear extrapolant function, and found very little difference between the resulting estimators (differing by at most 0.001; results not shown), and so present results for the quadratic extrapolant function only.

It should be noted that RC and MIME are supplied with different information from SIMEX-CC. In particular, our analyses have not directly incorporated the information that measurement error is known not to have zero-mean in RC and MIME; however, the direct modelling of the true covariate value takes this into account directly. In effect, we wanted to "level the playing field" by allowing each analytic method to have access to some information beyond the analytic sample (either a validation sample *or* information on the error distribution). Thus, SIMEX-CC was "permitted" to use the distribution of the error distribution, while RC and MIME were "permitted" to use validation data. Further, while the validation sample was relatively small, viewing the measurement error as a missing data problem, the fraction of missing data was not unreasonable; for example, for Scenario (a), the average

percentage of missing information (as calculated using the `mitools` package in `R`) ranged from 25-32% for each of the four parameters in the outcome model when $n = 100$. Sample R code for the Guassian outcome setting is provided online as supporting information.

5.3.3 Results of the Simulation Study

Results of the simulation study are shown in Figure 5–1 for the Gaussian outcome (Scenario (a) in Table 5–1) and for the Poisson outcome (Scenario (c) in Table 5–1); full results are given in Appendix E. Performance was measured by bias, empirical mean squared error (MSE) and coverage percentage (CP), where CP was calculated using bootstrap standard errors. For the Gaussian outcome settings, it is evident from the results that the naïve estimator is seriously biased. The SIMEX-CC estimator performs similarly to RC and better than MIME when the measurement error is correctly specified. When the measurement error distribution is misspecified, SIMEX-CC yields bias that is comparable to MIME, performing worse than RC but still superior to the naïve estimator. For Poisson outcome settings, SIMEX-CC performs considerably better than both RC and MIME, even when the measurement error distribution is misspecified. The coverage of the SIMEX-CC was at the nominal level (e.g., CP ranged from 92.3 to 96.9 in Tables 5–4 and 5–5 in Appendix E). Finally, we observe – unsurprisingly – that the impact of measurement error is considerably greater when the prevalence of the variable Z is greater, that is when more individuals have an error-prone measurement of U. We considered additional values for $P(Z = 1)$ in Scenario (a) for $n = 100$; we observed that the impact of this probability on the amount of bias due to measurement error in the naïve estimates varied across parameters (see Figure 5–2 in Appendix E). These results would suggest that the impact of measurement error is likely to be small in the SPOT data, where fewer than 5% of individuals in the sample are HIV-positive.

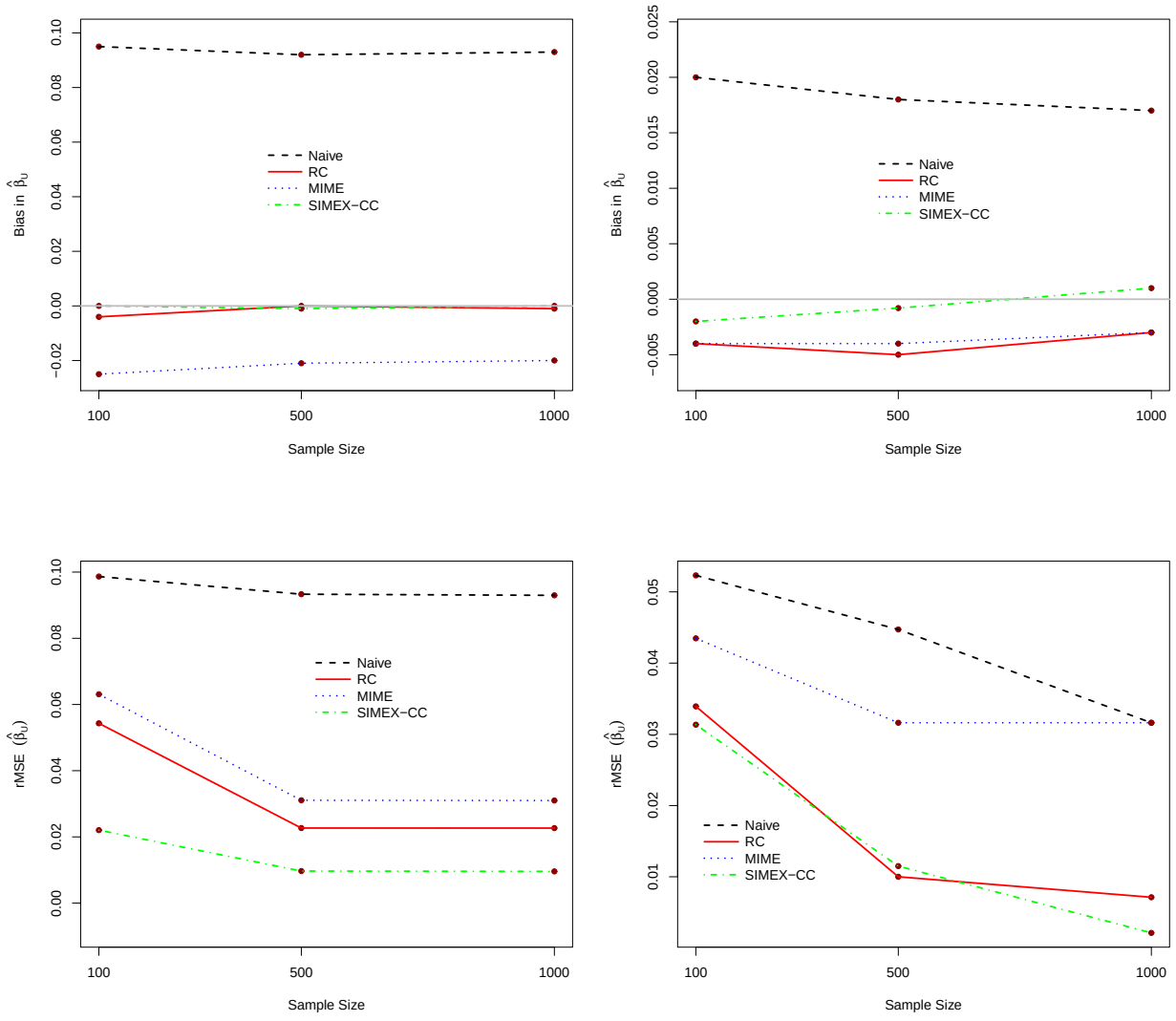


Figure 5–1: Comparison of measurement error correction methods in terms of the bias (top row) and rMSE (bottom row) of a regression parameter estimate as a function of sample size. Panels on the left-hand side provide results for a Gaussian outcome (Scenario (b) in Table 5–1) and panels on the right-hand side for a Poisson outcome (Scenario (d) in Table 5–1).

5.4 Analysis of the SPOT Data

The Montreal SPOT point of care testing site was opened in Montreal's "gay village" neighborhood in 2009. It promotes HIV testing and recruits participants through advertisements in gay magazines and web sites as well as through outreach activities. The site offers rapid, free, and anonymous HIV testing to the community of men who have sex with men, and administers an anonymous questionnaire that elicits information on socio-demographic characteristics, HIV testing behavior, sex life, attitude towards HIV, and socio-sexual profile. Up to April 2012, SPOT had tested and recruited 1803 men. In addition to providing questionnaire data, and all HIV-positive participants' blood underwent HIV RNA sequencing so that phylogenetic clustering could be used to determine the size of the cluster to which the HIV RNA sequence belongs [119]. Selected characteristics for 1803 men who have sex with men participating the SPOT study are given in Table 5–2. Characteristics are fairly similar between the HIV- positive and negative participants. Most HIV-positive participants have HIV RNA that is characterized as belonging to a larger cluster (indicating that at least 3 other individuals in the province of Quebec have an HIV RNA sequence that is in the same phylogenetic cluster).

We applied naïve regression and SIMEX-CC to determine whether there is any association between cluster size and number of sex partners in the SPOT data, as evidenced by a non-zero regression coefficient in a log-linear regression model. (A number of other variables were also explored; results were similar and thus not shown.) Since it is not possible to obtain a validation sample for the SPOT data, RC and MIME could not be applied. For SIMEX-CC, HIV status was taken to be the conditioning variable upon which the measurement error distribution depends. More specifically, a participant who is HIV-negative has cluster size of zero and there is no measurement error in this value. On the other hand, an HIV-positive participant's cluster size is subject to measurement error that is characterized by an undercounting of the truth. To apply SIMEX-CC, we must supply a distribution for the measurement error in the cluster size of HIV-positive participants. In 2012, the population over age 15 in the province

Table 5–2: Characteristics of the SPOT participants. Statistics shown are mean (standard deviation) for continuous or count variables and number (percentage) for categorical variables.

Variable	HIV-positive ($n = 34$)	HIV-negative ($n = 1769$)	All ($n = 1803$)
Age	33 (9.5)	32.6 (10.5)	32.6 (10.5)
Ethnic origin			
Anglo-Quebec	3 (8.8%)	213 (12.0%)	216 (12%)
Other	31 (91.2%)	1556 (88.0%)	1587 (88.0%)
Cluster size			
0	0	1767 (100%)	
1	10 (29.4%)	0	
2-3	3 (8.8%)	0	
> 3	21 (61.8%)	0	
Income			
< 30000	12 (35.3%)	712 (40.3%)	725 (40.2%)
≥ 30000	17 (50%)	932 (52.7%)	949 (52.6%)
Unwilling to report	5 (14.7%)	124 (7.0%)	129 (7.2%)
Education			
No degree	1 (2.9%)	19 (1.1%)	20 (1.1%)
High school or college	11 (32.4%)	649 (36.7%)	660 (36.6%)
University	18 (52.9%)	1052 (59.5%)	1070 (59.4%)
Other training	4 (11.8%)	49 (2.8%)	52 (2.9%)
No. of sex partner	5.8 (4.7)	5.9 (9.2)	5.9 (9.1)
No. of one night sex partners	4.3 (4.7)	3.8 (8.1)	3.8 (8.1)

of Quebec was $6,802,700$ (www.stat.gouv.qc.ca/statistiques/population-demographie/structure/104.htm), and the incidence of HIV was 7 per $100,000$ (www.inspq.qc.ca/publications/notice.asp?E=p\&NumPubli cation=1706); for additional details, see Appendix E. Thus, the total number of new cases of HIV in Quebec in 2012 is approximately 476. Allowing for 25% of individuals to be unaware of their status, and thus not included in the cluster size measurement suggests that measurement error that follows a distribution with mean 3 would yield an appropriate distribution of values to bring up the total of the cluster measurement values to include those ≈ 159 individuals who may be unaware of their HIV-positive status and thus not represented in the current cluster size measures. As the error is discrete, we make the simplifying assumption that the distribution is Poisson, so that a mean of 3 also informs us of the variance of the error.

Cluster size appears to have no significant association with number of sex partners (Table 5–3), whether accounting for measurement error or not. The similarity in the estimates from the naïve and SIMEX analysis is not surprising given the small proportion of the sample affected by measurement error. As a sensitivity analysis, we also considered measurement error distributions for the HIV- positive participants of Poisson(5) and Poisson(10) and found no meaningful change in the resulting estimates.

5.5 Discussion

In this paper, we have proposed an extension of the SIMEX which can accommodate measurement error that is covariate-dependent and may not have mean zero. While a number of methods are available for correcting measurement error, this approach can be used in settings where it is infeasible to collect validation or replicate data, a feature that is unique amongst frequentist methods, though a similar approach has been considered in a Bayesian framework [128]. Furthermore, SIMEX has other attractive properties: for example, unlike likelihood based approaches, SIMEX does not require any distributional assumptions regarding the mis-measured covariate or the outcome. This functional approach to measurement error offers considerable robustness.

Table 5–3: Naïve regression and SIMEX-CC (using quadratic or non-linear extrapolant function) to assess whether there is any relationship between phylogenetic cluster size and number of sex partners. See Table 5–2 for definitions of categorical variables.

	Naïve			SIMEX-CC quadratic			SIMEX-CC non-linear		
	$\hat{\beta}$	SE	p-value	$\hat{\beta}$	SE	p-value	$\hat{\beta}$	SE	p-value
Outcome: number of sex partners									
Cluster size	0.001	0.004	0.778	<0.001	0.007	0.999	<0.001	0.006	0.969
Age	0.018	0.001	<0.001	0.018	0.003	<0.001	0.018	0.003	<0.001
Not Anglo-Quebec	0.193	0.034	<0.001	0.193	0.102	0.050	0.193	0.102	0.050
Educ: HS/college	-0.015	0.101	0.879	-0.016	0.227	0.942	-0.016	0.227	0.943
Educ: university	-0.020	0.101	0.843	-0.020	0.222	0.924	-0.021	0.222	0.925
Educ: other training	0.127	0.116	0.272	0.128	0.379	0.784	0.128	0.379	0.734
Income $\geq$ 30000	-0.108	0.023	<0.001	-0.108	0.100	0.280	-0.108	0.100	0.280
Income not reported	0.013	0.042	0.752	0.013	0.134	0.919	0.013	0.134	0.920
Outcome: number of one night sex partners									
Cluster size	0.007	0.005	0.126	0.007	0.010	0.473	0.006	0.010	0.508
Age	0.023	0.001	<0.001	0.023	0.004	<0.001	0.023	0.004	<0.001
Not Anglo-Quebec	0.280	0.043	<0.001	0.280	0.140	0.046	0.280	0.140	0.046
Educ: HS/college	0.081	0.131	0.533	0.082	0.412	0.841	0.082	0.412	0.841
Educ: university	0.069	0.130	0.593	0.070	0.408	0.863	0.070	0.408	0.863
Educ: other training	0.193	0.146	0.186	0.193	0.561	0.729	0.194	0.561	0.728
Income $\geq$ 30000	-0.169	0.027	<0.001	-0.169	0.130	0.194	-0.169	0.130	0.194
Income not reported	-0.006	0.049	0.903	-0.006	0.182	0.972	-0.005	0.182	0.972

Our simulation studies suggest that SIMEX-CC performs at least as well as competing approaches when the measurement error distribution is correctly specified, and can even perform well in some instances with a mis-specified error distribution. However, in our simulations we made a deliberate choice to compare competing methods that, by the nature of the approaches, made use of different external information. Specifically, in our simulations, RC and MIME were supplied with validation data which SIMEX-CC did not use; in contrast, SIMEX-CC was supplied with information (sometimes imperfect) about the measurement error distribution. Whether these two different forms of information were in some sense equivalent or equally informative may be debated, and we would welcome further research into a metric for comparing the degrees of information provided by different forms of knowledge. Indeed, even when seeking guidance on an appropriate size of validation sample, we found the literature to be quite sparse.

It could be argued that knowledge of the measurement error mean and variance is often difficult or even infeasible to obtain. While such information is likely to be available for, say, well studied laboratory assays, it may be less readily available for a variety of other measurements – indeed, this was the case in the SPOT analysis. In some instances, external data may be used to inform the choice of mean and variance; we attempted to do so in the SPOT analysis. Because of this concern, we did not insist on perfect knowledge of the error distribution parameters, but rather evaluated the performance of SIMEX-CC under ideal and non-ideal conditions and compared it with the naïve approach, regression calibration, and multiple imputation for measurement error through simulations. Applying SIMEX-CC to the SPOT data, we did not detect any association between phylogenetic cluster size and the number of sex partners, however the number of HIV-positive men in the sample was small indicating that measurement error is unlikely to strongly affect results. The proposed method may nevertheless prove useful in other settings where measurement error is more pronounced, such as in laboratory assays.

