METHODOLOGICAL ISSUES IN DESIGN AND ANALYSIS OF STUDIES WITH CORRELATED DATA IN HEALTH RESEARCH

By

JINHUI MA, B.Sc., M.Sc.

A Thesis Submitted to the School of Graduate Studies in Partial Fulfilment of the Requirements for the Degree of Doctor of Philosophy

McMaster University DOCTOR OF PHILOSOPHY (2012) Hamilton, Ontario (Health Research Methodology ─ Biostatistics Specialization)

TITLE: Methodological Issues in Design and Analysis of Studies with Correlated Data in Health Research

AUTHOR: Jinhui Ma, B.Sc. (University of British Columbia), M.Sc. (McMaster University)

SUPERVISORS: Dr. Lehana Thabane and Dr. Parminder Raina

NUMBER OF PAGES: xii, 126

Abstract

Background and Objectives:

Correlated data with complex association structures arise from longitudinal studies and cluster randomized trials. In the former case, repeated measurements from the same subject are correlated. In the later case, measurements from the subjects within the same cluster possess identical characteristics. Both the longitudinal and the cluster randomized design are widely adopted in health research. However, some methodological challenges in the design and analysis of such studies or trials have not been overcome. In this thesis, we address three of the challenges: 1) *Power analysis for population based longitudinal study investigating gene-environment interaction effects on chronic disease:* For longitudinal studies with interest in investigating the gene-environment interaction in disease susceptibility and progression, rigorous statistical power estimation is crucial to ensure that such studies are scientifically useful and cost-effective since human genome epidemiology is expensive. However conventional sample size calculations for longitudinal study can seriously overestimate the statistical power due to overlooking the measurement error, unmeasured etiological determinants, and competing events that can impede the occurrence of the event of interest. 2) *Comparing the performance of different multiple imputation strategies for missing binary outcomes in cluster randomized trials*: Though researchers have proposed various strategies to handle missing binary outcome in cluster randomized trials (CRTs), comprehensive guidelines on the selection of the most appropriate or optimal strategy are not available in the literature. 3) *Comparison of population-averaged and cluster-specific models for the analysis of cluster randomized* *trials with missing binary outcome*: Both population-averaged and cluster-specific models are commonly used for analyzing binary outcomes in CRTs. However, little attention has been paid to their accuracy and efficiency when analyzing data with missing outcomes. The objective of this thesis is to provide researchers recommendations and guidance for future research in handling the above issues.

Methods:

Project 1: Motivated by the design of Canadian Longitudinal Study on Aging (CLSA), we designed a simulation study based on an irreversible illness-death model to investigate the power profile and minimum detectable hazard ratio (MDHR) for an environmental risk factor, a genotype risk factor, and their interaction on the transition intensity from healthy aging to different aging related chronic diseases. In this simulation study, we took into account the analytic complexity which could lead to an overestimation of the statistical power; therefore a realistic power profile for the CLSA was provided.

Project 2: Under the assumption of covariate dependent missingness, we investigated the performance of six strategies to handle missing binary outcomes in CRTs. These strategies are complete case analysis which omits those for whom data are missing, two standard multiple imputation (MI) strategies which ignore the clustering effect – logistic regression and Markov chain Monte Carlo (MCMC) method, two within-cluster MI strategies which impute missing values based on the observed data within the same cluster as the missing ones – logistic regression and MCMC method, and MI using

logistic regression with cluster as a fixed effect. The performance of these strategies was evaluated by bias, empirical standard error (ESE), root mean squared error (RMSE), and coverage.

Project 3: We conducted a simulation study to compare the accuracy and efficiency of population-averaged (i.e. generalized estimating equations (GEE)) and cluster-specific (i.e. random-effects logistic regression (RELR)) models for analyzing data from cluster randomized trials (CRTs) with missing binary responses. The clustering binary outcomes from CRTS were generated from a Beta-binomial distribution. Under the assumption of covariate dependent missingness, missing outcomes were handled by complete case, standard MI or within-cluster MI strategies. Data were analyzed using GEE method and RELR. Performance of the two methods was assessed by standardized bias (SB), ESE, RMSE, and coverage.

Results and Conclusion:

Project 1: Given statistical power of 80%, significance level of 0.05 for environmental risk exposures and 0.0001 for genotype risk exposures and gene-environment interactions, the design of CLSA, which involves 30,000 participants measured every three years for at least twenty years, enables moderate ($1.5<HR\leq2.0$) or large ($2.0<HR\leq3.0$) hazard ratio (HR) to be detected for environmental risk exposures. For genotype risk exposures, the CLSA is capable of detecting moderate HR only when the incidence of disease is high, or the prevalence of genotype risk factor is high (≥0.1). For gene-environment interactions,

even large HR can not be detected when the prevalence of genotype and environmental risk factors is low (<0.1). Misclassification on risk factors substantially reduces the statistical power. The HRs for the design involving data collection every three years are slightly larger than those obtained assuming exact event time is observed. Improvement on the study design and implementation and synthesis of information with other human genome studies are recommended to increase the capacity for investigating the effect of determinants on chronic diseases.

Project 2: Under the assumption of covariate dependent missingness and applying the generalized estimating equations approach for fitting the logistic regression, we showed that complete case analysis yields valid inferences when the percentage of missing outcomes is not large (<20%) for all the design of CRTs considered in this paper. Standard MI strategies can be adopted when the design effect is small (variance inflation factor [VIF] ≤3); however, they tend to underestimate the standard error of treatment effect when the design effect is large. Within-cluster MI strategy using logistic regression is valid for imputation of missing data from CRTs, especially when the cluster size is large (>50) and the design effect is large (VIF>3). In contrast, within-cluster MI strategy using MCMC method may yield biased estimates of treatment effect for CRTs with a small cluster size (≤50). MI using logistic regression with cluster as a fixed effect may substantially overestimate the standard error of the estimated treatment effect when the intracluster correlation coefficient is small. It may also lead to biased estimated treatment effect.

Project 3: GEE performs well on all four measures under the following scenarios: complete case analysis for CRTs with a small amount of missing data; standard MI for CRTs with VIF<3; within-cluster MI for CRTs with VIF≥3 and cluster size>50. In contrast, RELR does not perform well when either standard or within-cluster MI strategy is applied prior to the analysis.

Preface

This thesis is a "sandwich" thesis, which combined three individual projects prepared for publication in peer-reviewed journals. The following are contributions of J. Ma in all of the papers included in the dissertation: developing the research ideas and research questions; conducting all statistical analysis; writing all of the manuscripts; submitting the manuscripts; and responding to reviewers' comments. The work of this thesis was conducted between Winter 2008 and Summer 2012.

Acknowledgements

I may not know where life’s road will lead me, but walking with You, God, through this journey has given me strength.

I would like to express my deepest gratitude to my supervisors, Dr. Lehana Thabane and Dr. Paminder Raina, for their wise guidance, caring, patience, encouragement and providing me with an excellent atmosphere for doing research. They have been, and I hope they will continue to be my life-long mentors and friends. I would like to thank Dr. Joseph Beyene for his invaluable contributions to all aspects of the research presented in this thesis.

I would also like to thank professors John Petkau, Narayanaswamy Balakrishnan, Peter Macdonald. Their enthusiasm for statistics and its applications rubbed off on me and inspired me to become a statistician.

I am grateful to my God mother and father, Gloria Lam and Y.K. Lam, for their constant encouragement and support, which helped me through the most difficult time in my life.

Last but not least, I would like to thank my husband Raymond, my daughter Joy, my parents-in-law Guoxiu Li and Defu Tang, my parents Xiufang Yu and Chunshan Ma for their selfless dedication and love. I would like to thank my friends Peggy Lin, Weiwei Xiong, Lin Fang, and Patricia Carson, who have never refused to lend me a helping hand.

Thesis topics come and go, but the musical styling of Journey is forever.

Table of Contents

List of Tables

Chapter 2

Chapter 3

Chapter 4

List of Figures

Chapter 2

Chapter 4

Chapter 1

Introduction

Correlated data with complex association structures arise from longitudinal studies and cluster randomized trials (CRT). In the former case, repeated measurements from the same subject are correlated. In the later case, measurements from the subjects within the same cluster possess identical characteristics. Both the longitudinal and the cluster randomized designs are widely adopted in health research. The correlations cause a number of methodological issues in the design and analysis of such studies or trials.