5.6 Appendix for Manuscript II

5.6.1 Appendix D: Proof of Theorem

Let us consider the following simple linear regression model

$$Y_i = \beta_0 + \beta_1 U_i + \epsilon_i, \tag{5.4}$$

where true predictor U_i follows $\mathrm{N}(\mu_U, \sigma_U^2)$. We have another variable Z_i which is correctly measured. Suppose X_i is an imperfect measurement of U_i which is defined as

$$X_i = U_i - \delta_i = U_i - \delta_i(Z_i), \tag{5.5}$$

where δ_i depends on the error-free covariate Z_i and δ_i follows any distribution with mean μ_δ and variance σ_δ^2. Also, δ_i is assumed to be independent of U_i and Y_i. Note that under this measurement error specification, X_i is always less than or equal to U_i.

As noted above, B new covariates $X_{ib}(\lambda_k)$ are generated according to equation 5.2 so that the total measurement error variance is then the variance of $X_{ib}(\lambda_k)$, i.e. $\sigma_U^2 + \sigma_\delta^2(1+\lambda_k)^2$. For the b^{th} data set, regressing Y on $X_b(\lambda_k)$ gives the vector of naïve estimates $\hat{\beta}_b^{S-CC}(\lambda_k) = (\hat{\beta}_{0,b}(\lambda_k),\ \hat{\beta}_{1,b}(\lambda_k))^T$ of $\beta_b(\lambda_k)$ found via ordinary least squares (OLS), with the average estimate at each λ_k computed according to equation 5.3. Note that here OLS is used as Y is assumed to be continuous. Other forms of regression can be used for other outcome types.

To study the asymptotic mean of the average estimate of slope and intercept, we substitute (5.5) into (5.4), which gives

$$\begin{aligned} Y_i &= \beta_0 + \beta_1(X_i + \delta_i) + \epsilon_i \\ &= \beta_0 + \beta_1\{X_{ib}(\lambda_k) + \sqrt{\lambda_k}\delta_{ib} - (1+\sqrt{\lambda_k})\mu_\delta + \delta_i\} + \epsilon_i \\ &= \beta_0 + \beta_1 X_{ib}(\lambda_k) + \epsilon_i^*, \end{aligned}$$

where $\epsilon_i^* = \beta_1\{\sqrt{\lambda_k}\delta_{ib} - (1+\sqrt{\lambda_k})\mu_\delta + \delta_i\} + \epsilon_i$. For the b^{th} dataset, the naïve estimator of slope β_1 can be obtained by OLS, which yields

$$\begin{aligned}
\hat{\beta}_{1b}^{S-CC}(\lambda_k) &= \frac{\sum_{i=1}^{n}(X_{ib}-\bar{X}_b)(Y_i-\bar{Y})}{\sum_{i=1}^{n}(X_{ib}-\bar{X}_b)^2} \\
&= \frac{\sum_{i=1}^{n}\left\{(X_i-\bar{X}) - \sqrt{\lambda_k}(\delta_{ib} - \sum_{i=1}^{n}\frac{\delta_{ib}}{n})\right\}(Y_i-\bar{Y})}{\sum_{i=1}^{n}\left\{(X_i-\bar{X}) - \sqrt{\lambda_k}(\delta_{ib} - \sum_{i=1}^{n}\frac{\delta_{ib}}{n})\right\}^2} \\
&= \frac{S_{XY} - \sqrt{\lambda_k}S_{Y\delta_b}}{S_{XX} + \lambda_k S_{\delta_b\delta_b} - 2\sqrt{\lambda_k}S_{X\delta_b}}. \qquad (5.6)
\end{aligned}$$

The naïve estimator of the intercept is

$$\hat{\beta}_{0b}^{S-CC}(\lambda_k) = \bar{Y} - \hat{\beta}_{1b}^{S-CC}(\lambda_k)\bar{X}_b. \qquad (5.7)$$

At each λ_k, the estimators are defined as

$$\hat{\beta}_1^{\,S-CC}(\lambda_k) = E\left[\hat{\beta}_{1b}^{S-CC}(\lambda_k)|\{Y_i, X_i\}_{i=1}^n\right]$$

and

$$\hat{\beta}_0^{\,S-CC}(\lambda_k) = E\left[\hat{\beta}_{0b}^{S-CC}(\lambda_k)|\{Y_i, X_i\}_{i=1}^n\right],$$

where the expectation is in terms of the distribution of $\{\delta_{ib}\}$ only.

It then follows that

$$E\left[\hat{\beta}_1^{\,S-CC}(\lambda_k)\right] = E\left[\hat{\beta}_{1b}^{S-CC}(\lambda_k)\right] \qquad \text{and} \qquad E\left[\hat{\beta}_0^{\,S-CC}(\lambda_k)\right] = E\left[\hat{\beta}_{0b}^{S-CC}(\lambda_k)\right].$$

Using the fact that $S_{XY} \xrightarrow{P} \sigma_{XY}$, $S_{XX} \xrightarrow{P} \sigma_{XX}$, $S_{Y\delta_b} \xrightarrow{P} \sigma_{Y\delta_b}$, $S_{\delta_b\delta_b} \xrightarrow{P} \sigma_{\delta_b\delta_b}$, and $S_{X\delta_b} \xrightarrow{P} \sigma_{X\delta_b}$, we obtain

$$\hat{\beta}_{1b}^{S-CC}(\lambda_k) \xrightarrow{P} \frac{\sigma_{XY} - \sqrt{\lambda_k}\sigma_{Y\delta_b}}{\sigma_{XX} + \lambda_k\sigma_{\delta_b\delta_b} - 2\sqrt{\lambda_k}\sigma_{X\delta_b}}$$

and hence

$$\hat{\beta}_1^{S-CC}(\lambda_k) \xrightarrow{P} \frac{\sigma_{XY} - \sqrt{\lambda_k}\sigma_{Y\delta_b}}{\sigma_{XX} + \lambda_k\sigma_{\delta_b\delta_b} - 2\sqrt{\lambda_k}\sigma_{X\delta_b}}. \tag{5.8}$$

Here,

$$\begin{aligned}
\sigma_{XY} &= Cov(X,Y) = Cov(U,Y), \\
\sigma_{Y\delta_b} &= Cov(Y,\delta_b) = 0, \qquad \text{(as } Y \text{ and } \delta_b \text{ are independent)} \\
\sigma_{XX} &= Var(X) = Var(U+\delta) = \sigma_U^2 + \sigma_\delta^2, \qquad \text{(as } U \text{ and } \delta_b \text{ are independent)} \\
\sigma_{\delta_b\delta_b} &= Var(\delta_b) = \sigma_\delta^2 \\
\text{and} \qquad \sigma_{X\delta_b} &= Cov(X,\delta_b) = Cov(U+\delta,\delta_b) = 0 \qquad \text{(as } U \text{ and } \delta_b \text{ are independent).}
\end{aligned}$$

By substitution into (5.6), we obtain

$$\begin{aligned}
\hat{\beta}_1^{S-CC}(\lambda_k) &\xrightarrow{P} \frac{Cov(U,Y)}{\sigma_U^2 + (1+\lambda_k)\sigma_\delta^2} \\
&= \frac{Cov(U,Y)}{Var(U)} \frac{Var(U)}{\sigma_U^2 + (1+\lambda_k)\sigma_\delta^2} \\
&= \beta_1 \left[\frac{\sigma_U^2}{\sigma_U^2 + (1+\lambda_k)\sigma_\delta^2} \right].
\end{aligned}$$

Hence,

$$\lim_{\lambda_k \to -1} \text{plim}\hat{\beta}_1^{S-CC}(\lambda_k) = \beta_1.$$

Similarly, considering (5.7), it can be shown that

$$\lim_{\lambda_k \to -1} \text{plim}\hat{\beta}_0^{S-CC}(\lambda_k) = \beta_0.$$

5.6.2 Appendix E: Additional Results

5.6.2.1 Additional Numerical Results

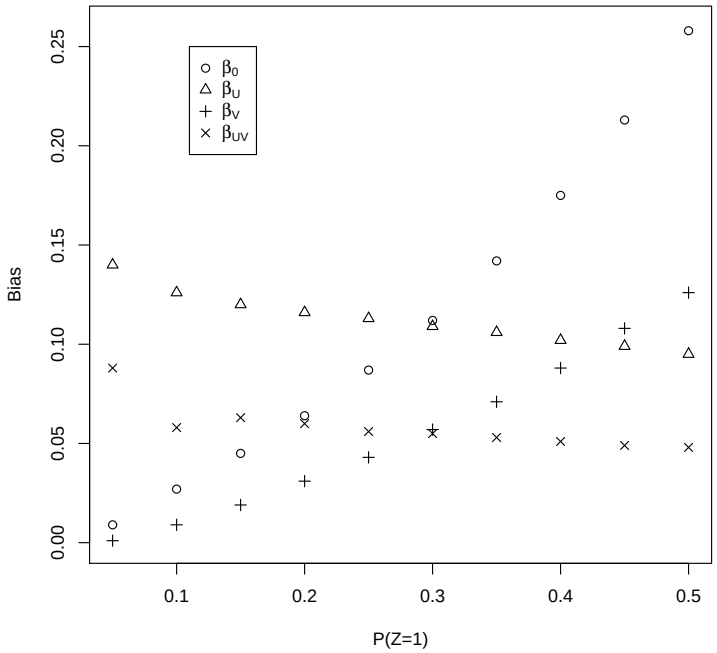


Figure 5–2: Bias in the naïve estimates as the proportion of the covariate subject to measurement error increases (Scenario (a), n=100).

Table 5–4: Simulation results for normally distributed outcomes when conditioning variable Z is common ($P(Z = 1) = 0.5$) or rare ($P(Z = 1) = 0.05$), and measurement error is correctly specified, underspecified and overspecified.

	Naïve		RC		MIME		SIMEX-CC correctly specified		SIMEX-CC under-specified		SIMEX-CC over-specified	
True Value	Bias	MSE	Bias	MSE	Bias	MSE	Bias	MSE	Bias	MSE	Bias	MSE
							$\delta \mid Z \sim P(1.5)$		$\delta \mid Z \sim P(1.5)$		$\delta \mid Z \sim P(1.5)$	
							$\delta_b \mid Z \sim P(1.5)$		$\delta_b \mid Z \sim P(0.75)$		$\delta_b \mid Z \sim P(3)$	
$P(Z = 1) = 0.5$, $n = 100$												
$\beta_0 = 1$ 0.260	0.093	82.7	-0.001	0.117	94.3	0.125	0.095	96.9	-0.001	0.023	94.3	0.116	0.037	-0.176	0.054
$\beta_U = 1$ 0.095	0.010	44.0	-0.004	0.003	92.3	-0.025	0.003	96.5	< 0.001	0.001	94.5	0.046	0.003	-0.086	0.008
$\beta_V = 1$ 0.126	0.048	93.9	-0.002	0.056	97.9	0.059	0.058	97.6	-0.003	0.025	94.7	0.055	0.029	-0.090	0.031
$\beta_{UV} = 0.5$ 0.048	0.003	55.5	-0.002	0.001	93.5	-0.012	0.001	97.4	-0.001	0.001	93.2	0.023	0.001	-0.040	0.002
$P(Z = 1) = 0.5$, $n = 500$															
$\beta_0 = 1$ 0.260	0.074	64.9	0.004	0.018	91.4	0.132	0.030	91.2	0.003	0.004	95.8	0.090	0.015	-0.127	0.026
$\beta_U = 1$ 0.092	0.009	41.9	< 0.001	0.001	93.7	-0.021	0.001	92.3	-0.001	< 0.001	95.0	0.033	0.002	-0.064	0.006
$\beta_V = 1$ 0.129	0.023	74.5	0.001	0.008	96.5	0.062	0.013	95.6	< 0.001	< 0.001	94.9	0.044	0.007	-0.064	0.011
$\beta_{UV} = 0.5$ 0.046	0.003	56.1	< 0.001	< 0.001	95.5	-0.010	< 0.001	89.1	< 0.001	< 0.001	93.6	0.016	0.001	-0.032	0.002
$P(Z = 1) = 0.5$, $n = 1000$															
$\beta_0 = 1$ 0.258	0.070	70.2	0.004	0.008	92.2	0.132	0.030	92.5	< 0.001	< 0.001	96.9	0.088	0.015	-0.125	0.024
$\beta_U = 1$ 0.093	0.009	39.1	-0.001	0.001	94.1	-0.020	0.001	91.5	< 0.001	< 0.001	95.9	0.033	0.001	-0.062	0.005
$\beta_V = 1$ 0.130	0.020	58.9	0.001	0.004	96.3	0.061	0.010	94.1	< 0.001	< 0.001	95.6	0.041	0.002	-0.062	0.010
$\beta_{UV} = 0.5$ 0.046	0.003	57.0	-0.001	0.001	96.1	-0.010	< 0.001	88.7	< 0.001	< 0.001	96.3	0.015	< 0.001	-0.032	0.002
$P(Z = 1) = 0.05$, $n = 100$															
$\beta_0 = 1$ 0.009	0.010	95.4	-0.007	0.069	95.1	-0.013	0.096	90.0	< 0.001	0.010	94.9	0.005	0.010	-0.006	0.010
$\beta_U = 1$ 0.140	0.095	40.6	-0.008	0.031	91.5	-0.074	0.017	92.8	-0.001	0.042	92.5	0.056	0.058	-0.112	0.033
$\beta_V = 1$ 0.001	0.010	95.8	-0.001	0.092	95.4	-0.001	0.129	97.7	-0.004	0.010	94.5	-0.002	0.011	-0.007	0.011
$\beta_{UV} = 0.5$ 0.088	0.177	65.2	0.016	0.032	91.3	-0.038	0.013	93.2	0.012	0.010	93.7	0.036	0.062	-0.042	0.052
$P(Z = 1) = 0.05$, $n = 500$															
$\beta_0 = 1$ 0.015	0.002	94.4	0.003	0.005	97.4	0.007	0.005	91.5	0.001	0.002	94.1	0.007	0.002	-0.009	0.004
$\beta_U = 1$ 0.117	0.015	35.1	-0.013	0.008	90.1	-0.021	0.006	92.5	-0.001	0.001	93.6	0.055	0.004	-0.102	0.018
$\beta_V = 1$ 0.006	0.002	94.6	-0.001	0.004	97.4	0.001	0.004	95.1	-0.001	0.002	94.2	0.002	0.002	-0.008	0.004
$\beta_{UV} = 0.5$ 0.059	0.005	43.0	-0.006	0.003	90.5	-0.011	0.002	92.0	-0.002	0.001	93.9	0.027	0.002	-0.051	0.020
$P(Z = 1) = 0.05$, $n = 1000$															
$\beta_0 = 1$ 0.014	0.001	93.9	0.002	0.001	96.2	0.007	0.001	91.5	-0.001	0.001	94.8	0.003	0.001	-0.015	0.001
$\beta_U = 1$ 0.114	0.013	25.1	-0.008	0.004	91.3	-0.019	0.004	92.2	-0.001	0.000	93.7	0.054	0.003	-0.095	0.010
$\beta_V = 1$ 0.006	0.001	96.8	0.001	0.001	97.3	0.001	0.001	98.0	-0.001	0.001	96.2	0.002	0.001	-0.005	0.001
$\beta_{UV} = 0.5$ 0.057	0.004	33.1	-0.004	0.001	89.6	-0.009	0.001	90.6	-0.001	0.001	93.5	0.026	0.001	-0.047	0.003

MIME, multiple imputation for measurement error; SIMEX-CC, simulation-extrapolation conditional on covariates; CP, coverage percentage; MSE, mean squared error; RC, regression calibration.

Table 5–5: Simulation results for Poisson distributed outcomes when the conditioning variable Z is common ($P(Z = 1) = 0.5$) or rare ($P(Z = 1) = 0.05$), and measurement error is correctly specified, underspecified, and overspecified.

	Naïve			RC			MIME			SIMEX-CC correctly specified			SIMEX-CC under-specified		SIMEX-CC over-specified	
True Value	Bias	MSE	CP	Bias	MSE	CP	Bias	MSE	CP	Bias	MSE	CP	Bias	MSE	Bias	MSE
										$\delta\|Z \sim P(5)$ $\delta_b\|Z \sim P(5)$			$\delta\|Z \sim P(5)$ $\delta_b\|Z \sim P(3.5)$		$\delta\|Z \sim P(5)$ $\delta_b\|Z \sim P(7)$	
$P(Z = 1) = 0.5$, $n = 100$																
$\beta_0 = 0.25$	0.327	0.109	58.4	0.259	0.070	81.8	0.374	0.209	79.9	-0.009	0.015	95.7	0.081	0.021	-0.114	0.032
$\beta_U = 0.10$	0.006	<0.001	77.8	-0.007	<0.001	80.4	-0.008	<0.001	81.6	<0.001	<0.001	95.7	0.002	<0.001	<0.001	<0.001
$\beta_V = 0.10$	0.071	0.014	90.1	0.027	0.016	95.6	0.085	0.049	94.2	<0.001	0.016	95.6	0.020	0.017	-0.022	0.023
$\beta_{UV} = 0.01$	<0.001	<0.001	88.8	-0.001	<0.001	92.7	-0.011	0.001	85.1	<0.001	<0.001	95.1	<0.001	<0.001	<0.001	<0.001
$P(Z = 1) = 0.5$, $n = 500$																
$\beta_0 = 0.25$	0.326	0.107	57.9	0.258	0.070	82.5	0.351	0.129	72.5	0.006	0.003	93.4	0.096	0.012	-0.106	0.015
$\beta_U = 0.10$	0.005	<0.001	79.6	-0.011	<0.001	71.4	-0.006	<0.001	84.2	<0.001	<0.001	95.1	0.002	<0.001	-0.002	<0.001
$\beta_V = 0.10$	0.070	0.008	80.1	0.063	0.007	85.8	0.075	0.019	87.7	0.002	<0.001	94.3	0.024	0.004	-0.027	0.005
$\beta_{UV} = 0.01$	-0.001	<0.001	82.5	-0.002	<0.001	84.7	-0.008	<0.001	81.0	<0.001	<0.001	94.6	<0.001	<0.001	<0.001	<0.001
$P(Z = 1) = 0.5$, $n = 1000$																
$\beta_0 = 0.25$	0.324	0.107	58.5	0.258	0.070	81.9	0.342	0.125	75.6	0.006	0.001	93.6	0.096	0.012	-0.106	0.013
$\beta_U = 0.10$	0.005	<0.001	80.2	-0.011	<0.001	70.3	-0.005	<0.001	86.1	<0.001	<0.001	95.0	0.001	<0.001	-0.002	0.000
$\beta_V = 0.10$	0.067	0.007	81.6	0.063	0.006	84.2	0.070	0.010	89.2	0.002	0.000	94.4	0.024	0.001	-0.025	0.003
$\beta_{UV} = 0.01$	-0.001	<0.001	82.9	-0.002	<0.001	86.1	-0.006	<0.001	87.1	<0.001	<0.001	94.7	<0.001	<0.001	<0.001	<0.001
$P(Z = 1) = 0.05$, $n = 100$																
$\beta_0 = 0.25$	0.009	0.010	92.6	-0.059	0.078	97.6	0.087	0.109	94.0	-0.007	0.009	93.7	-0.003	0.009	-0.004	0.009
$\beta_U = 0.10$	0.020	0.003	43.2	-0.004	0.001	85.1	-0.004	0.002	79.5	-0.002	0.001	92.5	0.004	0.001	-0.009	< 0.001
$\beta_V = 0.10$	0.005	0.009	93.2	0.001	0.094	98.5	0.007	0.008	98.7	0.002	0.009	93.2	0.002	0.009	0.005	0.001
$\beta_{UV} = 0.01$	-0.002	0.007	83.2	-0.001	0.002	92.4	-0.001	0.001	87.9	-0.001	0.002	93.6	-0.002	0.005	-0.002	< 0.001
$P(Z = 1) = 0.05$, $n = 500$																
$\beta_0 = 0.25$	0.026	0.002	71.3	0.019	0.003	93.1	0.019	0.003	82.9	-0.007	0.006	92.3	0.005	0.001	-0.006	0.009
$\beta_U = 0.10$	0.018	0.002	35.0	-0.005	< 0.001	80.2	-0.004	0.001	77.3	-< 0.001	< 0.001	94.3	0.005	< 0.001	-0.009	< 0.001
$\beta_V = 0.10$	0.007	< 0.001	92.7	0.005	< 0.001	95.1	0.005	0.003	81.9	0.002	0.005	92.7	0.002	0.001	0.004	0.001
$\beta_{UV} = 0.01$	0.002	<0.001	70.8	-< 0.001	<0.001	92.3	-0.001	<0.001	86.3	-0.002	0.002	92.8	0.001	<0.001	-0.001	< 0.001
$P(Z = 1) = 0.05$, $n = 1000$																
$\beta_0 = 0.25$	0.014	0.001	79.7	0.003	0.001	91.4	0.019	0.002	84.5	-0.001	< 0.001	95.0	0.003	< 0.001	-0.009	< 0.001
$\beta_U = 0.10$	0.017	0.001	30.0	-0.003	<0.000	83.0	-0.003	0.001	82.1	0.001	<0.001	95.0	0.004	<0.001	-0.005	< 0.001
$\beta_V = 0.10$	0.007	< 0.001	92.5	0.005	< 0.001	95.2	0.005	0.004	83.0	< 0.001	< 0.001	94.0	0.002	< 0.001	-0.003	< 0.001
$\beta_{UV} = 0.01$	0.002	< 0.001	69.9	<0.001	<0.001	92.7	-0.001	<0.000	86.8	<0.001	<0.001	94.0	0.001	< 0.001	-0.001	<0.001

MIME, multiple imputation for measurement error; SIMEX-CC, simulation-extrapolation conditional on covariates; CP, coverage percentage; MSE, mean squared error; RC, regression calibration.

5.6.2.2 Additional Details on the SPOT Analysis

The adult (age > 15) population of Quebec in 2012 was $6,802,700$ with an HIV incidence rate 7 per $100,000$ so that the total number of new cases of HIV in Quebec can be estimated as 6,802,700$\times$0.00007, or approximately 476 people. It has been suggested that in Canada, approximately one quarter of people who are living with HIV are not aware of their seropositive status. Thus we estimate that there are $476 \times 0.25 = 159$ who are not included in the QC genotyping program and therefore contribute to the undercounting of cluster size. Based on previous studies of cluster size distributions and the sizes of clusters in SPOT, we found that a Poisson(3) distribution would be sufficient to yield a distribution of cluster sizes that is similar those found in the literature and would account for the approximately 159 individuals who are estimated to be "missing" from the Quebec genotyping program.

Chapter 6
Manuscript III: New Challenges in HIV Research: Combining Phylogenetic Cluster Size and Epidemiologic Data

Preamble

In this final manuscript (in a sequence of three), I focus on demonstrating methods that can be applied to improve the generalizability of results from a study that is subject to non-probabilistic sampling scheme.

In the first two manuscripts, as the methodological component of the research, I extended the SIMEX method to accommodate errors with non-zero means, in order to apply it to the SPOT data to determine behavioural correlates of cluster size. Unfortunately, results from the SPOT study are limited because SPOT employed a "by convenience" sampling (recruitment) approach. I improve the generalizability of these results by adjusting for the sampling mechanism in SPOT using external information from another study of MSM in Quebec that used a probabilistic venue-based sampling method to recruit participants, the ARGUS study. Thus, in the third manuscript, I apply SIMEX-CC to the SPOT data while adjusting for the SPOT sampling scheme by calculating sampling weights based on common covariates and the venue-based sampling weights for comparable MSM in the ARGUS study.

This manuscript has been submitted for publication. The references for this article have been merged with the overall thesis bibliography.

Manuscript III: New Challenges in HIV Research: Combining Phylogenetic Cluster Size and Epidemiologic Data

Nabila Parveen[1], Erica E. M. Moodie[1], Bluma Brenner[2], Joseph Cox[1,3], and Gilles Lambert[3]

[1] McGill University, Department of Epidemiology, Biostatistics and Occupational Health

[2] Lady Davis Research Institute, Montreal, Quebec, Canada

[3] Prevention and Control of Infectious Diseases Sector, Regional Public Health Department, Centre intégré universitaire de santé et de services sociaux du Centre-Sud-de-l'Ile-de-Montréal

Abstract

An exciting new direction in HIV research is centered on using molecular phylogenetics to understand the social and behavioural drivers of HIV transmission. SPOT was a study evaluating the acceptability of HIV point of care testing offered to men who have sex with men (MSM) at a community-based site in Montreal that also collected data on socio-demographic and behavioural characteristics along with HIV transmission/phylogenetic cluster size. Participant recruitment in SPOT is by convenience sample. Moreover, the phylogenetic cluster size in SPOT is determined through incomplete information due to the absence of a sizable fraction of HIV-infected individuals in the population. Consequently, measurement error occurs in defining the transmission cluster size. In this paper, we use SPOT data to evaluate the association between HIV transmission cluster size and the number of sex partners for MSM, after adjusting for SPOT sampling scheme and correcting for measurement error in cluster size. The sampling weights for SPOT participants were calculated from an external source to fit a weight adjusted model, whereas measurement error was corrected using the simulation-extrapolation conditional on covariates method.

Keywords: HIV transmission cluster; probability sampling; sampling weights; measurement error; SIMEX method.

6.1 Introduction

The HIV pandemic is composed of complex sub-epidemics, each influenced by many biological, behavioral and cultural factors in susceptible populations. The concentrated epidemics in North America have been localized to specific at-risk populations such as men who have sex with men (MSM) and intravenous drug users. While antiretroviral therapy has increased the quality and length of life of individuals infected with HIV, and decreased transmissions [129], MSM remain disproportionately affected in Canada [56, 130–132]: 57% of incident cases of HIV are among MSM [7].

Early stage infection, often defined as within six months of infection, is thought to be a key window for HIV transmission [9,56,133–135], likely due to high concentrations of viremia in bodily fluids. While direct evidence on transmission chains is not measured, phylogenetic analyses have been used to provide insights into transmission networks by clustering individuals based on similarities of the HIV RNA with which they are infected [11–17, 56, 136]. Coupled with epidemiological data, a phylogenetic strategy may provide a unique window to discern HIV transmission, and to attempt to correlate personal characteristics (demographic, behavioral) with large transmission clusters, which are thought indicative of rapid HIV transmission [56, 136]. Our analysis focuses on investigating the behavioural correlates of phylogenetic cluster size using SPOT data. Given the complex and dynamic nature of rapid transmission events, our analysis does not aim to be causal in any sense as we are unable to ensure temporal ordering of some of the epidemiological variables that we consider and infection.

In this paper, we outline some methodological challenges that have arisen in attempting to combine phylogenetic and epidemiologic data, and demonstrate solutions to address these challenges. The first challenge is one of information. HIV-positive tests have been notifiable in Canada since 2004, however reporting is anonymous. Epidemiological data is available through a research questionnaire completed by patients participating in SPOT, a free and anonymous HIV testing service offered to MSM in Montreal, and HIV genotyping is performed

on blood samples of those found to be positive. Thus, the phylogenetic sequences from SPOT participants may be correlated with all sequences from the Quebec HIV Genotyping Cohort to determine the size of the phylogenetic network with which the individual's HIV clusters. The HIV Genotyping Cohort is part of a drug resistance programme, operational since 2002, that includes HIV *pol* sequences. However, as detailed below, the resulting clusters are known to be too small, and so measurement error in the cluster size must be taken into account. Further, while SPOT has a significant research component, the study recruited often by social networks and with the aim of providing HIV testing to sexually active MSM, so that the generalizability of findings from the analysis are uncertain. We will therefore supplement these data with another study of MSM in Montreal whose sampling design was venue-based.

6.2 Methods

We use SPOT data to evaluate the association between HIV transmission cluster size and the number of sex partners for MSM, after adjusting for SPOT sampling scheme and correcting for measurement error in cluster size. Below, we describe in detail our primary data sources, and the methods that we propose to overcome the two major challenges that we encountered in analysing the SPOT data: non-random sampling and measurement error.

6.2.1 Data Sources

6.2.1.1 The SPOT Study

SPOT is an HIV-testing program with a research component targeting MSM in Montreal, Quebec. Beginning in July 2009, SPOT offered free, anonymous, HIV point of care rapid tests to men who have sex with men at a community-based testing site close to Montreal's gay village. Participants were recruited provided they met the inclusion criteria: self-identification as male; at least 18 years of age; resident of Quebec, speaking and understanding French or English; anal sex with another man in the past 12 months; and unknown HIV status at the time of testing. SPOT promotion was also undertaken through outreach

activities organized by the RÉZO community organization in a range of community and social venues. Previously, it has been reported that a large number of participants learned of SPOT from friends [51].

In addition to free rapid tests, the SPOT project administered an anonymous questionnaire eliciting information on socio-demographic characteristics, HIV testing behaviour, and behavorial/lifestyle information; phylogenetic analyses are undertaken on anyone found to be HIV-positive. We analyze the data from 1803 men recruited up until 2013, 36 of whom were found to have HIV.

Two inferential challenges are encountered in SPOT. The first is one of sampling: A sampling frame is of course not available for the target population. The second challenge faced is one of measurement error. The Genotyping Cohort does not contain all phylogenetic information for all individuals with HIV in the province of Quebec. Individuals may not be included for a variety of reasons including not having been tested (either at all, or only outside of the province) or having viral load less than 400 copies per ml [119]. Consequently, measurement error occurs in defining the transmission cluster size and this measurement error is characterized by a systematic *under-counting* of the true cluster size due to the *absence* of the phylogenetic information on the HIV status or phylogenetic cluster of the individuals in the province [137]. However the under-counting exists only for those men who are HIV-positive; anyone free of HIV has a cluster size of 0. Thus, to obtain the valid inferences in investigating correlates of large phylogenetic clusters, it is necessary to account for the non-probabilistic sampling scheme employed by SPOT, and measurement error in cluster size. We propose doing so through venue-based sampling-type weighting using an *external* source of data in combination with a new method of measurement error correction designed specifically for settings in which validation data are unavailable and measurement error may depend on another covariate, namely HIV status: simulation-extrapolation conditional on covariates (SIMEX-CC) [138].

6.2.1.2 The ARGUS Study

The data used in this study were collected from the second wave of ARGUS [139], a second generation surveillance study designed to monitor trends in HIV, sexually transmitted infections, and risk behaviors among MSM living in the province of Quebec that occurred in 2008-2009. Participants were recruited using time-location sampling, or venue-based sampling, from venues including saunas, bars, coffee shops, sports and recreational groups where gay men interact, as well as a fixed study site. That is, at each venue (except for the fixed site), individuals were randomly sampled from among those present. Information was collected on the frequency with which such a venue was attended so that individuals could be reweighted according to the inverse of the likelihood of having been sampled. From the 42 sampling locations in the province of Quebec (37 of which were located in Montreal), 1873 individuals were recruited in the period around when the individuals in SPOT were recruited. A self-administered questionnaire was given, and a blood sample was also collected to perform screening tests for HIV, syphilis and Hepatitis C virus. The ARGUS questionnaire focused on a participant's socio-demographic characteristics, structure of his social network, sexual and other lifestyle activities. As ARGUS employed venue-based sampling, the data from ARGUS respondents, when appropriately reweighted, may better capture the target population of MSM in an urban Quebec setting.