In this thesis, three issues which have not been fully addressed in the literature are highlighted. First, for longitudinal studies with interest in investigating the gene-environment interaction in disease susceptibility and progression, rigorous statistical power estimation is crucial to ensure that such studies are scientifically useful and cost-effective since human genome epidemiology is expensive. However conventional sample size calculations for longitudinal study can seriously overestimate the statistical power due to overlooking the measurement error, unmeasured etiological determinants, and competing events that can impede the occurrence of the event of interest. Second, though researchers have proposed various strategies to handle missing binary outcome in cluster randomized trials (CRTs), comprehensive guidelines on the selection of the most appropriate or optimal strategy are not available in the literature. Third, both population-averaged and cluster-specific models are commonly used for analyzing binary outcomes in CRTs. However, little attention has been paid to their accuracy and efficiency when analyzing data with missing outcomes. The final objective of this thesis is to provide researchers recommendations and guidance for future research in handling the above issues.

Issue 1: Power analysis for population based longitudinal study investigating gene-environment interaction effects on chronic diseases

The longitudinal design provides possibilities for researchers to exploit any relationship between an antecedent cause and subsequent effects since participants are followed up into the future and the progress of their health conditions and risk exposures can be measured at the pre-specified time points. The advantages of the prospective longitudinal study over other study designs are well documented [1]. First, fewer subjects are required in a longitudinal study since the repeated measurements from the same subject are rarely perfectly correlated and consequently provide more independent information than a single measurement obtained from a single subject. Second, each subject can serve as his/her own control in longitudinal study which results in more efficient estimators of exposure-related effects. Third, through longitudinal design, researchers can separate the changes over time within subjects (i.e. aging effects) from differences between subjects at baseline (i.e. cohort effects). Fourth, individual changes or trends observed from longitudinal data allow researchers to understand the heterogeneity in a population and the determinants of changes at the individual level. However, longitudinal studies are time-consuming, expensive to conduct, and usually subject to attrition or loss to follow-up.

It is crucial that researchers know the minimum required sample size to provide a reliable answer to the primary research question(s) addressed when planning longitudinal studies. As pointed out by Muller *et al* [2], insufficient sample size can lead to inadequate sensitivity, whereas an excessive sample size can be a waste of time and money. There are several common factors that influence the determination of required sample size for any study design [3]. They are: 1) the study objectives (to provide reliable sample size calculation, an appropriate statistical test for the hypotheses of interest, which should be established to reflect the study objectives, is necessarily derived under the study design); 2) the type of endpoint/outcome (continuous, binary, categorical, or survival); 3) variation of the study population; 4) type I error, which is the probability of rejecting the null hypothesis when it is true, and type II error, which is the probability of not rejecting the null hypothesis when it is false; 5) the minimum clinically important effect size; and 6) measurement errors on the outcome and risk exposure. For longitudinal studies, some other factors, such as the sampling strategies, the length of follow-up, the frequency and timing of repeated measurements on participants, the presence and nature of the correlation between repeated measurements from the same subject, and the attrition due to mortality and loss to follow-up may also play important roles on sample size determination.

Current literature on sample size and statistical power estimation for longitudinal study

Researchers developed formulae and software to calculate the minimum required sample size or statistical power for longitudinal studies with either continuous or binary outcomes [2, 4, 5]. For studies with ordinal or categorical response, researchers could use methods or formulae for continuous responses, but adjust the detectable effect size by an efficiency loss [6, 7, 8], or use the method proposed by Kim *et al* [9] for repeated ordinal outcomes. For studies with survival outcome (i.e. time-to-event), most available software, tables and formulae for sample size calculation focused on a single event and assumed the survival time follows an exponential distribution, i.e. the hazard is constant over time [10]. However, the hazard may not necessarily be constant over the study period, especially for research on aging when the follow-up time is long. Heo *et al* showed that the power and sample size can be poorly estimated when the survival time is assumed to be exponential but, in fact, follows Weibull distribution with shape parameter $\rho > 1$ [11]. For multiple-event with completing risk, Ellen Maki [12] proposed methods of calculating the required sample size for clinical trials when two treatments are to be compared with respect to two or more competing risks. In her study, a flexible parametric Weibull model, non-uniform study entry, and an allocation ratio other than unity were considered. For longitudinal studies when the response variable is the stage of disease, and with a focus on the transition intensities, Hwang *et al* [13] discussed the sample size and statistical power of statistical tests on the ratio of transition intensities under the assumption of constant intensity ratio between the two groups. This study was based on the progressive model, which means the subjects may only make an instantaneous transition to the next severe stage, and all subjects start at the same stage. van den Hout and Matthews [14] investigated study design choices such as sample size, length of follow-up, and time intervals between measurements in a simulation study. Their simulation study used a reversible illness-death Model, a three-state Markov model whose transition intensities are allowed to depend on time since entry into the study.

Biological and technological advances over the past decade, such as the sequencing of the human genome, have increased researchers' ability to study aging in all its complexity. The importance of studying gene-environment interactions (G×E) in the context of aging related chronic diseases was emphasized since they typically occur as a result of the interaction between an individual's genetic make-up and detrimental environments [15]. Sample sizes to detect genetic main effects or G×E with sufficient statistical power are expected to be very large (up to several hundred thousand) [16]. Therefore, many genetic association studies were susceptible to lack of sufficient statistical power [17]. Very few longitudinal studies of aging to date have collected biomarker, genetic or epigenetic data to elucidate the process of aging, and to study how biological processes interact with physical and psychosocial environment to produce deleterious health outcomes. Unlike early association studies in which individuals were not tracked over time and all measurements on each participant were made at one point in time, the longitudinal design enables researchers to separate the changes over time within subjects from differences between subjects at baseline, and allows researchers to create the most comprehensive and insightful framework for understanding the mechanisms by which genome function can be altered during aging [18, 19].

Although massive reductions in genotyping costs, prospect cohort study remains limited by the cost of proper phenotyping [20]. Therefore, rigorous sample size and statistical power estimation are crucial to ensure that such kind of studies is scientifically useful and cost-effective. However, conventional sample size calculations for longitudinal study can seriously overestimate the statistical power due to overlooking the measurement error, unmeasured etiological determinants, and competing events that can impede the occurrence of the event of interest.

Contribution of our work to current literature

Motivated by the Canadian Longitudinal Study on Aging (CLSA) [21], we designed a simulation study based on the irreversible illness-death model. The objective of the study was to investigate the power profile and minimum detectable hazard ratio (MDHR) for the environmental risk factor, the genotype risk factor, and their interaction on the transition from the healthy to diseased state, while the transition from healthy to dead state was considered as a competing risk. By taking into account the measurement error on risk exposure and unmeasured etiological determinants, we tried to provide an accurate and realistic power profile for the CLSA. We also assessed the impact of sampling strategy, frequency and timing of repeated measurements, and attrition due to mortality and loss to follow-up on the power profile. Findings from the present study provide empirical guidance for designing newly initiated population genomics cohort studies.

Issue 2: Comparison of strategies to handle missing binary outcomes in CRTs

CRT is the “gold standard” for evaluating the effectiveness of medical interventions since it ensures subjects in treatment and control groups are similar in both measured and unmeasured attributes as long as the number of clusters and the subjects is large [22]. Moreover, allocating subjects as clusters may minimize potential contamination between subjects in different treatment arms. However, CRTs can be susceptible to some methodological problems in the design and analysis since subjects from the same cluster cannot be assumed to be independent. It is well recognized that cluster randomized trials may have substantially reduced statistical efficiency relative to trials that randomize the same number of individuals. The reduction in efficiency is quantified by the variance inflation factor (VIF) (also known as the design effect) defined as $1+(m-1)\ \rho$, where m is the average cluster size and ρ is the intracluster correlation coefficient, interpreted as the proportional of overall variation in response that can be accounted for by the between-cluster variation. One of the consequences of cluster randomization is that traditional or standard statistical analysis methods, which ignore the clustering effect, may increase the chance of making a type I error, a false positive difference between the interventions.

In addition, missing data may occur in some CRTs due to lengthy follow-up or lack of direct contact with individual patients [23, 24]. A further complication is that all individuals in a cluster may be missing, which is likely to occur in CRTs with a small cluster size. Missing data may weaken the power of a trial, and cause bias depending on why the data are missing.

Current literature on strategies to handle missing outcomes in cluster randomized trials

Multiple imputation (MI) has been widely applied to missing data problems. The implementation of this procedure is provided in many commercially available software packages; however, most of them are developed based on the independent data assumption, which may not be valid to handle the clustering data from CRTs. Taljaard *et al* [25] evaluated imputation strategies for missing continuous outcomes in CRTs via simulation, assuming the data were missing completely at random (MCAR). They concluded that if the intracluster correlation coefficient (ICC) is small (<0.005), ignoring the clusters may yield acceptable Type I error; however, if the ICC is large, ignoring the clustering will lead to severe inflation of Type I error. Andridge [26] investigated the impact of fixed-effects modeling of clusters in MI for CRTs with continuous outcomes assuming outcomes are missing completely at random (MCAR) or missing at random (MAR). She showed that incorporation of clustering using fixed effects for clusters can lead to severe overestimation of variance of group means, and the overestimation is more severe when cluster sizes and ICCs are small. We [27] compared several strategies for handling missing binary outcomes in CRTs under the assumption of MCAR and covariate dependent missingness and found that within-cluster and across-cluster MI strategies — which take into account intracluster correlation — provide more conservative treatment effect estimates compared to MI strategies which ignore the clustering effect.

Though researchers have proposed various strategies, comprehensive guidelines on the selection of the most appropriate or optimal strategy for handling missing binary outcomes in CRTs are not available in the literature. The generalizability of the conclusions from our previous study [27] to other design settings may be limited since the simulation study was based on a real dataset which has a relatively large cluster size and intracluster correlation coefficient (ICC). Moreover, we compared different imputation strategies through the odds ratios (OR) and corresponding 95% confidence intervals (CIs) for the estimated treatment effect, and the kappa statistics for agreement between imputed datasets and the real dataset. Other evaluation criteria, such as bias, root mean squared error, and coverage probability etc., are considered more informative to assess the accuracy and efficiency of different imputation strategies.

Contribution of our work to current literature

We evaluated the performance of various strategies for missing binary outcomes in CRTs under different design settings. Under the assumption of covariate dependent missingness (CDM), we focused on the following strategies: complete case analysis, two standard MI strategies – logistic regression and Markov chain Monte Carlo (MCMC) method, two within-cluster MI strategies – logistic regression and MCMC method, and MI strategy using logistic regression with the cluster as a fixed effect. Using the generalized estimating equations (GEE) approach for fitting the population-averaged model for clustered binary data, we compared the performance of these strategies using bias, root mean squared error (RMSE), coverage probability of nominal 95% CI, and empirical standard error of the estimated treatment effect. Findings from this study provide researchers with quantitative evidence to guide the selection of an appropriate strategy to deal with missing binary outcomes based on the design settings of CRTs.

Issue 3: Comparison of population-averaged and cluster-specific models for the analysis of cluster randomized trials with missing binary outcomes

For CRTs with binary outcomes, population-averaged (PA) models (also called marginal models) and cluster-specific (CS) models (also called conditional models), which account for the clustering effect, were proposed in the literature to analyze binary data in CRTs. A representative of PA models is the generalized estimating equations (GEE) method, which estimates the PA intervention effect. A representative of CS models is the random-effects logistic regression (RELR), which estimates the CS intervention effect. Both GEE method and RELR are advocated to analyze data in CRTs since they allow for the possible imbalance of both cluster-level and individual-level characteristics to be incorporated into the analysis.

Current literature on the performance of GEE method and RELR in the analysis of binary outcomes in CRTs

Some attention has been paid in the literature to the performance of GEE approach and RELR in the analysis of binary outcomes in CRTs. Austin [28] compared their statistical powers through a simulation study in which the minimum number of clusters examined was 26 (13 clusters per trial arm). The results showed that the differences between the two methods were negligible in most settings. Bellamy *et al*. also conducted a series of simulation studies comparing their statistical power [29]. They examined settings in which the total number of clusters was 10, 20, 30 or 50, where the mean number of subjects per cluster was either 10 or 100, the ICC was 0.1, the response proportion in the control arm was 0.23 and the response proportions in the intervention arm were 0.09, 0.13, 0.18, 0.23 or 0.28. The study showed that the difference between the two models diminished as the number of clusters increased. In particular, the difference would negligible if the total number of clusters was at least 30. However, if the total number of clusters was 10 or 20, RELR had moderately lower power than GEE method. Ukoumunne *et al*. [30] compared the accuracy of estimated treatment effect and confidence interval coverage of several methods for analyzing binary outcomes in CRTs through a simulation study. They showed that GEE method has acceptable properties as long as its downward bias of the standard error is corrected when the number of clusters is small. The RELR was not assessed in their simulation study.

When missing data occur, common strategies to handle the missing data are to ignore them (complete case analysis), or to replace them with single or multiple plausible values (multiple imputation) and then conduct the statistical analysis. The impact of missing data on estimating treatment effect and its confidence interval depends on the mechanism which causes the data to be missing, the strategy to handle missing data, and the statistical model used for analysis. However, the accuracy and efficiency of the GEE method and the RELR are still unknown when multiple imputation (MI) is applied prior to the analysis.

Contribution of our work to current literature

We compared the accuracy and efficiency of PA and CS models, in particular, the GEE method and the RELR respectively, for analyzing data from CRTs with missing binary outcomes. In this simulation study, different design settings of CRTs and percentage of missing data were considered to mimic the scenarios commonly encountered in practice. The performance of the GEE method and RELR was compared in terms of standardized bias (SB), empirical standard error (ESE), root mean squared error (RMSE), and coverage. Findings from this study provide researchers recommendations and guidance on selecting appropriate imputation and analysis method to avoid poor inference. It will improve the design and analysis of CRTs with missing binary outcomes, and close the gap between statistical knowledge and its application in empirical settings.

Outline of the thesis

This thesis is a sandwich of three papers mapped to each of the issues described above. The three papers are separated into different chapters beginning with Chapter 2.

In Chapter 2, we investigated the statistical power profile for the environmental risk factor, the genotype risk factor, and their interactions on the transition from the healthy to diseased state based on the design of the CLSA, and provided guidance on designing similar population based longitudinal studies.

In Chapter 3, we compared the performance of six strategies for handling missing binary outcomes in CRTs. Quantitative evidence was provided to guide the selection of an appropriate strategy to deal with missing binary outcomes based on the design settings of CRTs.

In Chapter 4, we assessed the performance of GEE method and RELR when the missing binary outcome was handled by different missing data strategies. Findings from this study provided health researchers guidance on selecting appropriate imputation and analysis method to avoid poor inference.

Lastly, in Chapter 5, the key findings, limitations and implications of the thesis were summarized.

References

1. Donald RH, Robert DG, *Longitudinal data analysis*, John Wiley & Sons, Inc. New Jersey, 2006.

2. Muller KE, LaVange LM, Ramey SL, and Ramey CT, Power calculations for general linear multivariate models including repeated measures applications. *Journal of the American Statistical Association* 1992; 87: 1209-1226.

3. Julious SA, *Sample size for clinical trials*, Chapman and Hall, Boca Raton, 2009.

4. Diggle P, Heagerty P, Liang K, Zeger S, *Analysis of longitudinal data,* 2nd edition, Oxford University Press, New York, 2002.

5. Snijders TAB, Bosker RJ, Standard errors and sample sizes for two-level research. *Journal of Educational Statistics* 1993; 18:237-259.

6. Snijders TAB, Bosker RJ, Standard errors and sample sizes for two-level research. *Journal of Educational Statistics* 1993; 18:237-259.

7. Armstrong BG, Sloan M, Ordinal regression models for epidemiologic data. *American Journal of Epidemiology* 1989; 129(1): 191-204.

8. Stromberg U, Collapsing ordered outcome categories: A note of concern. *Am J Epidemiol* 1996; 144(4): 421-424.

9. Kim HY, Williamson JM, Lyles CM, Sample-size calculations for studies with correlated ordinal outcomes. *Stat Med* 2005; 24(19): 2977-2987.

10. Schoenfeld DA, Richter JR, Nomograms for calculating the number of patients needed for a clinical trial with survival as an endpoint. *Biometrics* 1982; 38(1): 163-170.

11. Heo M, Faith MS, Allison DB, Power and sample size for survival analysis under the weibull distribution when the whole lifespan is of interest. *Mech Ageing Dev* 1998; 102(1): 45-53.