6.2.2 Statistical Methods

6.2.2.1 Venue-Based Sampling

Venue-based sampling is one of the commonly used sampling techniques to recruit hard-to-reach populations, including MSM. It is a probabilistic method in which individuals are sampled probabilistically at particular times in fixed venues (e.g., clubs, bars and gyms) [29]. The sampling framework contains venue-day-time (VDT) units that represent the potential totality of venues, days, and times. For instance, a VDT unit may be a specified time period of five hours on a Friday in a particular venue.

Venue-based sampling typically begins with an on-field team interviewing key service providers and members of the target population to identify a range of VDT units in which to locate the members of the target population. The research team then visits the venues, counting the number of individuals present, prepares a list of potentially eligible VDT units, and estimating the population size for each VDT unit. Upon building the sampling frame, the sample is selected in two stages. In the first stage, *venues* are selected as primary sampling units using simple or stratified sampling with probabilities proportional to the estimated size of the population captured in each venue. In the second stage, a sample of *participants* from the selected venues is drawn using systematic (random) sampling [27]. There are many advantages to venue-based sampling, such as (i) it allows the calculation of a selection probability for each individual in the sample; (ii) unlike convenience sampling, it greatly diminishes the arbitrary selection of venues or individuals, and provides a replicable sampling selection method; and (iii) it does not require a comprehensive enumeration of individual members of the target population so long as all members of the population can be assumed to be reached at the sampled venues at different times. Venue-based sampling is not without costs: it requires intensive fieldwork to visit and map VDTs. Moreover, bias or low generalizability can occur if key venues are missed, or members of the target population do not (as assumed) frequent the venues included in the sampling frame.

Venue-based Sampling Weights in the ARGUS Study:

We calculated sampling weights in ARGUS based on the reported frequency of attending the venue from which a man was sampled. E.g., men who were sampled from a café received a weight 60, 15 or 3.75 if they reported visiting cafés where MSM socialize less than 1 time per month, 1-3 times per month, or 1-3 times per week, respectively. These weights were obtained as the inverse of the frequency per day with which the venues were attended, as follows:

- Attending the café <once/month, we took this to be attendance of once every two months, and gave a weight of $[1/(60 \text{ days})]^{-1} = 60$.

- Attending the café 1-3 times/month, we took this to be an attendance of twice per month, and gave a weight of $[2/(30 \text{ days})]^{-1} = 15$.
- Attending the café 1-3 times/week, we took this to be an attendance of 8 times/month, and gave a weight of $[8/(30 \text{ days})]^{-1} = 3.75$.

Weights were calculated similarly for men recruited from each of the social sites included in the VDT sample (bars, saunas, etc.). For the fixed study site, all participants were recruited, and hence the sampling weight for these individuals was 1. Therefore, each participants in the ARGUS study received a weights of 1, 3.75, 15 or 60.

Weight Adjustment in the SPOT Study:

As ARGUS and SPOT recruit from the sample population, we leveraged the information in the venue-based sampling weights from ARGUS to create venue-based sampling weights for SPOT. Specifically, we built a model to estimate venue-based weights for SPOT using the weight from ARGUS study as dependent variable in a regression model which took all common covariates in the SPOT and ARGUS studies as predictors (see Table 6–1). Then, using the covariates in SPOT, we predicted the most probable weight for each SPOT participant. In our primary analysis, we used a multinomial logistic regression; a sensitivity analysis using linear regression to model the venue-based weights. Note that the number of sex partners in ARGUS was counted for the last 6 months whereas it was counted in SPOT for the last 3 months. We, therefore, multiplied the number of sex partners in SPOT by 2.

6.2.2.2 Simulation-Extrapolation Conditional on Covariates

Over the last three decades, several measurement error correction methods including as regression calibration [58, 108], multiple imputation [21, 22], and simulation-extrapolation [58, 68] have been proposed. Each of these approaches, with the exception of simulation-extrapolation, requires either a validation sample or replicate data for some fraction of the observed sample. In the SPOT study, phylogenetic cluster size is under-counted due to unobserved (untested) individuals. Therefore, obtaining a validation sample is both ethically

and practically infeasible as it would require the testing of *all* members of the population of interest (in our case, all MSM resident in the province of Quebec). Under this circumstance, that is in the absence of validation or replicated data, simulation-extrapolation is the most avenue for correcting the undercounting of the cluster size data.

The simulation-extrapolation method developed by Cook and Stefanski [68] is a simulation based technique for estimating and reducing bias due to additive measurement error. The method was further generalized/extended by Carroll et al. [58] and Stefanski and Cook [70]. Simulation-extrapolation is a two-step estimation procedure consisting of a simulation step and an extrapolation step. Estimates are obtained by adding additional measurement error (in known increments) to the mis-measured data in a resampling-like stage, computing estimates from the contaminated data, establishing a trend between these estimates and the variance of the added measurement errors, and extrapolating this trend back to the case of no measurement error. This method requires the knowledge of the distribution of the measurement error (which may be known in cases such as a laboratory assay, or estimated using external or validation data).

Let U_i, $i = 1, \ldots, n$ is the unobserved true explanatory variable; an error-prone version X_i is available. Consider $X_i = U_i + \delta_i$, where $\delta_i \sim N(0, \sigma_\delta^2)$ and δ_i is independent of U_i, Y_i. SIMEX proceeds in two steps. In the first simulation step, artificial measurement error is added to X_i and B new covariates $X_{i,b}(\lambda_k)$ are generated through $X_{i,b}(\lambda_k) = X_i + \sqrt{\lambda_k}\delta_{ib}$, where $b = 1, \ldots, B$; $k = 1, \ldots, K$ and $i = 1, \ldots, n$ for values of λ_k chosen by the analyst and $\{\delta_{i,b}\}_{b=1}^{B}$ are independent computer simulated random numbers from $N(0, \sigma_\delta^2)$. It can be shown that the variance of $X_{i,b}(\lambda_k)$ is $\sigma_U^2 + (1 + \lambda_k)\sigma_\delta^2$, which increases with λ_k. For each λ_k, let $\hat{\beta}_b(\lambda_k)$ denote the vector of naïve estimates obtained by regressing Y on $X_{i,b}(\lambda_k)$. Using B estimates for each λ_k, an average estimate can be obtained as $B^{-1}\sum_{b=1}^{B}\hat{\beta}_b(\lambda_k)$. By regressing $\hat{\beta}_b(\lambda_k)$ on λ_k, and extrapolating back to $\lambda_k = 1$, we find the estimate $\hat{\beta}(-1)$ corresponding to the error $\sigma_U^2 + (1 + \lambda_k)\sigma_\delta^2 = \sigma_U^2$ (i.e., the error free setting). Typically, $\hat{\beta}_b(\lambda_k)$ is regressed on λ_k assuming either a quadratic or a non-linear relationship (e.g., a lowess smoother [140]).

Parveen et al. [138] extended SIMEX to accommodate measurement error distributions that (i) depend on other covariate and (ii) need not to have mean zero, called simulation-extrapolation conditional on covariates (SIMEX-CC). They expressed the imperfect measurement of U_i as $X_i = U_i - \delta_i$, where the conditional distribution of the error δ_i given a correctly measured variable Z_i was $P(\delta_i|Z_i = z_i) = f_z$ with a finite mean (μ_δ) and variance (σ^2_δ). In SIMEX-CC, the simulated covariate $X_{i,b}(\lambda_k)$ was taken to be

$$X_{ib}(\lambda_k) = X_i - \sqrt{\lambda_k} \times \delta_{ib} + (1 + \sqrt{\lambda_k}) \times \mu_\delta, \tag{6.1}$$

where $\mu_\delta = E(\delta_i|Z_i)$; $b = 1, \ldots, B$; $k = 1, \ldots, K$ and $i = 1, \ldots, n$. Note that under this measurement error specification, X_i is always less than or equal to U_i.

In the SPOT data setting, we consider U_i to be the (unobserved) true cluster size and Z_i to be the HIV status of the participants. All HIV-negative participants belong to a cluster size zero, and there is no measurement error in this cluster size. Measurement error was only present in the cluster size of HIV-positive participants. Thus measurement error in cluster size depends on another covariate: the HIV status of the participants and, as such, we applied the SIMEX-CC method.

Thus, to undertake our analysis of the epidemiological correlates of phylogenetic cluster size in the SPOT data, we considered a weighted regression of Y on $X_{i,b}(\lambda_k)$ in the simulation step of SIMEX-CC method, where weights were the estimated venue-based sampling weights estimated with the external information provided from ARGUS, as described above.

6.3 The SPOT Analysis

Methods

In this analysis, our main focus was on the association between the phylogenetic cluster size and the number of sex partners in the SPOT data. We adopted log-linear models to assess whether there was any evidence of a relationship between cluster size on number of sex partners, when adjusting for age, ethnicity, education, and income as potential confounders. Selected characteristics of the study samples from SPOT and ARGUS subjects are presented

in Table 6–1; for a detailed breakdown of key covariates in the SPOT study by HIV status, please see Table 6–3 in Appendix F, where similar distributions of characteristics are observed between the HIV-positive and negative participants. As noted in the previous section, to adjust for the SPOT sampling scheme in the analysis stage, we used external information from the ARGUS study. The distribution of covariates varies between the SPOT and ARGUS studies, with ARGUS participants being more commonly of French-Canadian origin, less likely to hold a university degree, and more likely to have snorted or smoked cocaine.

The sampling weight for ARGUS participants were directly calculated using the frequency of venue attendance. We then use multinomial logistic regression to predict the sampling weights in ARGUS using the variables listed in Table 6–1, which the variables common to both the SPOT and ARGUS studies. The fitted model was then used to estimate sampling weights for the SPOT study participants. In the SPOT study, there were some missing data. For the variables age, ethnicity, cluster size, number of sex partners and number of one night sex partners the number of missing data points were 6, 3, 2, 137 and 50 respectively. The missing (predicted) sampling weights were resulted from the missing data for the predictors (all common variables in the SPOT and ARGUS data) in the prediction model. Since fewer than 8% of data were missing, our primary analysis used complete cases only. We, however, checked the sensitivity of results to this choice by re-analyzing the data following imputation.

To apply SIMEX-CC, we must supply the method with the mean and variance of the measurement error distribution of cluster size of the HIV-positive individuals. We estimate this distribution using the following external data: The adult population of Quebec in the year from which the SPOT data were taken was 6,802,700 [115] with an HIV incidence rate of 7/100,000 [116] so that the total number of incident cases of HIV in Quebec can be estimated as $6,802,700 \times 0.00007 \approx 476$. It has been suggested that in Canada, approximately 25% of people who are living with HIV do not know their infection status [117]. In fact, the proportion of MSM in Montreal who are unaware of their HIV status is likely lower (13%

Table 6–1: Characteristics of the ARGUS and SPOT participants. Statistics shown are mean (standard deviation) for continuous or count variables, and number (percentage) for categorical variables.

Variables	ARGUS ($n = 1873$)	SPOT ($n = 1803$)
Age	40.45 (12.4)	32.6 (10.5)
Ethnic origin		
French-Canadian	1390 (74%)	883 (49%)
English-Canadian	133 (7%)	216 (12%)
Other origin	350 (19%)	704 (39%)
Education		
Completed college or below	968 (52.1%)	680 (37.7%)
Completed university	848 (45.6%)	1070 (59.4%)
Other training	42 (2.3%)	53 (2.9%)
Income in the previous year (before taxes)		
$< 30,000$ CAD	722 (38.9%)	725 (40.2%)
$\geq 30,000$ CAD	1107 (59.6%)	949 (52.6%)
Unwilling to report	29 (1.6%)	129 (7.2%)
Gay or homosexual	1633 (87.9%)	1498 (83.5%)
No. of sex partner[†]	median = 8 (IQR =18)	median = 3 (IQR = 4)
No. of one night sex partners[†]	median = 4 (IQR =16)	median = 1 (IQR = 4)
Used condom at last intercourse	1332 (71.7%)	1343 (74.6%)
Current no. of gay or homosexual friends		
None	50 (2.7%)	53 (3%)
Less than half	545 (29.6%)	667 (37.2%)
Half	478 (25.9%)	591 (33%)
Most	771 (41.8%)	480 (26.8%)
Any previous HIV tests	1667 (89.7%)	1596 (89.1%)
Drugs use[‡]		
Snorted or smoked cocaine		
Never	1418 (76.3%)	1552 (93.4%)
Occasionally	440 (23.7%)	110 (6.6%)
Snorted or smoked heroin		
Never	1839 (99%)	1662 (99.9%)
Occasionally	19 (1.1%)	2 (0.1%)
Snorted ketamine		
Never	1700 (91.5%)	1617 (97.2%)
Occasionally	158 (8.5%)	46 (2.7%)
Snorted crystal meth		
Never	1813 (97.6%)	1637 (98.4%)
Occasionally	45 (2.5%)	26 (1.5%)

[†]In the last 3 months for SPOT, the last 6 for ARGUS

[‡]In the last 6 months for SPOT

in ARGUS), however we opt for the higher bound as previous work has suggested that it is better to overestimate rather than underestimate measurement error variance [138]. Thus, we estimate that there are $476 \times 0.25 \approx 159$ who are "missing" from the genotyping program database, and thus contributing to the under-counting of cluster size. Based on previous studies of cluster size distributions (see, for example, [10, 11]) and the sizes of clusters in SPOT, we found through trial and error that a Poisson(3) distribution was appropriate to yield a distribution of cluster sizes that was similar those found in the literature and would account for the approximately 160 individuals who are estimated to be missing from the Quebec genotyping program.

Results

In Table 6–2, we compare the results of four different models: (i) a naïve unweighted model (neither adjusted for the sampling scheme nor corrected for measurement error); (ii) a naïve weighted model (not corrected for measurement error, but adjusted for the sampling scheme); (iii) an unweighted SIMEX-CC analysis (not adjusted for sampling scheme but corrected for measurement error); and (iv) a weighted SIMEX-CC (adjusted for both the sampling scheme and measurement error). It is evident that cluster size has no association with the total number of sex partners across all models. However, notable differences appear in the analysis of the number of one-night sex partners: cluster size appears to be associated with the number of one night partners when adjusting for sampling weights, however the association is very weak in magnitude across all models and not significant when measurement error is taken into account. In this analysis, adjusting for the sampling confers greater changes in the estimates than correcting for measurement error – perhaps unsurprisingly as there are relatively few individuals living with HIV and hence whose cluster size is subject to measurement error.

To assess the sensitivity of the results to the measurement error distribution, missing data, and the estimation of the sampling weight, we re-analyzed the SPOT data (i) assuming measurement error was distributed with a mean and variance of 5; (ii) considering all missing

variables as the modal (most common) values for binary variables, and median for continuous and count variables; and (iii) modelling the distribution of sampling weights as via linear regression. We found no meaningful changes in the resulting estimates (see Tables 6–4 to 6–6 in Appendix F).

6.4 Discussion

In this paper, we have demonstrated how phylogenetic and epidemiological data may be combined in an effort to leverage new insights into the correlates of HIV phylogenetic cluster size in MSM using two high quality data sources: a rich dataset offering information on both phylogenetic clustering and epidemiological covariates, and another offering similar epidemiological data and yet is one of the few studies in the Canadian MSM population to have used a probabilistic recruitment approach (time-location sampling). To leverage the complementary strengths of these two datasets, we used of considerable external information, adjusting SPOT by assuming it could be viewed through the lens of venue-based sampling and addressing measurement error in cluster size using the SIMEX-CC, which can accommodate measurement error that is covariate-dependent and may not have mean zero. In this case study, we have thus demonstrated important methodological tools that can be employed in a wide array of settings, most particularly in studies of hard-to-reach populations and studies where measurement error can be well-characterized.