12. Maki E, Power and sample size considerations in clinical trials with competing risk endpoints. *Pharm Stat* 2006; 5(3): 159-171.

13. Hwang WT, Brookmeyer R, Design of panel studies for disease progression with multiple stages. *Lifetime Data Anal* 2003; 9(3): 261-274.

14. van den Hout A, Matthews FE, A piecewise-constant markov model and the effects of study design on the estimation of life expectancies in health and ill health. *Stat Methods Med Res* 2009; 18(2): 145-162.

15. Grigorenko EL. The Inherent Complexities of Gene–Environment Interactions. *Journals of Gerontology: SERIES B* 2005; 60B (Special Issue I):53–64.

16. Collins FS. The case of a US prospective cohort study of genes and environment. *Nature* 2004, 249, 475–477.

17. Khoury MJ, Little J, Gwinn M, Ioannidis JP. On the synthesis and interpretation of consistent but weak gene-disease associations in the era of genome-wide association studies. *Int J Epidemiol* 2007;36:439–45.

18. Feinberg AP. Phenotypic plasticity and the epigenetics of human disease. *Nature* 2007 May 24; 447(7143):433-40.

19. Calvanese V, Lara E, Kahn A, et al. The role of epigenetics in aging and age-related diseases. *Ageing Res Rev* 2009 Oct; 8(4):268-76.

20. Thompson EA. Human genetics: overview. In: Palmer L, ed. Biostatistical genetics and genetic epidemiology. Chichester, United Kingdom: John Wiley and Sons Ltd, 2002:386–90.

21. Raina PS, Wolfson C, Kirkland SA, Griffith LE, Oremus M, et al. The canadian longitudinal study on aging (CLSA). *Can J Aging* 2009; 28(3): 221-229.

22. Donner A, Klar N, *Design and Analysis of Cluster Randomisation Trials in Health Research*, Arnold Publishing Co. London, 2000.

23. Syme, SL, Life style intervention in clinic-based trials. *American Journal of Epidemiology* 1978; 108, 87–91.

24. Donner A, Brown KS, Brasher P, A methodological review of non-therapeutic intervention trials employing cluster randomization, 1979–1989. *International Journal of Epidemiology* 1990; 19(4), 795–800.

25. Taljaard M, Donner A, Klar N, Imputation strategies for missing continuous outcomes in cluster randomized trials. *Biom J* 2008; 50(3): 329-345.

26. Andridge RR, Quantifying the impact of fixed effects modeling of clusters in multiple imputation for cluster randomized trials. *Biom J* 2011; 53(1): 57-74.

27. Ma J, Akhtar-Danesh N, Dolovich L, Thabane L, CHAT investigators. (2011) Imputation strategies for missing binary outcomes in cluster randomized trials. *BMC Med Res Methodol* 11: 18.

28. Austin PC, A comparison of the statistical power of different methods for the analysis of cluster randomization trials with binary outcomes. *Statistics in Medicine* 2007; 26: 3550-3565.

29. Bellamy SL, Gibbard R, Hancock L, Howley P, Kennedy B, Klar N, Lipsitz S, Ryan L, Analysis of dichotomous outcome data for community intervention studies. *Statistical Methods for Medical Research* 2000; 9:135–59.

30. Ukoumunne OC, Carlin JB, Gulliford MC. A simulation study of odds ratio estimation for binary outcomes from cluster randomized trials. Stat Med. 2007; 26(18):3415-28

CHAPTER 2

Power Analysis for Population Based Longitudinal Study Investigating Gene-Environment Interaction Effects on Chronic Diseases

Jinhui Ma [1, 2, 3], Lehana Thabane [1, 3, 4, 5], Joseph Beyene [1], Parminder Raina [1, 2 *]

[1] Department of Clinical Epidemiology and Biostatistics, McMaster University, Hamilton, Ontario, Canada

[2] McMaster University Evidence-based Practice Center, Hamilton, Ontario, Canada

[3] Biostatistics Unit, St. Joseph's Healthcare Hamilton, Hamilton, Ontario, Canada

[4] Centre for Evaluation of Medicines, St Joseph’s Healthcare Hamilton, Ontario, Canada

[5] Population Health Research Institute, Hamilton Health Sciences, Hamilton, Ontario, Canada

Corresponding author:

Parminder Raina

50 Main Street East, Room 310

Hamilton, Ontario, Canada L8N 1E9

Email: praina@mcmaster.ca

Summary

Background

The Canadian Longitudinal Study on Aging (CLSA) was launched as a platform to investigate complexities of the aging process with one of major interests in investigating effects of gene-environment interaction on the incidence of diseases. Rigorous statistical power estimation is crucial to ensure that such a study is scientifically useful and cost-effective, since human genome epidemiology is expensive. However, conventional sample size calculations for longitudinal study can seriously overestimate the statistical power due to overlooking the measurement error, unmeasured etiological determinants, and competing events that can impede the occurrence of the event of interest.

Methods

Based on an irreversible illness-death model, this simulation-based study takes into account the above analytic complexity and provides both accurate and realistic power profile for the CLSA.

Results

Given statistical power of 80%, significance level of 0.05 for environmental risk exposure and 0.0001 for genotype risk exposure and gene-environment interaction, the design of CLSA, which involves 30,000 participants measured every three years for at least twenty years, enable moderate ($1.5<HR\leq2.0$) or large ($2.0<HR\leq3.0$) hazard ratio (HR) to be detected for environmental risk exposures. For genotype risk exposure, the CLSA is capable of detecting moderate HR only when the incidence of disease is high, or

the prevalence of genotype risk factor is high (≥0.1). For gene-environment interaction, even large HR cannot be detected when the prevalence of genotype and environmental risk factors is low (<0.1). Misclassification on risk factors substantially reduces the statistical power. The HRs for designs involving data collection every three years are slightly larger than those obtained assuming exact event time is observed.

Discussion

Improvement on the study design and implementation, synthesis of information with other human genome studies are recommended to increase the capacity for investigating the effect of determinants on chronic diseases.

Keywords: longitudinal study, illness-death model, statistical power, minimum detectable hazard ratio, left truncation, human genome

1. Introduction

Biological and technological advances over the past decade, such as the sequencing of the human genome, have increased researchers' ability to study aging in all its complexity. The importance of studying gene-environment interactions (G×E) in the context of aging related chronic diseases has been emphasized since they typically occurred as a result of the interaction between an individual's genetic make-up and detrimental environments [1]. Sample sizes to detect genetic main effects or G×E with sufficient statistical power are expected to be extraordinarily large (up to several hundred thousand) [2]. Therefore, many genetic association studies were susceptible to lack of sufficient statistical power [3]. Very few longitudinal studies of aging to date have collected biomarker, genetic or epigenetic data to elucidate the process of aging and how biological processes interact with physical and psychosocial environment to produce deleterious health outcomes. Unlike early association studies in which individuals were not tracked over time and all measurements on each participant were made at one point in time, the longitudinal design enables researchers to separate the changes over time within subjects (i.e. aging effects) from differences between subjects at baseline (i.e. cohort effects). It also allows researchers to create the most comprehensive and insightful framework for understanding the mechanisms by which genome function can be altered during aging [4, 5].

Although massive reductions in genotyping costs, prospect cohort study remains limited by the cost of proper phenotyping [6]. Therefore, rigorous sample size or statistical power estimation is crucial to ensure that such a study is scientifically useful and cost-effective. Determinants of the required sample size for a longitudinal study include: (1) the study objectives (to provide reliable sample size calculation, an appropriate statistical test for the hypotheses of interest, which should be established to reflect the study objectives, is necessarily derived under the study design); (2) the type of endpoint/outcome (continuous, binary, categorical, or survival); (3) variation of the study population; (4) type I error, which is the probability of rejecting the null hypothesis when it is true, and type II error, which is the probability of not rejecting the null hypothesis when it is false; (5) the minimum clinically important effect size; and (6) measurement error in outcomes and risk exposures; (7) length of follow-up; (8) frequency and timing of repeated measurements on participants; (9) the presence and nature of the correlation between repeated measurements for the same subject; (10) the attrition due to mortality and loss to follow-up; and (11) unmeasured etiological determinants. However, conventional sample size calculations for longitudinal studies can seriously overestimate the statistical power due to overlooking some of the above determinants of sample size, especially the measurement error, unmeasured etiological determinants, and competing events that can impede the occurrence of the event of interest.

The Canadian longitudinal study on aging (CLSA) is a national multi-disciplinary study and has been launched as a platform to investigate the complexities of the aging process and improve the understanding of the transitions and trajectories of healthy aging [7]. The CLSA will consist of a national stratified random sample of 50,000 Canadian women and men between the age of 45 and 85 at the time of recruitment (baseline). Participants will undergo repeated waves of data collection every three years and will be followed for at least twenty years, or until death. All participants will be asked to provide a common set of information on demographic, social, physical/clinical, psychological, economic, and health service utilization aspects relevant to health and aging. Of the 50,000 participants, 30,000 (the CLSA comprehensive cohort) will also be asked to provide additional in-depth information through physical examinations and biological specimen collection. The choice of measurement frequency, i.e. every three years, balances the need to have a short enough interval to capture important changes and map trajectories with the practical consideration of the time required to complete a wave of data collection. The inclusion of study participants as young as 45 years of age at baseline is motivated by the desire to capture mid-life experiences prospectively, since important changes are known to influence outcomes later in life occur during this period. The lower age limit will also permit inclusion of individuals who are part of the baby boom cohort (i.e., those born between 1946 and 1964), who will be 47 to 65 of age in 2011. The upper limit includes individuals entering their senior years who are making the transition into retirement, who are already retired, and who have already reached old age. In the CLSA Comprehensive, self-reported diagnosis of chronic conditions will be supplemented with a disease-specific questionnaire and physical test measures.

Based on an irreversible illness-death model, the objective of this simulation-based study is to determine the statistical power profile of the CLSA, explore how to increase the statistical power through improving the design and implementation of the longitudinal study, and provide empirical guidance for designing new initiated population genomics cohort studies.

2. Methods and simulation

Both the simulation and analytical models are based on an irreversible illness-death model and implemented in a combination of SAS 9.2 (Cary, NC) and R 2.11 to achieve high computational efficiency.

The irreversible illness-death model is widely used in the medical literature to describe the progression of incurable diseases over time between three states: "healthy", "diseased", and "dead" (absorbing state). In this paper, our interest lies in the transition from "healthy" to "diseased", while the transition from "healthy" to "dead" is considered as a competing risk. Full specification of the mathematical models used for simulating and analyzing data are provided in Appendix I.

The risk of developing an age-related chronic disease for a subject increases over time. This must be captured in the statistical analysis especially when the follow-up time is long. Therefore, the transition time between two given states is assumed to follow a Weibull distribution with shape parameter larger than one in this simulation study. In addition, the time a subject initially comes under observation in a population-based cohort study usually does not coincide with the time when the subject becomes at risk of the disease of interest. Therefore, the time when a study is started (baseline) may not be an appropriate time origin in survival analysis. Alternatively, a specific age, such as 45 in this simulation study, may be a reasonable choice of time origin since the aging process starts at that time as conventionally believed. In this case, the elapsed time from the specified age to the event of interest is the survival time, and the delayed entry (subjects enter the study after the specific age) is considered as left-truncation occurring at the age of entry into the study [8]. Throughout this paper, we define state “healthy” as free of a particular disease and assume it starts from age 45. Age minus 45 is taken as the time scale for transition from “healthy” to “diseased” and from “healthy” to “dead”. The delayed entry is considered as left-truncation occurring at the age of entering into the CLSA.

In this simulation, parameters are carefully chosen to mimic the evolution of the CLSA comprehensive cohort. An instantaneous loss to follow-up rate of 0.005 per year is assumed (according to the information provided by Statistics Canada for the National Population Health Survey (NPHS) for the period 1994-1995 to 2000-2001), which results in about 8% participants lost to follow-up by the end of the CLSA. To incorporate this in the simulation, we assume the time to loss to follow-up follows an exponential distribution with rate parameter of 0.005. Both environmental and genotype risk factors are assumed to be dichotomous, which leads to the least statistical power compared with continuous measurements of these risk factors. Choices of the prevalence of both risk factors are 0.01, 0.1, and 0.2 to present very rare, common, and very common risk exposures respectively. To keep the simulation study simple yet representative, power profile of the CLSA for detecting three diseases are investigated. They are diabetes, dementia, and Parkinson’s disease, which present diseases with relatively quick, relatively slow, and very slow progression from “healthy” to “diseased”. Choices of the prevalence of diseases are 0.02 for dementia and Parkinson’s disease, and 0.14 for diabetes. Assuming 30,000 subjects are randomly sampled from the Canadian population between the age of 45 to 85, the expected number of prevalent cases at baseline, and death and incident cases at each year during the study period can be estimated based on the disease prevalence, incidence, and mortality of this population. The cumulative expected number of incident cases for these diseases during the follow-up period is presented in Figure 1. The Weibull scale and shape parameters for transition from “healthy” to “diseased” or “dead” can then be obtained by fitting Weibull regression without adding covariates (examples and R code are provided in Appendix II). For transition from “healthy” to “diseased”, the Weibull scale and shape parameters are 65 and 2.0 for diabetes, 48 and 5.6 for dementia, 130 and 3.3 for Parkinson’s disease. For transition from “healthy” to “dead”, the Weibull scale and shape parameters are fixed at 42 and 4.3. A log-normal frailty, modeled through a random effect with variance reflecting a 10-fold ratio in baseline risk between individuals on 97.5% and 2.5% population percentile, is assumed when simulating the data to present the unmeasured etiological determinants. Misclassification of exposure can be categorized into non-differential (the probability or degree of misclassification is the same among diseased and non-diseased subjects) and differential (the probability or degree of misclassification is not the same among the diseased and non-diseased) [9]. It is typically thought to be non-differential in cohort studies since exposure assessment is independent of the diagnosis of diseases. Choices of misclassification rates for environmental and genotype risk exposures are 0.1 and 0.01 respectively. Repeated measurements are assumed to be taken at year 0 (baseline), 3, 6, 9, 12, 15, 18, and 21. The hazard ratio for main effects is set to vary from 1 to 3. When investigating the power profile for gene-environment interaction, we fix hazard ratios for both main effects at 1.5 and vary the hazard ratio for interaction from 1 to 10. To achieve a reasonable degree of precision, 1000 datasets for each scenario are simulated. Within each dataset, thirty thousand subjects are generated to present the sample size of the comprehensive cohort of the CLSA. Details of simulation procedures are illustrated in Figure 2.

4. Results

Figures 3, 4, 5 represent the statistical power profile for the environmental risk factor, the genotype risk factor, and the gene-environment interaction on detecting disease with relatively quick (diabetes), relatively slow (Dementia), and very slow (Parkinson's disease) progress from "healthy" to "diseased". The significance level for environmental risk factor, genotype risk factor, and their interaction are defined as 0.05, 0.0001, and 0.0001 respectively. The power profiles indicate shows the following: (1) higher prevalence of risk factor is associated with higher statistical power; (2) misclassification on risk factors substantially reduces the statistical power, especially for risk factors with low prevalence (≤0.01); (3) the hazard ratios (HR) for designs involving data collection every three years are slightly larger comparing with those obtained assuming the exact event time can be observed; and (4) for disease with relatively fast progression from "healthy" to "diseased", the statistical power is higher comparing with disease with slow progression.

The minimum detectable hazard ratio (MDHR) is defined as the smallest hazard ratio that can be detected with specific sample size (30,000) and statistical power (80%) at a chosen level of statistical significance. Based on the power profiles, we obtain

MDHRs for genotype risk factor, environmental risk factor, and their interaction (present in Table 1). We categorize the MDHR into four categories: small (1<MDHR≤1.5), moderate (1.5<MDHR≤2.0), large (2.0<MDHR≤3.0), and very large (MDHR>3.0). The results show that the design of CLSA, which involves 30,000 participants measured every three years for at least twenty years, enable small or moderate HR to be detected for environmental risk exposure when the prevalence of risk factor is relatively high (≥0.1). When the prevalence of environmental risk factor is low (≤0.01), only large or very large HR can be detected. For genotype risk exposure, the CLSA is capable of detecting moderate or large HR only when the incidence of disease is high (i.e. relatively fast progression from "healthy" to "diseased"), or the prevalence of genotype risk factor is high (≥0.1). For gene-environment interaction, the CLSA is able to detect large or very large HR when the prevalence of risk factor is high (≥0.1). However, even very large HR may not be detected when the prevalence of both genotype and environmental risk factors are low (≤0.01).