Our analysis is subject to several limitations. First, while we have drawn on two large studies of urban MSM in Canada, power is limited due to the small number of people who were HIV-positive and for whom both phylogenetic cluster size and epidemiological data were available. More importantly, "cluster size" is not a static measure, but rather one that evolves with new transmissions and the dynamics that drive large clusters are highly complex and may vary from cluster to cluster. For instance, cluster size may be driven in one cluster by a large number of individuals engaging with a small number of sexual partners within the period of high infectivity shortly after seroconversion, and in driven in another by a small number of individuals with many sexual partners and transmission events. As time

Table 6–2: Results from unweighted and weighted naïve models, and unweighted and weighted SIMEX-CC where the cluster measurement error is assumed to be distribution Poisson(3), to assess the correlation between cluster size and number of sex partners. See Table 6–3 in Appendix F for definitions of categorical variables.

	Naïve Unweighted			Naïve Weighted			SIMEX-CC Unweighted			SIMEX-CC Weighted		
	$\hat{\beta}$	SE	CI	$\hat{\beta}$	SE	CI	$\hat{\beta}$	SE	CI	$\hat{\beta}$	SE	CI
Total Number of Sex Partners												
Cluster size	0.002	0.005	(-0.008, 0.012)	-0.002	0.002	(-0.006, 0.002)	0.002	0.008	(-0.014, 0.018)	-0.001	0.013	(-0.026, 0.024)
Age	0.016	0.001	(0.014, 0.018)	0.004	< 0.001	(0.002, 0.006)	0.016	0.004	(0.008, 0.024)	0.004	0.004	(-0.004, 0.012)
Not English-Canadian†	0.179	0.035	(0.110, 0.248)	0.208	0.021	(0.167, 0.249)	0.179	0.096	(-0.009, 0.367)	0.209	0.099	(0.015, 0.403)
Educ: university	0.016	0.022	(-0.027, 0.059)	-0.040	0.007	(-0.054, -0.026)	0.016	0.081	(-0.143, 0.175)	-0.040	0.250	(-0.530, 0.450)
Educ: other training	0.165	0.066	(0.036, 0.294)	0.009	0.020	(-0.030, 0.048)	0.164	0.352	(-0.526, 0.854)	0.008	0.367	(-0.711, 0.727)
Income $\geq$ 30000	-0.073	0.023	(-0.118, -0.028)	0.186	0.008	(0.170, 0.202)	-0.073	0.093	(-0.255, 0.109)	0.186	0.150	(-0.108, 0.480)
Income not reported	0.034	0.043	(-0.050, 0.118)	0.113	0.014	(0.086, 0.140)	0.034	0.140	(-0.240, 0.308)	0.113	0.114	(-0.110, 0.336)
Total Number of One-night Sex Partners												
Cluster size	0.005	0.005	(-0.005, 0.015)	0.006	0.002	(0.002, 0.010)	0.005	0.011	(-0.017, 0.027)	0.007	0.014	(-0.020, 0.034)
Age	0.022	0.001	(0.020, 0.024)	0.007	<0.001	(0.005, 0.009)	0.022	0.004	(0.014, 0.030)	0.007	0.005	(-0.003, 0.017)
Not English-Canadian†	0.228	0.044	(0.142, 0.314)	0.266	0.027	(0.213, 0.319)	0.228	0.147	(-0.060, 0.516)	0.269	0.144	(-0.013, 0.551)
Educ: university	0.017	0.027	(-0.036, 0.070)	0.046	0.010	(0.026, 0.066)	0.017	0.104	(-0.187,0.221)	0.047	0.230	(-0.404, 0.498)
Educ: other training	0.179	0.079	(0.024, 0.334)	0.036	0.026	(-0.015, 0.087)	0.178	0.528	(-0.857, 1.213)	0.032	0.428	(-0.807, 0.871)
Income $\geq$ 30000	-0.135	0.028	(-0.190, -0.080)	0.125	0.010	(0.105, 0.145)	-0.136	0.120	(-0.371, 0.100)	0.125	0.152	(-0.173, 0.423)
Income not reported	0.026	0.052	(-0.076, 0.128)	0.100	0.017	(0.067, 0.133)	0.026	0.171	(-0.309, 0.361)	0.095	0.144	(-0.187, 0.377)

Abbreviations: SIMEX-CC, Simulation-Extrapolation Conditional on Covariates; SE, Standard Error; CI, Confidence Interval

†Due to uncertainty as to the most true underlying ethnic composition of the MSM community in Quebec we opted to consider English-Canadian versus other to reduce the influence of this variable on the sampling weights.

progresses, the measurement error in cluster size may decrease: with each new HIV-infected person being tested in the province, cluster size may be updated. However, simply adding precision to some *final* measure of a cluster size may not provide the relevant information: ideally, we want to correlate an individual's socio-demographic and lifestyle characteristics *at the time of infection* with the size of the cluster *at that same point in time*. Phylogenetics offer some insights into when an individual was infected, but not with enough precision to accurately determine the size of a cluster at the point at which an particular member of that cluster was infected. Finally, our analysis implicitly assumed that ARGUS had recruit from the same MSM population as SPOT. While demographics are broadly similar in the two populations, ethnic origin did vary, with more SPOT participants being of non-Canadian origin and cocaine use differed, likely due to recruitment from saunas, bars, and sex clubs. ARGUS sought to recruit lifetime MSM, not only those who were currently sexually active. If this is the case that the venues from which ARGUS sampled did not cover the *entire* MSM community then the weighting scheme used will have adjusted SPOT to look more like ARGUS participants, but neither study sample will represent, or generalize to, the entire MSM community in Montreal or the province of Quebec. However, while it is possible that some social venues may not have been identified, the extensive formative research and environmental scan used to map the MSM community is a strength of the ARGUS study that minimizes this as a potential concern.

As in previous work [57], large cluster size has not been found to correlate with sexual risk behavior. This underscores the importance of finding other indicators such as, perhaps, real-time phylogenetic monitoring are needed to identify early stage infection and better understand transmission dynamics among MSM. Clearly, our work only scratches the very surface of the potential links between rapid transmission as hinted at by larger cluster size and individual-level characteristics, but is an important first step in posing the question and offering solutions to some of the methodological hurdles that may be faced.

6.5 Appendix for Manuscript III

Appendix F: Additional Results from the Data Analysis

Table 6–3: SPOT participant characteristics. Statistics shown are mean (standard deviation) for continuous and count variables, and number (percentage) for categorical variables.

Variable	HIV-positive ($n = 34$)	HIV-negative ($n = 1769$)	All ($n = 1803$)
Age	33 (9.5)	32.6 (10.5)	32.6 (10.5)
Ethnic origin			
English-Canadian	3 (8.8%)	213 (12.0%)	216 (12.0%)
Other	31 (91.2%)	1556 (88.0%)	1587 (88.0%)
Cluster size			
0	0	1767 (100%)	
1	10 (29.4%)	0	
2-3	3 (8.8%)	0	
> 3	21 (61.8%)	0	
Income			
< 30000	12 (35.3%)	712 (40.3%)	725 (40.2%)
≥ 30000	17 (50%)	932 (52.7%)	949 (52.6%)
unwilling to report	5 (14.7%)	124 (7.0%)	129 (7.2%)
Education			
No degree	1 (2.9%)	19 (1.1%)	20 (1.1%)
High school or college	11 (32.4%)	649 (36.7%)	660 (36.6%)
University	18 (52.9%)	1052 (59.5%)	1070 (59.4%)
Other training	4 (11.8%)	49 (2.8%)	52 (2.9%)
No. of Sex Partner	5.8 (4.7)	5.9 (9.2)	5.9 (9.1)
No. of One night Sex Partners	4.3 (4.7)	3.8 (8.1)	3.8 (8.1)

Table 6–4: **Sensitivity Analysis 1: Measurement Error Distribution is Poisson(5)**– Summary results from unweighted and weighted naïve regression and unweighted and weighted SIMEX-CC to assess the relationship between phylogenetic cluster size and number of sex partners. See Table 6–3 in Appendix F for definitions of categorical variables.

	Naïve Unweighted			Naïve Weighted			SIMEX-CC Unweighted			SIMEX-CC weighted		
	$\hat{\beta}$	SE	CI	$\hat{\beta}$	SE	CI	$\hat{\beta}$	SE	CI	$\hat{\beta}$	SE	CI
Outcome: Number of Sex Partners												
Cluster size	0.002	0.005	(-0.007, 0.011)	-0.002	0.002	(-0.006, 0.002)	0.001	0.008	(-0.015, 0.017)	-0.001	0.011	(-0.023, 0.021)
Age	0.016	0.001	(0.014, 0.018)	0.004	< 0.001	(0.002, 0.006)	0.016	0.003	(0.010, 0.022)	0.004	0.004	(-0.004, 0.012)
Not Anglo-Quebec	0.179	0.035	(0.110, 0.247)	0.208	0.021	(0.167, 0.249)	0.178	0.099	(-0.016, 0.372)	0.209	0.099	(0.015, 0.403)
Educ: university	0.016	0.022	(-0.027, 0.058)	-0.040	0.007	(-0.054, -0.026)	0.016	0.091	(-0.162, 0.194)	-0.040	0.255	(-0.540, 0.460)
Educ: other training	0.165	0.066	(0.034, 0.294)	0.009	0.020	(-0.030, 0.049)	0.165	0.376	(-0.572, 0.902)	0.007	0.328	(-0.636, 0.650)
Income $\geq$ 30000	-0.073	0.023	(-0.119, -0.028)	0.186	0.008	(0.170, 0.201)	-0.073	0.089	(-0.247, 0.101)	0.186	0.150	(-0.108, 0.480)
Income not reported	0.034	0.043	(-0.050, 0.118)	0.113	0.014	(0.086, 0.140)	0.034	0.135	(-0.231, 0.299)	0.113	0.122	(-0.126, 0.352)
Outcome: Number of One Night Sex Partners												
Cluster size	0.005	0.005	(-0.005, 0.015)	0.006	0.002	(0.002, 0.010)	0.005	0.010	(-0.015, 0.024)	0.008	0.013	(-0.017, 0.033)
Age	0.022	0.001	(0.019, 0.020)	0.007	< 0.001	(0.005, 0.009)	0.022	0.005	(0.012, 0.031)	0.007	0.005	(-0.003, 0.017)
Not Anglo-Quebec	0.228	0.044	(0.142, 0.314)	0.266	0.027	(0.213, 0.320)	0.228	0.139	(-0.044, 0.500)	0.271	0.157	(-0.037, 0.579)
Educ: university	0.017	0.027	(-0.036, 0.070)	0.046	0.010	(0.026, 0.065)	0.017	0.107	(-0.193, 0.227)	0.047	0.220	(-0.384, 0.478)
Educ: other training	0.179	0.079	(0.024, 0.334)	0.036	0.026	(-0.015, 0.087)	0.179	0.510	(-0.821, 1.178)	0.030	0.389	(-0.732, 0.792)
Income $\geq$ 30000	-0.135	0.028	(-0.190, -0.080)	0.125	0.010	(0.105, 0.145)	-0.136	0.121	(-0.373, 0.101)	0.125	0.148	(-0.165, 0.415)
Income not reported	0.026	0.052	(-0.076, 0.128)	0.096	0.017	(0.063, 0.129)	0.026	0.180	(-0.327, 0.379)	0.095	0.154	(-0.207, 0.397)

Abbreviations: SIMEX-CC, Simulation-Extrapolation Conditional on Covariates; SE, Standard Error; CI, Confidence Interval

Table 6–5: **Sensitivity Analysis 2: Imputing Missing Data** – Summary results from unweighted and weighted naïve regression and unweighted and weighted SIMEX-CC to assess the relationship between phylogenetic cluster size and number of sex partners. See Table 6–3 in Appendix F for definitions of categorical variables.

	Naïve Unweighted			Naïve Weighted			SIMEX-CC Unweighted			SIMEX-CC Weighted		
	$\hat{\beta}$	SE	CI	$\hat{\beta}$	SE	CI	$\hat{\beta}$	SE	CI	$\hat{\beta}$	SE	CI
Outcome: Number of Sex Partners												
Cluster size	0.003	0.005	(-0.007, 0.013)	-0.003	0.002	(-0.007, 0.001)	0.002	0.008	(-0.014, 0.018)	-0.002	0.014	(-0.029, 0.025)
Age	0.018	0.001	(0.016, 0.020)	0.005	<0.001	(0.003, 0.007)	0.018	0.004	(0.010, 0.026)	0.005	0.004	(-0.003, 0.013)
Not English-Cdn	0.197	0.034	(0.130, 0.264)	0.225	0.020	(0.186, 0.264)	0.197	0.092	(0.017, 0.377)	0.226	0.086	(0.057, 0.395)
Educ: university	-0.005	0.021	(-0.046, 0.036)	-0.041	0.007	(-0.055, -0.027)	-0.005	0.079	(-0.160, 0.150)	-0.041	0.218	(-0.468, 0.386)
Educ: other training	0.121	0.056	(0.011, 0.231)	0.116	0.017	(0.083, 0.149)	0.121	0.244	(-0.357, 0.599)	0.115	0.256	(-0.387, 0.617)
Income $\geq$ 30000	-0.105	0.023	(-0.150, -0.060)	0.160	0.008	(0.144, 0.176)	-0.105	0.092	(-0.285, 0.075)	0.160	0.129	(-0.093, 0.413)
Income not reported	-0.010	0.040	(-0.088, 0.068)	0.104	0.013	(0.079, 0.129)	-0.010	0.126	(-0.257, 0.237)	0.103	0.114	(-0.120, 0.326)
Outcome: Number of One Night Sex Partners												
Cluster size	0.007	0.005	(-0.003, 0.017)	0.005	0.002	(0.001, 0.009)	0.007	0.013	(-0.018, 0.032)	0.006	0.014	(-0.021, 0.033)
Age	0.024	0.001	(0.022, 0.026)	0.009	< 0.001	(0.007, 0.011)	0.024	0.005	(0.014, 0.034)	0.009	0.005	(-0.001, 0.019)
Not English-Cdn	0.277	0.043	(0.193, 0.361)	0.330	0.027	(0.277, 0.383)	0.277	0.142	(0.001, 0.555)	0.334	0.145	(0.050, 0.618)
Educ: university	-0.017	0.026	(-0.068, 0.034)	0.044	0.009	(0.026, 0.062)	-0.017	0.108	(-0.229, 0.195)	0.044	0.223	(-0.393, 0.481)
Educ: other training	0.154	0.067	(0.023, 0.285)	0.248	0.020	(0.208, 0.287)	0.154	0.345	(-0.522, 0.830)	0.245	0.312	(-0.367, 0.857)
Income $\geq$ 30000	-0.183	0.028	(-0.238, -0.128)	0.112	0.010	(0.092, 0.132)	-0.184	0.127	(-0.433, 0.065)	0.113	0.153	(-0.187, 0.413)
Income not reported	-0.024	0.049	(-0.120, 0.072)	0.115	0.016	(0.084, 0.146)	-0.024	0.184	(-0.385, 0.337)	0.115	0.164	(-0.206, 0.436)

Abbreviations: SIMEX-CC, Simulation-Extrapolation Conditional on Covariates; SE, Standard Error; CI, Confidence Interval

Table 6–6: **Sensitivity Analysis 3: Estimating Sampling Weights Using a Linear Rather than Multinomial Regression Model** – Summary results from unweighted and weighted naïve regression and unweighted and weighted SIMEX-CC to assess the relationship between phylogenetic cluster size and number of sex partners. See Table 6–3 in Appendix F for definitions of categorical variables.