5. Discussion

In this paper, we investigate the MDHR for environmental and genotype risk exposures and their interaction on the transition from "healthy" to "diseased" for the CLSA, considering the transition from "healthy" to "dead" as a competing risk. Our results show that the design choice of measuring subjects every three years for at least twenty years slightly increases the MDHRs compared with the design assuming the exact time of occurrence of disease is known. It suggests the frequency and timing of the repeated measurements in the CLSA may be a reasonable choice in the sense of reducing the cost substantially but not losing much statistical power. The effect of frequency and timing of repeated measurements is closely related to the mean sojourn in "healthy" and "diseased" states. For example, if the frequency of repeated measurements, which is three year for the CLSA, is considerably larger than the mean sojourn in the state of diseased, it is very likely that both the transition from "healthy" to "diseased" and the subsequent transition to "dead" occur within the same data collection interval. Consequently, the statistical power will decrease since the transition from "healthy" to "diseased" will not be observed. Therefore, researchers should obtain prior knowledge about the mean sojourn time in each of the transient states and use this knowledge to guide the design choice of the frequency and timing of repeated measurements. In addition, the larger the time interval between two adjacent time-points for data collection, the higher chance that the subjects are lost to follow-up within that time interval. In this case, any transition within that time interval will not be observed. Though we illustrate that the frequency and timing of the CLSA may be a reasonable design choice, it may not be optimal. When assessments are expensive, increasing the frequency of repeated measurements will increase the cost. Therefore, the optimal design may only be determined once a cost function is specified. The above findings also suggests that without increasing the cost, higher statistical power may be achieved through increasing the frequency of measurements for subjects with high risk of developing diseases or loss to follow-up and slightly decreasing the frequency of measurements for other subjects. Since the incidence of aging related chronic diseases is usually higher for older people, to achieve higher statistical power, researcher may recruit higher proportion of seniors for achieving more incident cases in a relatively short period. However, the above finding suggests that increasing the number of seniors in the sample may also cause a decrease in power due to two reasons: (1) since more people developed diseases at baseline, the transition from "healthy" to "diseased" will not be observed during the follow-up for those subjects; and (2) the incidence of new cases may not be observed since seniors are more likely to develop the disease and then die or lost to follow-up before the next measurement in 3 years comparing with mid-age subjects.

On the other hand, larger MDHRs when assuming subjects are under repeated measurements every three years comparing with those when assuming subjects are under continuous monitoring implies that smaller MDHRs can be achieved by increasing the frequency of the repeated measurements. This is consistent with the findings from van den Hout *et al* [10]. However, they also conclude that it is not always necessary to have a long follow-up for the study or a large sample size, and relatively short follow-up time can still be used if the time intervals between measurements are not too wide. Our findings do not lead to the same conclusion. This is because their simulation was based on a different statistical model ─ reversible illness-death model. This model is commonly used to model the progression of reversible diseases. Since a subject may experience several transitions from "healthy" to "diseased" and "diseased" to "healthy" during the course of the study, large amount of events or state transitions may be observed through relatively small sample size but frequent measurements. However, for irreversible illness-death model, which is used to model the progression of irreversible diseases, increasing the frequency of the measurements may only lead to limited increase in the number of events. Among the three factors which may influence the statistical power of a study ─ length of follow-up (i.e. the duration of the study), frequency of measurements, or sample size, increasing the frequency or duration may only increase power to an upper limit that depends on the progression process of the disease for the study population, whereas increasing sample size can raise power toward 1.0. In general, increasing the sample size means the cost of recruitment increases; increasing the length of follow-up increases the risk of attrition and the cost of tracking participants. When assessments are expensive, increasing the frequency of repeated measurements will increase the cost. Therefore, the optimal design may only be determined once a cost function is specified. Further investigation could be done to determine what design choices are most efficient for the CLSA.

In this simulation study, we find that misclassification of the environmental and genotype risk exposures substantially increases the MDHR. This is consistent with the finding from Garcia-Closas [11] *et al* that misclassification of environmental or genotype risk factors can substantially increase the sample size required to evaluate gene-environment interaction in case-control studies. The magnitude of the increase in sample size is highly dependent on the misclassification rate on the risk exposures. Therefore, improving the accuracy in measuring both genotype and environmental risk exposures is critical, especially for valid assessment of gene-environment interaction.

This project has some limitations. First, we assume the loss-to-follow-up rate is constant overtime. In practice, it may change with time and other variables. Second, accrual period is not considered since we assume all subjects enter the study at the beginning of the CLSA. However, the influence of ignoring accrual period on the estimation of MDHR may be compensated by assuming all the subjects are followed up for 21 years in the simulation study. Third, we only estimate the MDHR for time-independent risk exposures. Fifth, we consider the misclassification of the risk exposure in this simulation study. However, the measurement error for the response variable, i.e. the accuracy of the disease diagnosis, is not considered in the present study.

To the best of our knowledge, this project is the first attempt to investigate the power profile of a population based longitudinal study using the illness-death model on detecting the environmental risk factor, genotype risk factor, and their interaction on the health state transitions. This simulation study provides a realistic power by taking into account the measurement error, unmeasured etiological determinants, and competing events that can impede the occurrence of the event of interest, which are usually ignored by traditional sample size and statistical power calculation.

6. Conclusions

Improvement on the study design and implementation, synthesis of information with other human genome studies are recommended to increase the capacity for investigating the effect of determinants on chronic diseases. Findings from the present study provide guidance on designing similar population based longitudinal studies.

Competing interests

The authors declare that they have no competing interests.

Acknowledgements

This study was supported in part by funds from the Canadian Longitudinal Study on Aging **(**which **is** funded in part by the Canadian Institutes of Health Research (CIHR)**)**, Canadian Network and Centre for Trials Internationally (CANNeCTIN) program, and the Drug Safety and Effectiveness Cross-Disciplinary Training (DSECT) Program in the form of studentship on training awards. Dr. Lehana Thabane is a clinical trials mentor for the Canadian Institutes of Health Research (CIHR). Dr. Parminder Raina holds a Raymond and Margaret Labarge Chair in Research and Knowledge Application for Optimal Aging, a Canada Research Chair in GeroScience.

References

1. Grigorenko EL. The Inherent Complexities of Gene–Environment Interactions. *Journals of Gerontology: SERIES B* 2005; 60B (Special Issue I):53–64.
2. Collins FS. The case of a US prospective cohort study of genes and environment. *Nature* 2004, 249, 475–477.
3. Khoury MJ, Little J, Gwinn M, Ioannidis JP. On the synthesis and interpretation of consistent but weak gene-disease associations in the era of genome-wide association studies. *Int J Epidemiol* 2007;36:439–45.

4. Feinberg AP. Phenotypic plasticity and the epigenetics of human disease. *Nature* 2007 May 24; 447(7143):433-40.

5. Calvanese V, Lara E, Kahn A, et al. The role of epigenetics in aging and age-related diseases. *Ageing Res Rev* 2009 Oct; 8(4):268-76.

6. Thompson EA. Human genetics: overview. In: Palmer L, ed. Biostatistical genetics and genetic epidemiology. Chichester, United Kingdom: John Wiley and Sons Ltd, 2002:386–90.

7. Raina PS, Wolfson C, Kirkland SA, Griffith LE, Oremus M, et al. The canadian longitudinal study on aging (CLSA). *Can J Aging* 2009; 28(3): 221-229.

8. Lamarca R, Alonso J, Gomez G, Munoz A. Left-truncated data with age as time scale: An alternative for survival analysis in the elderly population. *J Gerontol A Biol Sci Med Sci* 1998; 53(5): M337-43.

9. Blair A, Stewart P, Lubin JH, Forastiere F. Methodological issues regarding confounding and exposure misclassification in epidemiological studies of occupational exposures. *Am J Ind Med* 2007; 50(3): 199-207.

10. van den Hout A, Matthews FE. A piecewise-constant markov model and the effects of study design on the estimation of life expectancies in health and ill health. *Stat Methods Med Res* 2009; 18(2): 145-162.

11. Garcia-Closas M, Rothman N, Lubin J. Misclassification in case-control studies of gene-environment interactions: Assessment of bias and sample size. *Cancer Epidemiol Biomarkers Prev* 1999; 8(12): 1043-1050.

Table 1. Minimum detectable hazard ratio for risk exposures

Characteristic of risk factor			Frequency of Outcome Measurement	Minimum detectable hazard ratio* Progress from Healthy to Diseased		
Risk factor	Prevalence	Measurement error		Relatively quick	Relatively slow	Very slow
Environment	0.01	0%	Continuous	1.57	2.12	2.42
		0%	Every 3 years	1.61	2.27	2.48
		10%	Every 3 years	>3.00	>3.00	>3.0
	0.1	0%	Continuous	1.21	1.27	1.40
		0%	Every 3 years	1.22	1.29	1.42
		10%	Every 3 years	1.46	1.61	1.67
	0.2	0%	Continuous	1.13	1.21	1.30
		0%	Every 3 years	1.14	1.23	1.31
		10%	Every 3 years	1.18	1.29	1.48
Genotype	0.01	0%	Continuous	1.95	>3.00	>3.00
		0%	Every 3 years	2.00	>3.00	>3.00
		1%	Every 3 years	>3.00	>3.00	>3.00
	0.1	0%	Continuous	1.46	1.55	1.60
		0%	Every 3 years	1.50	1.61	1.64
		1%	Every 3 years	1.62	1.71	1.75
	0.2	0%	Continuous	1.24	1.39	1.55
		0%	Every 3 years	1.24	1.43	1.56
		1%	Every 3 years	1.25	1.48	1.60
Gene-environment interaction	0.01[+]	0%	Continuous	>10.00	>10.00	>10.00
		0%	Every 3 years	>10.00	>10.00	>10.00
		10%,1%[#]	Every 3 years	>10.00	>10.00	>10.00
	0.1[+]	0%	Continuous	1.85	3.32	4.73
		0%	Every 3 years	2.00	3.56	5.21
		10%, 1%[#]	Every 3 years	>10.00	>10.00	>10.00
	0.2[+]	0%	Continuous	1.47	2.17	2.60
		0%	Every 3 years	1.52	2.31	2.62
		10%, 1%[#]	Every 3 years	2.28	3.31	3.37

Note:

* Minimum detectable hazard ratio is defined as the smallest hazard ratio that can be detected with statistical power of 80% at the level of statistical significance of 0.05 for environmental risk factor and 0.0001 for genotype risk factor and gene-environment interaction.

[+] Both environmental and genotype risk factor have the same specified prevalence.

[#] Misclassification rate for environmental and genotype risk factors are 10% and 1% respectively.

Figure 1. Expected number of incident cases for different diseases

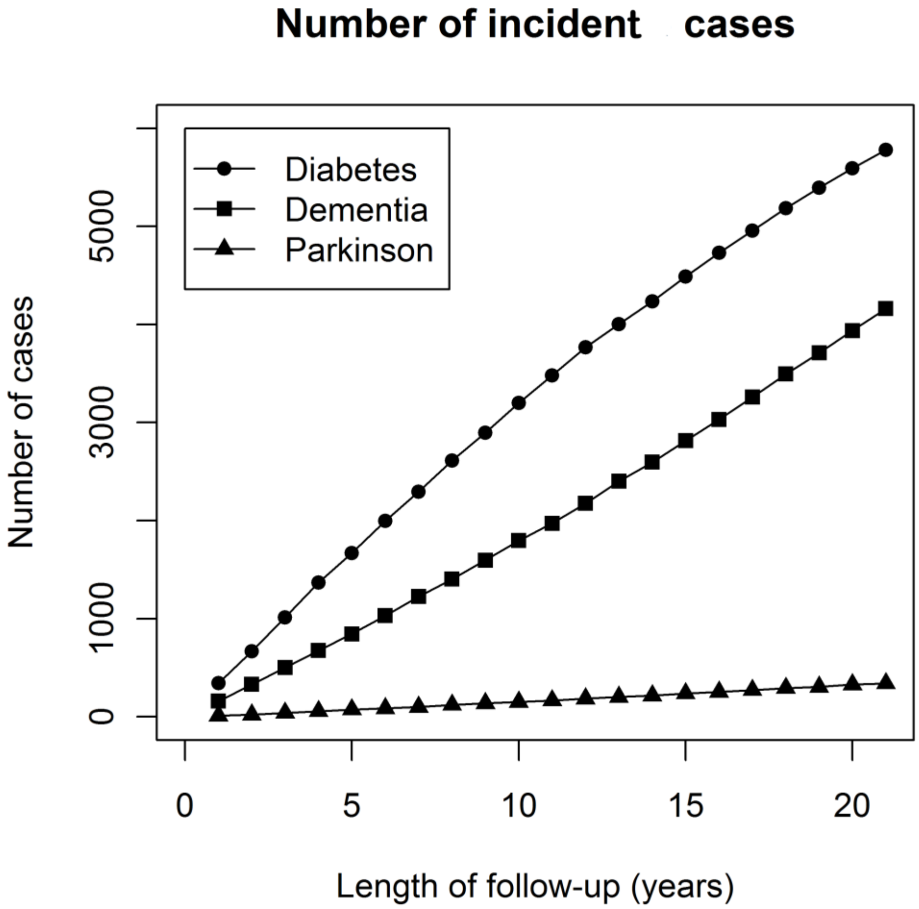
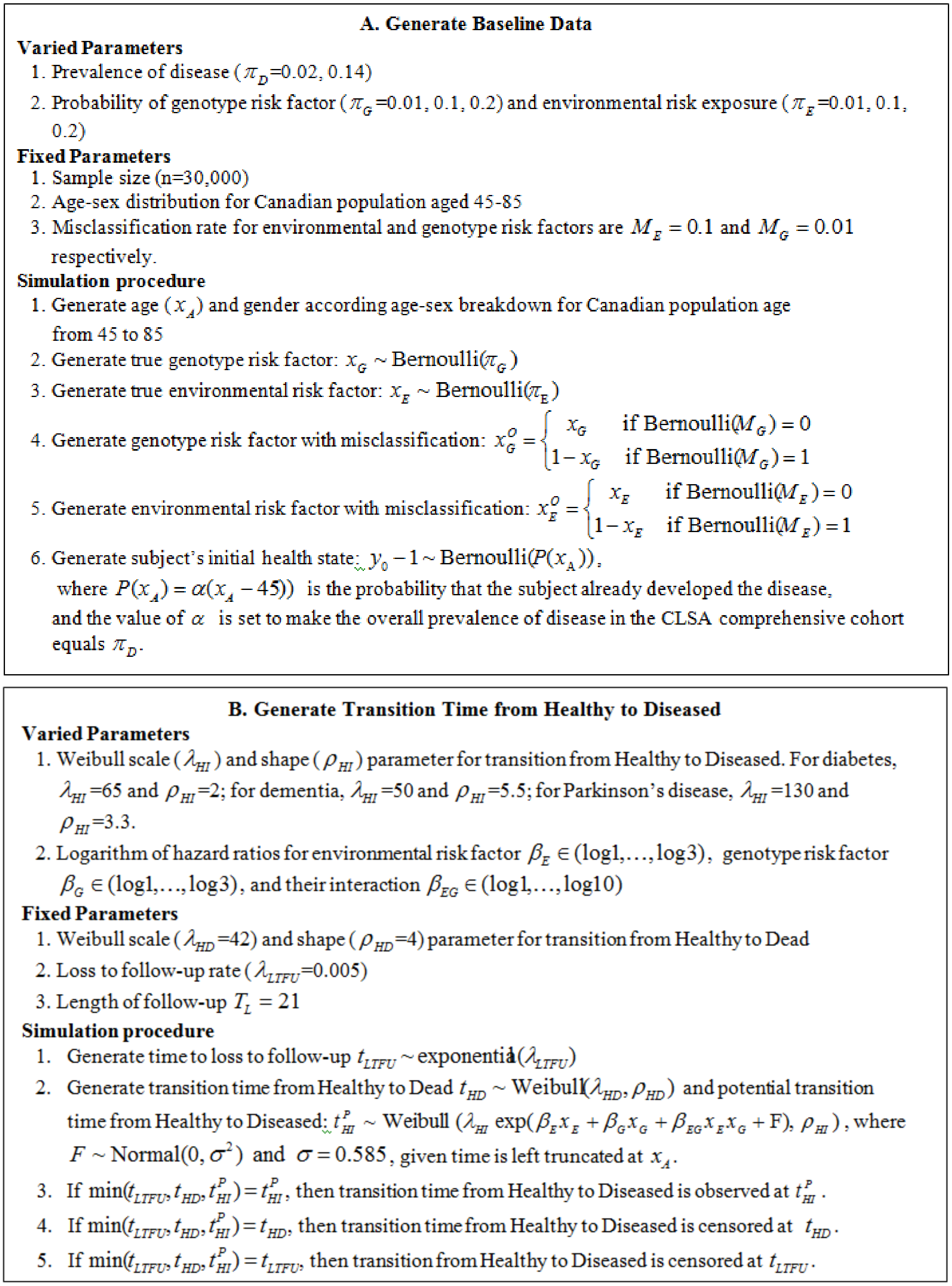