	Naïve Unweighted			Naïve Weighted			SIMEX-CC Unweighted			SIMEX-CC Weighted		
	$\hat{\beta}$	SE	CI	$\hat{\beta}$	SE	CI	$\hat{\beta}$	SE	CI	$\hat{\beta}$	SE	CI
Outcome: Number of Sex Partners												
Cluster size	0.002	0.005	(-0.008, 0.012)	0.004	0.001	(0.002, 0.006)	0.001	0.008	(-0.015, 0.017)	0.003	0.009	(-0.015, 0.021)
Age	0.016	0.001	(0.014, 0.018)	0.015	< 0.001	(0.013, 0.017)	0.016	0.004	(0.008, 0.024)	0.015	0.003	(0.009, 0.021)
Not English-Cdn	0.179	0.035	(0.110, 0.248)	0.185	0.010	(0.165, 0.205)	0.178	0.106	(-0.030, 0.386)	0.185	0.081	(0.026, 0.345)
Educ: university	0.016	0.022	(-0.027, 0.058)	0.009	0.005	(-0.001, 0.019)	0.016	0.086	(-0.153, 0.185)	0.009	0.087	(-0.162, 0.180)
Educ: other training	0.165	0.067	(0.034, 0.296)	0.032	0.017	(-0.001, 0.065)	0.165	0.354	(-0.529, 0.859)	0.032	0.320	(-0.595, 0.659)
Income $\geq$ 30000	-0.073	0.023	(-0.118, -0.028)	-0.048	0.006	(-0.060, -0.036)	-0.073	0.096	(-0.261, 0.115)	-0.048	0.102	(-0.248, 0.152)
Income not reported	0.034	0.043	(-0.050, 0.118)	0.062	0.011	(0.040, 0.084)	0.034	0.132	(-0.225, 0.293)	0.062	0.132	(-0.197, 0.321)
Outcome: Number of One Night Sex Partners												
Cluster size	0.005	0.005	(-0.005, 0.015)	0.009	0.001	(0.007, 0.011)	0.005	0.013	(-0.020, 0.030)	0.008	0.014	(-0.019, 0.035)
Age	0.022	0.001	(0.020, 0.024)	0.021	< 0.001	(0.020, 0.023)	0.022	0.004	(0.014, 0.030)	0.021	0.004	(0.013, 0.029)
Not English-Cdn	0.228	0.044	(0.142, 0.314)	0.221	0.013	(0.196, 0.246)	0.228	0.139	(-0.044, 0.500)	0.221	0.123	(-0.020, 0.462)
Educ: university	0.017	0.027	(-0.036, 0.070)	0.014	0.007	(0.001, 0.028)	0.017	0.105	(-0.189, 0.223)	0.014	0.099	(-0.180, 0.208)
Educ: other training	0.179	0.079	(0.024, 0.333)	-0.017	0.021	(-0.058, 0.024)	0.178	0.527	(-0.855, 1.211)	-0.017	0.437	(-0.874, 0.840)
Income $\geq$ 30000	-0.135	0.028	(-0.190, -0.080)	-0.112	0.007	(-0.126, -0.098)	-0.136	0.117	(-0.365, 0.093)	-0.112	0.121	(-0.359, 0.125)
Income not reported	0.026	0.052	(-0.076, 0.128)	0.065	0.013	(0.040, 0.090)	0.026	0.198	(-0.362, 0.414)	0.065	0.189	(-0.305, 0.435)

Abbreviations: SIMEX-CC, Simulation-Extrapolation Conditional on Covariates; SE, Standard Error; CI, Confidence Interval

Chapter 7
Discussion

This doctoral research is based on the data from SPOT, a study in Montreal that offers HIV testing to MSM who show no signs of HIV or AIDS, and collects data on sociodemographic and behavioural characteristics. The main focus of my research was to address some of the methodological challenges that arise when studying correlates of HIV phylogenetic cluster size in MSM by combining phylogenetic and epidemiological data. The two major challenges that my research dealt with within the SPOT data were (1) measurement error in the cluster size, and (2) the use of a non-probability sampling scheme. The measurement error in forming the sexual cluster size arises due to not being able to obtain blood samples from all HIV-positive MSM in Quebec. The sampling scheme in SPOT was a convenience approach, relying heavily on personal contacts between members of the MSM community. Because the sampling frame is unknown and the sampling mechanism is non-probabilistic, findings from the study may suffer from a lack of generalizability.

While several competing approaches are available for correcting measurement error, the most suitable method in the SPOT data setting is the SIMEX method of Cook and Stefanski [68] because it does not require validation or repeat measurement data. However, the direct application of SIMEX method is not possible because the standard SIMEX development was limited to mean zero random errors. In the context of under-counted measures in the cluster size of SPOT data, SIMEX would need to be extended to alternative error distributions that has non zero mean. Further, there exists little literature on adjusting for convenience sampling approaches, but a significant literature on survey sampling methods which, with adequate external information, could prove helpful in analyzing the SPOT data. Below, I briefly revisit the objectives and findings of my doctoral research and highlight the

contributions to the field. I then outline limitations of my work and directions for further research before concluding.

Objectives and Findings

In the first manuscript, I extended and validated the SIMEX method to the case where errors can have any known parametric distribution with non-zero mean which better mimics the undercounting cluster size in the SPOT data. The non-zero mean SIMEX (NZM-SIMEX) estimators were theoretically proven to be consistent in continuous outcome, linear model settings. To empirically evaluate the performance of the NZM-SIMEX procedure, a number of simulation studies were carried out by varying: outcome distribution, covariate distribution, specification of error distribution (correctly-, over- and under- specified), and sample size. Simulation results suggested that the NZM-SIMEX performed very well in reducing bias as compared to the naïve method that ignores measurement error. The NZM-SIMEX was found to perform best when the measurement error distribution was correctly specified. It performed better than the naïve method even when the measurement error distribution was mis-specified, although exhibits some bias in those cases. These results were in line with the findings in Cook and Stefanski [68]: there is some evidence that SIMEX is robust to the specification of error distribution. The NZM-SIMEX was applied to the SPOT data to examine correlates of HIV phylogenetic cluster size. No statistically significant association was observed between the cluster size and the demographic and behavioural covariates of interest, indicating that these characteristics have not been shown to help identify and, subsequently, break the link of HIV transmissions within large clusters.

In the second manuscript, I further extended the NZM-SIMEX to the case where measurement error in a target variable depends on other error free covariates. This extension was undertaken so as to make use of the whole of the SPOT data such that both HIV-positive and HIV-negative MSM can be included in the analysis. In the SPOT study, measurement

error in the phylogenetic cluster size depends on a correctly measured covariate: HIV status. For HIV-negative MSM there is no measurement error in their cluster size: the size is known to be 0. Measurement error occurs only in the measurement of the phylogenetic cluster size ofthe HIV-positive MSM. I compared the performance of SIMEX conditional on covariates (SIMEX-CC) to other well known methods such as regression calibration and multiple imputation for measurement error. Simulation results suggested that for correctly specified measurement error distribution, SIMEX-CC performed at least as well as competing approaches. Even in mis-specified cases, SIMEX-CC can sometimes perform better than other methods. Applying SIMEX-CC to the SPOT data, no association was detected between phylogenetic cluster size and the sexual behavioural characteristics. Overall conclusions from the SIMEX-CC and naïve methods were similar. This may be due to the small number of HIV-positive MSM in the sample (who contributes to forming cluster size); measurement error is thus unlikely to strongly affect the results.

Finally, in the third manuscript, I demonstrated methods that can be used to improve the generalizability of results from a study whose sampling scheme is non-probabilistic, where the covariate of interest is prone to systematic *undercounting*. In the SPOT study, participants were recruited often by word of mouth, using a convenience sampling approach aimed at MSM often not reached by traditional HIV testing services. As such, the study cannot be said to have employed any specific probabilistic sampling scheme. Thus, in this manuscript, I again adopted SIMEX-CC for correcting measurement error and considered sampling weights derived from an external data source to adjust for sampling scheme. Specifically, the sampling weights for the SPOT study were obtained from a prediction model constructed by utilizing another HIV study of MSM in Montreal (the ARGUS study) for which the sampling mechanism was known by design. While adjustment for sampling design did provide some notable changes in estimates, the analysis still unable to reveal any important correlation between the phylogenetic cluster size and the number of sexual partners of a MSM.

Contribution

This dissertation research contributes to the realm of biostatistics both theoretically and practically. I developed a measurement error correction technique that can deal with the cases where measurement error distribution (i) can have non-zero mean; and (ii) may depend on another correctly measured covariate in the sample. Further, I demonstrated how an external source of information can be used to improve the generalizability of the results in a study whose sampling scheme is not probabilistic. I validated the methods via a series of simulation studies and compared the performance with other popular methods. The main advantage of the proposed method is that unlike most other methods, it does not require a validation or replicated sample, provided that reasonable knowledge of the measurements error distribution is available to the analyst. The methods that I developed were applied to study the correlates of phylogenetic clusters in the SPOT data. Although the methods that I developed were motivated by the analysis of SPOT data, they can also be used in situations whenever measurement error occurs due to systematic *undercounting* or *overcounting*, or when measures are known to vary by, say, sex or other easy to measure variables. Further, the approach to adjusting for the sampling distribution may also be widely applicable provided external data are available. In some cases, such as studies of vulnerable or hard-to-reach populations, it may be unreaslistic to obtain data from another study that used a probabilistic-based sampling and recruitment scheme. However, this approach could be used in a variety of settings – even where the sampling scheme was well-understood but there were many non-responders – if, for instance, population census data could be used to re-weight the observed sample.

Limitations

This research has several limitations. While detailed limitations were discussed in each manuscript separately, here I briefly outline the overall limitations common to the three manuscripts.

The major drawback of the NZM-SIMEX and SIMEX-CC is that both require "reasonable" knowledge of the measurement error distribution. In the case of mis-specified error distribution, my simulation results suggested that it may be safer to overestimate the variability of the measurement error for NZM-SIMEX. On the other hand, the simulation studies for SIMEX-CC were designed to mimic the SPOT setting. As such, a significant limitation is that the performance of the SIMEX-CC has not been fully explored under a wide range of settings. I considered only the case where one sub-population's measures were subject to error. However, as in laboratory measurements, there may be cases where men and women's measurements have different error distributions. In such situations, the performance of SIMEX-CC under the mis-specified error distribution has not yet been examined.

The overall results from the SPOT study may not be generalizable to the entire population of MSM in Montreal, but only to those MSM who frequent gay social venues. In the first manuscript, I included only HIV-positive MSM, thereby reducing the sample size to a greater extent (from $n = 1803$ to $n = 34$). Therefore, conclusion from the first manuscript is likely to be affected by limited power as well as lack of generalizability. The lack of a larger sample also prevented inclusion of more covariates in the mean model.

I assumed that error in cluster size is due to primarily missing of HIV-positive individuals, but I have ignored errors due to the phylogenetic clustering itself, which may be subject to uncertainty too. However, for the data analysis, I considered a range of plausible error distributions, and conclusions were unchanged. This suggested that taking additional sources of error into account may not distort the overall conclusion of the analysis.

In the theoretical development of NZM-SIMEX and SIMEX-CC, I considered the simple settings: (i) including only one independent variable; (ii) additive error was independent of the value of both mis-measured observed covariate and the unobserved (true) covariate. While the extension to the multiple covariates would be quite straightforward, it is however more challenging to incorporate the situation where measurement error depends on both the

observed and unobserved (true) values of the mis-measured covariate. To do so would require strong, untestable assumptions similar to the "missing not at random assumption" [21].

To improving the generalizability of results, I made use of data from two studies of MSM in Canada. While both studies are large, there is limited power because of the small number of HIV-positive MSM in SPOT.

In the third manuscript, the analysis implicitly assumed that ARGUS had recruit from the same population as SPOT. However, there is evidence to suggest that ARGUS may have recruited from a different subset of the MSM community, as 81% of ARGUS participants reported being Franco- or Anglo-Quebecois with only 19% from other countries, where as in SPOT, 39% were of non-Canadian origin. This suggests that SPOT may have reached a different members of the MSM community such as new immigrants. If this is the case – that the venues from which ARGUS sampled did not adequately cover the entire MSM community – then the weighting scheme that I used will simply have adjusted SPOT to look more like ARGUS participants, but neither study sample will represent, or generalize to, the entire MSM community in Montreal. Being aware of this difference in the country of origin of the two studies' participants and not having any further information to suggest which distribution might better characterize the target population with respect to background, I chose a very rough dichotomization of the ethnic background variable to minimize the impact of this variable on the venue-based sampling weights.

An additional limitation is that the SPOT data cannot be assumed to be independent and identically distributed. A large number (26.6%) of SPOT participants were referred by friends. Therefore, they may have similar demographic and behavioural characteristics. Note that due to the anonymous nature of the SPOT study, it is not possible to build contact network information among study participants, and thus measures of between-person correlation are not available within the study.

An important, and significant, limitation of this work is that cluster size was treated as a static measure. However, its very nature, cluster size would be expected to grow as time progresses. Thus, a cluster could be "large" simply because it is an older, more established cluster and this cluster may contain a mix of newly infected individuals as well as individuals infected many years prior. In my analyses, I used cluster measurement data from the time of analysis, not the time of infection and, as such, may unable to observe any relationships that might exist between epidemiologic covariates and rapidly-expanding clusters. Ideally, we would want to be able to measure, and then correlate, (i) the size of the cluster and (ii) socio-demographic and lifestyle data (from questionnaires) immediately prior to the at the point at which a person becomes infected with HIV, along with the same socio-demographic and lifestyle information from individuals who are not infected as some 'comparable' point in time (e.g. for a MSM who does not acquire HIV, matched for age and calendar year). Having access to those data would ensure temporal ordering of the lifestyle variables prior to HIV infection, and provide a more comparison group of HIV-negative individuals who are more similar at a clinically relevant time-point.

Future Research

While this thesis develops a measurement error (in the covariate) correction technique by extending the SIMEX method to the cases where measurement errors are not necessarily mean zero, users of SIMEX may, however, benefit from further work in the areas described below.

Measurement error in discrete variables invokes a *misclassification* problem. Use of misclassified covariates and responses in regression model will lead to inconsistent estimators of covariate effects [141, 142]. Although various techniques, including methods based on maximum likelihood or quasi-likelihhod, pseudo-likelihoods [143–145], estimating functions [146, 147] and Bayesian methods [148] are available, SIMEX can also be used for this purpose. As discussed previously, unlike many other methods, SIMEX does not require either

internal or external validation samples or replication studies for estimating the parameters of the misclassification matrix [58]. The SIMEX method either assumes the misclassification matrix is known or uses an estimated misclassification matrix. The development of SIMEX method in misclassification of covariates has been limited to binary covariates [88]. Therefore, from the SPOT data perspective, it might be of interest to develop misclassification SIMEX (MC-SIMEX) when the predictor has multiple categories or follows some other discrete distribution, which may alternatively be viewed as a situation in which there are multiple, correlated binary predictors (the category indicators). The resulting methods can then be applied to analyze SPOT data focusing on a typically used categorized version of cluster size, labelling clusters as "unique", "small", and "large".

Measurement error in SPOT, as discussed previously, actually occurs due to missing data (from the untested/unknown-status individuals). Further research can also be conducted to build a model of an MSM community to test what parameters might realistically correspond to the data observed in SPOT by generating a population of all HIV-positive MSM in Montreal. One can begin with a simplest possible model which requires: (i) the population size; (ii) the number and size of the sexual clusters in Montreal; (iii) the probability of being tested; and (iv) the distribution of covariates. With plausible ranges for these parameters, it would be possible to create a population of HIV-positive MSM, generate their covariate information, and know whether or not each man is tested. Based on those who were tested, it would be possible to know how many people in their cluster have also been tested and use this as a measure of the observed (mis-measured) cluster size. Knowing also how many in the cluster have not been tested gives knowledge of the true cluster size. One can then compute the bias in the associations between covariates and cluster size based on using the true and the mis-measured cluster size. The literature [149–152] and the SPOT data can be used to help inform reasonable ranges for each of the parameters.

The SIMEX to accommodate systematic *undercounting* or *overcounting* can also be developed for the correlated data situation in the generalized linear mixed effects models (GLMMs) setting. Further extension of SIMEX to the non-zero mean error distribution can be made in the time-to-event data situation under the Cox proportional hazard model framework.

As noted in the limitations, all analyses were undertaken at a fixed point in time. However, phylogenetic cluster size is a measure that is dynamic in time. To truly harness the information available in the phylogenetic data, dynamic models will need to be employed to understand what factors contribute to (i) risk of infection, (ii) the growth of a cluster, and (iii) the interplay of individual-level social networks, as the data available are almost surely not independent.

Final Comments

There is growing interest in correlating HIV phylogenetic clustering data with epidemiological data. However, such analyses face several methodological challenges, one of which is the systematic *undercounting* of the true sexual network cluster size. This measurement error causes bias and can thus lead to incorrect inferences. Overall, my dissertation research contributes to the biostatistical literature by developing a general measurement error correction technique that can deal with systematic errors that need not have mean zero, where measurement error distribution may depend on another correctly measured covariate. My work has also demonstrated how external data may be used to improve the generalizability of results from a study whose sampling scheme was not probabilistic. The methods were applied to the SPOT data to study to correlates of the phylogenetic cluster size.

It is hoped that the developed methods will ultimately enhance the existing methodology to analyze with phylogenetic or related data in order to make valid inferences.

Bibliography

[1] http://www.amfar.org/about-hiv-and-aids/facts-and-stats/statistics--worldwide/.

[2] William NR, Steven BM. *Environmental and Occupational Medicine, 4th ed.* Lippincott Williams and Wilkins: 530 Walnut Street, Philadelphia, 19106, USA, 2007.

[3] Gallo R. A reflection on HIV/AIDS research after 25 years. *Retrovirology,* 2006; **3**:72.

[4] Mandell GL, Bennett JE, Dolin R. *Mandell, Douglas, and Bennett's Principles and Practice of Infectious Diseases (7th ed.).* Philadelphia, PA: Churchill Livingstone Elsevier: USA, 2010.

[5] Shilts R. *And the Band Played On: Politics, People, and the AIDS Epidemic.* St. Martin's Press: New York, USA, 1987.

[6] http://www.phac-aspc.gc.ca/aids-sida/publication/survreport/estimat08-eng.php.

[7] Yang Q, Ogunnaike-Cooke S, Halverson J, et al. Estimated national HIV incidence rates among key sub-populations in Canada. *Presented at 25th Annual Canadian Conference on HIV/AIDS Research, Winnipeg, Canada*, 2016.

[8] Hutchinson AB, Farnham PG, Dean HD, Ekwueme DU, del Rio C, Kamimoto L, Kellerman SE. The economic burden of HIV in the United States in the era of highly active antiretroviral therapy: evidence of continuing racial and ethnic differences. *Journal of Acquired Immune Deficiency Syndrome* 2006; **43(4)**:451–457.

[9] Brenner BG, Roger M, Stephens D, Moisi D, Hardy I, Weinberg J, Turgel R, Charest H, et al. Transmission clustering drives the onward spread of the HIV epidemic among men who have sex with men in Quebec. *Journal of Infectious Disease* 2011; **204(7)**:1115–1119.

[10] Leigh BAJ, Lycett SJ, Weinert L, Hughes GJ, Fearnhill E, Dunn DT, et al. Transmission network parameters estimated from HIV sequences for a nationwide epidemic. *Journal of Infectious Disease* 2011; **204(9)**:1463–1469.

[11] Lewis F, Hughes GJ, Rambaut A, Pozniak A, Andrew J, Leigh Brown L. Episodic sexual transmission of HIV revealed by molecular phylodynamics. *PLOS Medicine* 2008; **5(3)**:e50.

[12] Hué S, Clewley JP, Cane PA, Pillay D. HIV-1 pol gene variation is sufficient for reconstruction of transmissions in the era of antiretroviral therapy. *AIDS* 2004; **18(5)**:719–728.

[13] Brenner BG, Roger M, Routy JP, Moisi D, Ntemgwa M, Matte C, Baril JG, Thomas R, et al. High rates of forward transmission events after acute/early HIV-1 infection. *Journal of Infectious Diseases* 2007; **195(7)**:951–959.

[14] Bezemer D, Sighem VA, Lukashov VV, van der Hoek L, Back N, Schuurman R, Boucher CA, Claas EC, et al. Transmission networks of HIV-1 among men having sex with men in the Netherlands. *AIDS* 2010; **24(2)**:271–282.

[15] Hughes GJ, Fearnhill E, Dunn D, Lycett SJ, Rambaut A, Andrew J, Brown L. Molecular phylodynamics of the heterosexual HIV epidemic in the United Kingdom. *PLOS Pathogens* 2009; **5(9)**:e1000 590.

[16] Yerly S, Junier T, Gayet-Ageron A, Amari EB, von Wyl V, Gnthard HF, Hirschel B, Zdobnov E, et al. The impact of transmission clusters on primary drug resistance in newly diagnosed HIV-1 infection. *AIDS* 2009; **23(11)**:1415–1423.

[17] Levy I, Mor Z, Anis E, Maayan S, Leshem E, Pollack S, Chowers M, Mor O, Riesenberg K, et al. Men who have sex with men, risk behavior, and HIV infection: integrative analysis of clinical, epidemiological, and laboratory databases. *Clinical Infectious Diseases* 2011; **52(11)**:1363–1370.

[18] Brenner BG, Wainberg MA, Roger M. Phylogenetic inferences on HIV-1 transmission: Implications for the design of prevention and treatment interventions. *AIDS* 2013; **27**:1045–1057.

[19] Brenner BG, Wainberg MA. Future of phylogeny in prevention. *Journal of Acquired Immune Deciency Syndrome* 2013; **2**:s248–54.

[20] Messer K, Natarajan L. Maximum likelihood, multiple imputation and regression calibration for measurement error adjustment. *Statistics in Medicine* 2008; **27**:6332–6350.

[21] Rubin DB. *Multiple Imputation for Nonresponse in Surveys*. John Wiley and Sons, Inc, 1987.

[22] Cole SR, Chu H, Greenland S. Multiple-imputation for measurement-error correction. *International Journal of Epidemiology* 2006; **35**:1074–1081.

[23] Padilla MA, Divers J, Vaughan LK, Allison DB, Tiwari HK. Multiple imputation to correct for measurement error in admixture estimates in genetic structured association testing. *Human Heridity* 2009; **68**:65–72.

[24] Apanasovich TV, Raymond JC, Maity A. SIMEX and standard error estimation in semiparametric measurement error models. *Electron Journal of Statistics* 2009; **3**:318–348.

[25] Goodman LA. Snowball sampling. *Annals of Mathematical Statistics* 1961; **32**:148–170.

[26] Watters JK, Biernacki P. Targeted sampling: options for the study of hidden populations. *Social Problems* 1989; **36**:416–430.

[27] Magnani R, Sabin K, Saidel T, Heckathorn D. Review of sampling hard-to-reach and hidden populations for HIV surveillance. *AIDS* 2005; **19**:s67–72.

[28] Muhib FB, Lillian SL, Stueve A, Miller RL, Ford WL, Johnson WD, Smith PJ, for Youth Study Team CIT. A venue-based method for sampling hard-to-reach populations. *Public Health Rep* 2011; **116**:216–222.

[29] Gustafson P, Gilbert M, Xia M, Michelow W, Robert W, Trussler T, McGuire M, Paquette D, Moore DM, Gustafson R. Impact of statistical adjustment for frequency of venue attendance in a venue-based survey of men who have sex with men. *American Journal of Epidemiology* 2013; **177(10)**:1157–1164.

[30] Guo Y, Li X, Fang X, Lin X, Song Y, Jiang S, Stanton B. A comparison of four sampling methods among men having sex with men in China: implications for HIV/STD surveillance and prevention. *AIDS Care* 2011; **23(11)**:1400–1409.

[31] Sepkowitz KA. AIDS – the first 20 years. *The New England Journal of Medicine* 2001; **344(23)**:1764–1772.

[32] Sharp PM, Hahn BH. Origins of HIV and the AIDS pandemic. *Cold Spring Harbor Perspectives in Medicine* 2011; **1(1)**:a006 841.

[33] Gao F, Bailes E, Robertson DL, Chen Y, Rodenburg CM, Michael SF, Cummins LB, Arthur L, et al. Origin of HIV-1 in the chimpanzee Pan troglodytes troglodytes. *Nature* 1999; **397(6718)**:436–441.

[34] Keele BF, van Heuverswyn F, Li YY, Bailes E, Takehisa J, Santiago ML, Bibollet-Ruche F, Chen Y, et al. Chimpanzee reservoirs of pandemic and nonpandemic HIV-1. *Science* 2006; **313(5786)**:523–526.

[35] Reeves JD, Doms RW. Human Immunodeficiency Virus Type 2. *Journal of General Virology* 2002; **83(6)**:1253–1265.

[36] Tarantola D. Reducing HIV/AIDS risk, impact and vulnerability. *Bulletin of the World Health Organization,* 2000; **78**:2.

[37] *Fact Sheet 2016*. UNAIDS, 2016.

[38] Zhou X, Lin Z, Ma H. Phylogenetic detection of numerous gene duplications shared by animals, fungi and plants. *Genome Biology* 2010; **11**:R38.

[39] Asten LV, Verhaest I, Lamzira S, Aguado IH, Zangerle R, Boufassa F, Rezza G, Broers B, et al. Spread of hepatitis C virus among European injection drug users infected with HIV: a phylogenetic analysis. *Journal of Infectious Disease* 2004; **189(2)**:292–302.

[40] Waters ER, Vierling E. The diversification of plant cytosolic small heat shock proteins preceded the divergence of mosses. *Molecular Biology and Evolution* 1999; **16(1)**:127–139.

[41] Burgin MJ, Casal JJ, Whitelam GC, Sánchez RA. A light regulated pool of phytochrome and rudimentary high irradiance responses under far red light in *pinus elliottii and seudotsuga menziesii. Jounial of Experimental Botany* 1999; **50(335)**:831–836.

[42] Kolukisaoglu HU, Marx S, Wiegmann C, Hanelt S, Schneider-Poetsch HAW. Divergence of the phytochrome gene family predates angiosperm evolution and suggests that selaginella and equisetum arose prior to psilotutm. *Jounial of Molecular Evolution* 1995; **41**:329–337.

[43] Mathews S, Sharrock RA. Phytochrome gene diversity. *Plant, Cell and Environment* 1997; **20(6)**:666–671.

[44] McDowell JM, Huang S, McKinney EC, An YQ, Meagher RB. Structure and evolution of the actin gene family in Arabidopsis thaliana. *Genetics* 1997; **142**:587–602.

[45] Miazaki AS, Gastauer M, Meira-Neto JAA. Environmental severity promotes phylogenetic clustering in campo rupestre vegetation. *Acta Botanica Brasilica* 2015; **29(4)**:561–566.

[46] Leitner T, Escanilla D, Frazen C, Uhlen M, Albert J. Accurate reconstruction of a known HIV-1 transmission history by phylogenetic tree analysis. *Proceedings of the National Academy of Sciences of the United States of America,* 1996; **93(20)**:10 864–10 869.

[47] Scaduto DI, Brown JM, Haaland WC, Zwickl DJ, Hillis DM, Metzker ML. Source identification in two criminal cases using phylogenetic analysis of HIV-1 DNA sequences. *Proceedings of the National Academy of Sciences of the United States of America,* 2010; **107(50)**:21 242–21 247.

[48] Bernard EJ, Azad Y, Vandamme AM, Weait M, Geretti AM. HIV forensics: pitfalls and acceptable standards in the use of phylogenetic analysis as evidence in criminal investigations of HIV transmission. *HIV Medicine* 2007; **8(6)**:382–387.

[49] Chalmet K, Staelens D, Blot S, Dinakis S, Pelgrom J, Plum J, Vogelaers D, Vandekerckhove L, et al. Epidemiological study of phylogenetic transmission clusters in a local HIV-1 epidemic reveals distinct differences between subtype B and non-B infections. *BMC Infectious Diseases,* 2010; **10**:262.

[50] Aldous JL, Pond SK, Poon A, Jain S, Qin H, Kahn JS, Kitahata M, Rodriguez B, et al. Characterizing HIV transmission networks across the United States. *Clinical Infectious Disease* 2012; **55(8)**:1135–1143.

[51] Otis J, McFadyen A, Haig T, Blais M, Cox J, Brenner B, Rousseau R, Émond G, et al. Beyond condoms: risk reduction strategies among gay, bisexual, and other men

who have sex with men receiving rapid HIV testing in Montreal, Canada. *AIDS and Behavior* 2016; **20**:2812–2826.

[52] Emond G, Ghayas F, Otis J, Girard ME, et al. A rapid HIV testing intervention for MSM in a community setting attracts a high proportion of MSM born outside of Canada. *Annual Canadian Conference on HIV/AIDS Research (CAHR). Saskatoon, Canada*, 2010.

[53] Haig T, Thiboutot C, Emond G, Fadel G, Wainberg M, Rousseau R, Otis J, et al. Promoting new HIV testing options to MSM in Montreal: SPOTs communications campaign. *Annual Canadian Conference on HIV/AIDS Research (CAHR). Saskatoon, Canada*, 2010.

[54] Veillette-Bourbeau L, Otis J, Blais M, et al. Opportunities and challenges in implementing a new HIV prevention paradigm: rapid HIV testing for MSM in Montreal at SPOT. *Annual Canadian Conference on HIV/AIDS Research (CAHR). Toronto, Canada*, 2011.

[55] Emond G, Otis J, Blais M, et al. MSM not previously tested for HIV: baseline data from the SPOT project. *Annual Canadian Conference on HIV/AIDS Research (CAHR). Toronto, Canada*, 2011.

[56] Brenner BG, Wainberg MA. Future of phylogeny in prevention. *Journal of Acquired Immune Deficiency Syndrome* 2013; **2**:S248–54.

[57] Brenner BG, Ibanescu RI, Hardy I, Stephens D, Otis J, Moodie E, Grossman Z, et al. Large cluster outbreaks sustain the HIV epidemic among MSM in Quebec. *AIDS* 2017; **31**:707–717.

[58] Carroll R, Ruppert D, Stefanski L. *Measurement Error in Nonlinear Models*. Chapman and Hall, London, UK, 1995.

[59] Frisch R. *Statistical Colzfl cteizce Analysis by earzs of Conzplete Regression Systems*. Oslo: University Economics Institute, Norway, 1934.

[60] Regier MD, Moodie EEM, Platt RW. The effect of error-in-confounders on the estimation of the causal parameter when using marginal structural models and inverse probability-of-treatment weights: a simulation study. *The International Journal of Biostatistics* 2014; **10(1)**:1–15.

[61] Rosner B, Spiegelman D, Willett WC. Correction of logistic regression relative risk estimates and confidence intervals for measurement error: the case of multiple covariates measured with error. *American Journal of Epidemiology* 1990; **134(4)**:734–745.

[62] J A Hausman JA, Abrevaya J, Scott-Morton FM. Misclassification of the dependent variable in a discrete-response setting. *Journal of Econometrics* 1998; **87**:239–269.

[63] Fuller WA. *Measurement Error Models*. John Wiley and Sons, New York, USA, 1987.

[64] Freedman LS, Midthune D, Carroll RJ, Kipnis V. A comparison of regression calibration, moment reconstruction and imputation for adjusting for covariate measurement error in regression. *Statistics in Medicine* 2008; **27**:5195–5216.

[65] Guolo A, Brazzale AR. A simulation-based comparison of techniques to correct for measurement error in matched case-control studies. *Statistics in Medicine* 2008; **27**:3755–3775.

[66] Rosner B, Willett WC, Spiegelman D. Correction of logistic regression relative risk estimates and confidence intervals for systematic within-person measurement error. *Statistics in Medicine* 1989; **8(9)**:1051–1069.

[67] Carrol RJ, Stefanski LA. Approximate quasilikelihood estimation in models with surrogate predictors. *Journal of American Statistical Association* 1990; **85**:652–663.

[68] Cook JR, Stefanski LA. Simulation-extrapolation estimation in parametric measurement error models. *Journal of the American Statistical Association* 1994; **89**:1314–1328.

[69] Carroll RJ, Küchenhoff H, Lombard F, Stefanski LA. Asymptotics for the SIMEX estimator in structural measurement error models. *Journal of the American Statistical Association* 1996; **91**:242–250.

[70] Stefanski LA, Cook J. Simulation-extrapolation: the measurement error jackknife. *Journal of the American Statistical Association* 1995; **90**:1247–1256.

[71] Stefanski LA, Bay JM. Simulation-extrapolation deconvolution of finite population cumulative distribution function estimators. *Biometrika* 1996; **83**:407–417.

[72] Devnarayan V. Simulation-extrapolation method for heteroscedastic measurement error models with replicate measurements. PhD Thesis, North Carolina State University, Raleigh, NC 1996.

[73] Carroll RJ, Stefanski LA. Asymptotic theory for the SIMEX estimator in measurement error models (STMA V39 4250). *Advances in Statistical Decision Theory and Applications* 1997; **1**:151–164.

[74] Eckert RS, Carroll RJ, Wang N. Transformations to additivity in measurement error models. *Biometrics* 1997; **53**:262–272.

[75] Küchenhoff H, Carroll RJ. Segmented regression with errors in predictors: semi-parametric and parametric methods. *Statistics in Medicine* 1997; **16(1-3)**:169–188.

[76] Luo M, Stokes L, Sager T. Estimation of the CDF of a finite population in the presence of acalibration sample. *Environmental and Ecological Statistics* 1998; **5**:277–289.

[77] Lin X, Carroll RJ. SIMEX variance component tests in generalized linear mixed measurement error models. *Biometrics* 1999; **55**:613–619.

[78] Devanarayan V, Stefanski LA. Empirical simulation-extrapolation for measurement error models with replicate measurements. *Statistics and Probability Letters* 2002; **59**:219–225.

[79] Li Y, Lin X. Functional inference in frailty measurement error models for clustered survival data using the SIMEX approach. *Journal of the American Statistical Association* 2003; **98**:191–203.

[80] Wang N, Lin X, Gutierrez RG, Carroll RJ. Bias analysis and SIMEX approach in generalized linear mixed measurement error models. *Journal of the American Statistical Association* 1998; **93**:249–261.

[81] Yi GY, Tan X, Li R. Variable selection and inference procedures for marginal analysis of longitudinal data with missing observations and covariate measurement error. *The Canadian Journal of Statistics* 2015; **43(4)**:498–518.

[82] Hu B, Li L, Wang X, Greene T. Nonparametric multi-state representations of survival and longitudinal data with measurement error. *Statistics in Medicine* 2012; **31(21)**:10.1002/sim.5369.

[83] Mallick R, Fung K, Krewski D. Adjusting for measurement error in the Cox proportional hazards regression model. *Journal of Cancer Epidemiology* 2002; **7(4)**:155–164.

[84] Greene WF, Cai J. Measurement error in covariates in the marginal hazards model for multivariate failure time data. *Biometrics* 2004; **60(4)**:987–996.

[85] Hu C, Lin DY. Semiparametric failure time regression with replicates of mismeasured covariates. *Journal of the American Statistical Association* 2004; **99**:105–118.

[86] He W, Yi GY, Xiong J. Accelerated failure time models with covariates subject to measurement error. *Statistics in Medicine* 2007; **26**:4817–4832.

[87] He W, Xiong J, Yi GY. SIMEX R package for accelerated failure time models with covariate measurement error. *Journal of Statistical Software* 2012; **46**:Code Snippet 1.

[88] Küchenhoff H, Mwalili SM, Lesaffre E. A general method for dealing with misclassification in regression: the misclassification SIMEX. *Biometrics* 2006; **62(1)**:85–96.

[89] Küchenhoff L, Lederera W, Lesaffreb E. Asymptotic variance estimation for the misclassification SIMEX. *Computational Statistics and Data Analysis* 2007; **62(1)**:85–96.

[90] Guolo A. A double SIMEX approach for bivariate random-effects meta-analysis of diagnostic accuracy studies. *Medical Research Methodology* 2017; **17(1)**:6.

[91] Wang Y, Ma Y, Carroll RJ. Variance estimation in the analysis of microarray data. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 2009; **71(2)**:425–445.

[92] Wang Y, Miller DJ, Clarke R. Approaches to working in high-dimensional data spaces: gene expression microarrays. *British Journal of Cancer* 2008; **98(6)**:1023–1028.

[93] Drummond M, Stewart H. *A Review of European applications of artificial intelligence to space*. NASA Technical Memorandum: Ames Research Center, Moffett Field, California, 1993.

[94] Rubin DB. Multiple imputation after 18+ years. *Journal of the American Statistical Association* 1996; **91(434)**:473–489.

[95] Little R. Missing-data adjustments in large surveys. *Journal of Business and Economic Statistics* 1988; **6**:287–296.

[96] Thoresen M, Laake P. The use of replicates in logistic measurement error modelling. *Scandinavian Journal of Statistics* 2003; **30**:625–636.

[97] Dellaportas P, Stephens DA. Bayesian analysis of error-in-variables regression models. *Biometrics* 1995; **51**:1085–1095.

[98] Huang Y, Chen R, Dagne G. Simultaneous Bayesian inference for linear, nonlinear and semiparametric mixed-effects models with skew-normality and measurement errors in covariates. *International Journal of Biostatistics* 2011; **7(1)**:8.

[99] Gelman A, Hill J. *Data Analysis Using Regression and Multilevel/Hierarchical Models.* Cambridge University Press, Newyork, 2006.

[100] Muff S, Riebler A, Held L, Rue H, Saner P. Bayesian analysis of measurement error models using integrated nested Laplace approximations. *Journal of the Royal Statistical Society: Series C (Applied Statistics)* 2015; **64**:231–252.

[101] Rue H, Martino S, Chopin N. Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 2009; **71**:319–392.

[102] Heckathron D. Respondent-driven sampling: a new approach to the study of hidden populations. *Social Problems* 1997; **44**:174–199.

[103] Heckathron D. Respondent driven sampling II: deriving valid population estimates from chain-referral samples of hidden populations. *Social Problems* 2002; **49**:11–34.

[104] Kalton G. Sampling flows of mobile human populations. *Survey Methodology* 1991; **17**:181–194.

[105] Lohr SL. *Sampling: Design and Analysis, Second Edition, 2nd ed.* Brooks/Cole: Boston, USA, 2010.

[106] `http://www.cdc.gov/std/stats/sti-estimates-fact-sheet-feb-2013.pdf`.

[107] `http://www.spottestmontreal.com`.

[108] Gleser LJ. Improvements of naïve approach to estimation in nonlinear errors-in-variables regression models. *Contemporary Mathematics* 1990; **112**:99–114.

[109] Kim J, Gleser LJ. SIMEX approaches to measurement error in ROC studies. *Communications in Statistics - Theory and Methods* 2000; **29(11)**:2473–2491.

[110] Slate EH, Bandyopadhyay D. An investigation of the MC-SIMEX method with application to measurement error in periodontal outcomes. *Statistics in Medicine* 2009; **28**:3523–3538.

[111] Heid IM, Lamina C, Küchenhoff H, Fischer G, Kloop N, Kolz M, Grallert H, Vollmert C, et al. Estimating the single nucleotide polymorphism genotype misclassification from routine double measurements in a large epidemiologic sample. *American Journal of Epidemiology* 2008; **168**:878–889.

[112] Costas L, Infante-Rivard C, Zock JP, Tongeren MV, Boffetta P, Cusson A, Robles C, Casabonne D, et al. Occupational exposure to endocrine disruptors and lymphoma risk in a multi-centric European study. *British Journal of Cancer* 2015; **112**:1251–1256.

[113] Shang Y. Measurement error adjustment using the SIMEX method: an application to student growth percentiles. *Journal of Educational Measurement* 2012; **49**:446–465.

[114] Allodji RS, Thiúbaut ACM, Leuraud K, Rage E, Henry S, Laurier D, Bénichou J. The performance of functional methods for correcting non-Gaussian measurement error within Poisson regression: corrected excess risk of lung cancer mortality in relation to radon exposure among French uranium miners. *Statistics in Medicine* 2012; **31**:4428–4443.

[115] `http://www.stat.gouv.qc.ca/statistiques/population-demographie/structure/104.htm`.

[116] Global HIV and AIDS statistics. `http://www.avert.org/global-hiv-and-aids-statistics` 2016.

[117] `http://www.catie.ca/en/fact-sheets/epidemiology/epidemiology-hiv-canada`.

[118] `http://www.inspq.qc.ca/publications/notice.asp?E=p&NumPublication=1706`.

[119] Brenner BG, Moodie EM. HIV sexual networks: the Montreal experience. *Statistical Communications in Infectious Diseases* 2012; **4(1)**.

[120] Volz EM, Frost SDW. Inferring the source of transmission with phylogenetic data. *PLoS Computational Biology* 2013; **9(12)**:e1003 397.

[121] Pagani F, Panteghini M. Biological variation in serum activities of three hepatic enzymes. *Clinical Chemistry* 2001; **47(2)**:355–356.

[122] Carrol RJ, Stefanski LA. Approximate quasilikelihood estimation in models with surrogate predictors. *Journal of the American Statistical Association* 1990; **85**:652–663.

[123] Buonaccorsi J. *Measurement Error: Models, Methods and Applications*. Chapman and Hall, CRC Press: Boca Raton, FL, 2010.

[124] Carroll R, Ruppert D, Stefanski L, Crainiceanu C. *Measurement Error in Nonlinear Models: A Modern Perspective*. Chapman and Hall, CRC Press: Boca Raton, FL, 2006.

[125] Blackwell M, Honaker J, King J. A unified approach to measurement error and missing data: overview and applications. *Sociological Methods and Research* 2015; **1**:1–39.

[126] Efron B, Tibshirani R. *An Introduction to the Bootstrap*. Chapman and Hall, CRC Press: Boca Raton, FL, 1993.

[127] Trezo EP. Bayesian hierarchical model for the study of clustered data with cluster level sources of measurement. PhD Thesis, McGill University, QC, Canada 2015.

[128] Hong H, Rudolph KE, Stuart EA. Bayesian approach for addressing differential covariate measurement error in propensity score methods. *Psychometrika* 2016; **4(1)**:1–19.

[129] Granich R, Crowley S, Vitoria M, Lo YR, Souteyrand Y, Dye C, Gilks C, Guerma T, et al. Highly active antiretroviral treatment for the prevention of HIV transmission. *Journal of the International AIDS Society* 2010; **13**:1.

[130] Remis RS, Alary M, Liu J, Kaul R, Palmer RWH. HIV transmission among men who have sex with men due to condom failure. *Journal of the International AIDS Society* 2014; **9(9)**:e107 540.

[131] Remis RS, Palmer RW. Testing bias in calculating HIV incidence from the serologic testing algorithm for recent HIV seroconversion. *AIDS* 2009; **23(4)**:493–503.

[132] Cain R, Collins E, Bereket T, George C, Jackson R, Li A, Prentice T, Travers R. Challenges to the involvement of people living with HIV in community-based HIV/AIDS organizations in Ontario, Canada. *AIDS Care* 2014; **26**:263–266.

[133] Cohen MS, Chen YQ, , McCauley M, Gamble T, Hosseinipour MC, Kumarasamy N, Hakim JG, Kumwenda J, et al. Prevention of HIV-1 infection with early antiretroviral therapy. *New England Journal of Medicine* 2011; **365**:493–505.

[134] Powers KA, Ghani AC, Miller WC, Hoffman IF, Pettifor AE, Kamanga G, Martinson FEA, Cohen MS. The role of acute and early HIV infection in the spread of HIV-1 in Lilongwe, Malawi: Implications for Test and Treat and other transmission prevention strategies. *The Lancet* 2011; **378(9787)**:256–268.

[135] Wainberg MA, Brenner BG. The impact of HIV genetic polymorphisms and subtype differences on the occurrence of resistance to antiretroviral drugs. *Molecular Biology International* 2012; :Article ID 256 982.

[136] Brenner BG, Wainberg MA. Future of phylogeny in prevention. *Journal of Acquired Immune Deficiency Syndrome* 2013; **2**:S248–54.

[137] Parveen N, Moodie EEM, Brenner BG. The non-zero mean SIMEX: improving estimation in the face of measurement error. *Observational Studies* 2015; **1**:91–123.

[138] Parveen N, Moodie EEM, Brenner BG. Correcting covariate-dependent measurement error with non-zero mean. *Statistics in Medicine* 2017; **NA**:NA.

[139] Lambert G, Cox J, Hottes TS, Tremblay C, Frigault LR, Alary M, Otis J, Remis RS, et al. Correlates of unprotected anal sex at last sexual episode: analysis from a surveillance study of men who have sex with men in Montreal. *AIDS and Behavior* 2011; **15**:584–595.

[140] Cleveland WS. Robust locally weighted regression and smoothing scatterplots. *Journal of the American Statistical Association* 1979; **74(368)**:829–836.

[141] Yi G, Cook R. Errors in the measurement of covariates. *Encyclopedia of Biostatistics* 1998; **3**:1741–1748.

[142] Carroll R. Measurement error in epidemiological studies. *Encyclopedia of Biostatistics* 1998; **3**:2491–2519.

[143] Carroll R, Gail M, Lubin J. Measurement error in epidemiological studies. *Journal of the American Statistical Association* 1993; **88**:185–199.

[144] Lawless J, Kalbfleisch JD, Wild C. Semiparametric methods for response-selective and missing data problems in regression. *Journal of the Royal Statistical Society: Series B* 1999; **61**:413–438.

[145] Hanfelt J, Liang KY. Approximate likelihood for generalized linear errors-in-variables models. *Journal of the Royal Statistical Society: Series B* 1999; **59**:627–637.

[146] Nakamura T. Corrected score functions for errors-in-variables models: methodology and application to generalized linear models. *Biometrika* 1990; **77**:127–137.

[147] Pepe M, Fleming T. A nonparametric method for dealing with mismeasured covariate data. *Journal of the American Statistical Association* 1991; **86**:108–113.

[148] Mwalili S, Lesaffre E, Declerck D. A Bayesian ordinal logistic regression model to correct for interobserver measurement error in a geographical oral health study. *Journal of Royal Statistical Society, Series C* 2005; **54**:77–93.

[149] Parsons JT, Severino J, Nanin J, Punzalan JC, Von Sternberg K, Missildine W, Frost D. Positive, negative, unknown: assumptions of HIV status among HIV-positive men who have sex with men. *AIDS education and prevention : official publication of the International Society for AIDS Education* 2006; **18(2)**:139–49.

[150] Jin FY, Prestage G, Law MG, Kippax S, Van de Ven S, Rawsthorne P, Kaldor JM, Grulich AE. Predictors of recent HIV testing in homosexual men in Australia. *HIV Medicine* 2002; **3(4)**:271–276.

[151] MacKellar DA, Valleroy LA, Secura GM, Bartholow BN, McFarland W, Shehan D, Ford W, LaLota M, et al. Repeat HIV testing, risk behaviors, and HIV seroconversion

among young men who have sex with men: a call to monitor and improve the practice of prevention. *Journal of Acquired Immune Deficiency Syndromes* 2002; **29(1)**:76–85.

[152] Shira M, Armistead LP, Kalichman S. Predictors of HIV antibody testing among gay, lesbian, and bisexual youth. *Journal of Adolescent Health* 2000; **26(4)**:252–257.