Figure 2. Simulation procedure



A. Generate Baseline Data

Varied Parameters

1. Prevalence of disease (π_D=0.02, 0.14)
2. Probability of genotype risk factor (π_G=0.01, 0.1, 0.2) and environmental risk exposure (π_E=0.01, 0.1, 0.2)

Fixed Parameters

1. Sample size (n=30,000)
2. Age-sex distribution for Canadian population aged 45-85
3. Misclassification rate for environmental and genotype risk factors are $M_E = 0.1$ and $M_G = 0.01$ respectively.

Simulation procedure

1. Generate age (x_A) and gender according age-sex breakdown for Canadian population age from 45 to 85
2. Generate true genotype risk factor: $x_G \sim \text{Bernoulli}(\pi_G)$
3. Generate true environmental risk factor: $x_E \sim \text{Bernoulli}(\pi_E)$
4. Generate genotype risk factor with misclassification: $x_G^O = \begin{cases} x_G & \text{if Bernoulli}(M_G) = 0 \\ 1 - x_G & \text{if Bernoulli}(M_G) = 1 \end{cases}$
5. Generate environmental risk factor with misclassification: $x_E^O = \begin{cases} x_E & \text{if Bernoulli}(M_E) = 0 \\ 1 - x_E & \text{if Bernoulli}(M_E) = 1 \end{cases}$
6. Generate subject's initial health state: $y_0 - 1 \sim \text{Bernoulli}(P(x_A))$, where $P(x_A) = \alpha(x_A - 45))$ is the probability that the subject already developed the disease, and the value of α is set to make the overall prevalence of disease in the CLSA comprehensive cohort equals π_D.

B. Generate Transition Time from Healthy to Diseased

Varied Parameters

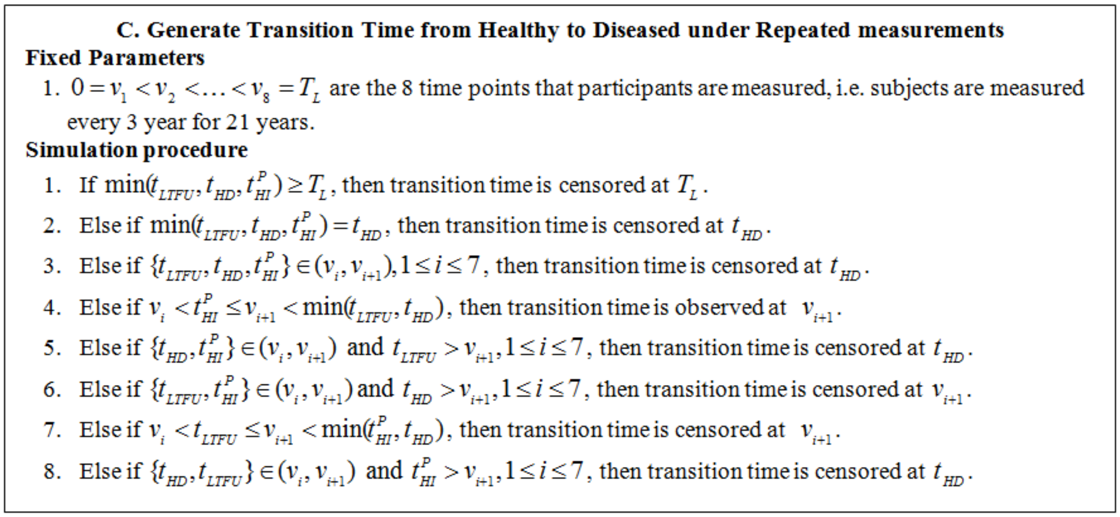
1. Weibull scale (λ_{HI}) and shape (ρ_{HI}) parameter for transition from Healthy to Diseased. For diabetes, λ_{HI}=65 and ρ_{HI}=2; for dementia, λ_{HI}=50 and ρ_{HI}=5.5; for Parkinson's disease, λ_{HI}=130 and ρ_{HI}=3.3.
2. Logarithm of hazard ratios for environmental risk factor $\beta_E \in (\log 1, \ldots, \log 3)$, genotype risk factor $\beta_G \in (\log 1, \ldots, \log 3)$, and their interaction $\beta_{EG} \in (\log 1, \ldots, \log 10)$

Fixed Parameters

1. Weibull scale (λ_{HD}=42) and shape (ρ_{HD}=4) parameter for transition from Healthy to Dead
2. Loss to follow-up rate (λ_{LTFU}=0.005)
3. Length of follow-up $T_L = 21$

Simulation procedure

1. Generate time to loss to follow-up $t_{LTFU} \sim \text{exponential}(\lambda_{LTFU})$
2. Generate transition time from Healthy to Dead $t_{HD} \sim \text{Weibull}(\lambda_{HD}, \rho_{HD})$ and potential transition time from Healthy to Diseased: $t_{HI}^P \sim \text{Weibull}\ (\lambda_{HI} \exp(\beta_E x_E + \beta_G x_G + \beta_{EG} x_E x_G + \text{F}), \rho_{HI})$, where $F \sim \text{Normal}(0, \sigma^2)$ and $\sigma = 0.585$, given time is left truncated at x_A.
3. If $\min(t_{LTFU}, t_{HD}, t_{HI}^P) = t_{HI}^P$, then transition time from Healthy to Diseased is observed at t_{HI}^P.
4. If $\min(t_{LTFU}, t_{HD}, t_{HI}^P) = t_{HD}$, then transition time from Healthy to Diseased is censored at t_{HD}.
5. If $\min(t_{LTFU}, t_{HD}, t_{HI}^P) = t_{LTFU}$, then transition from Healthy to Diseased is censored at t_{LTFU}.



C. Generate Transition Time from Healthy to Diseased under Repeated measurements

Fixed Parameters

1. $0 = v_1 < v_2 < \ldots < v_8 = T_L$ are the 8 time points that participants are measured, i.e. subjects are measured every 3 year for 21 years.

Simulation procedure

1. If $\min(t_{LTFU}, t_{HD}, t_{HI}^P) \geq T_L$, then transition time is censored at T_L.
2. Else if $\min(t_{LTFU}, t_{HD}, t_{HI}^P) = t_{HD}$, then transition time is censored at t_{HD}.
3. Else if $\{t_{LTFU}, t_{HD}, t_{HI}^P\} \in (v_i, v_{i+1}), 1 \leq i \leq 7$, then transition time is censored at t_{HD}.
4. Else if $v_i < t_{HI}^P \leq v_{i+1} < \min(t_{LTFU}, t_{HD})$, then transition time is observed at v_{i+1}.
5. Else if $\{t_{HD}, t_{HI}^P\} \in (v_i, v_{i+1})$ and $t_{LTFU} > v_{i+1}, 1 \leq i \leq 7$, then transition time is censored at t_{HD}.
6. Else if $\{t_{LTFU}, t_{HI}^P\} \in (v_i, v_{i+1})$ and $t_{HD} > v_{i+1}, 1 \leq i \leq 7$, then transition time is censored at v_{i+1}.
7. Else if $v_i < t_{LTFU} \leq v_{i+1} < \min(t_{HI}^P, t_{HD})$, then transition time is censored at v_{i+1}.
8. Else if $\{t_{HD}, t_{LTFU}\} \in (v_i, v_{i+1})$ and $t_{HI}^P > v_{i+1}, 1 \leq i \leq 7$, then transition time is censored at t_{HD}.

Figure 3. Power profile for environmental risk factor

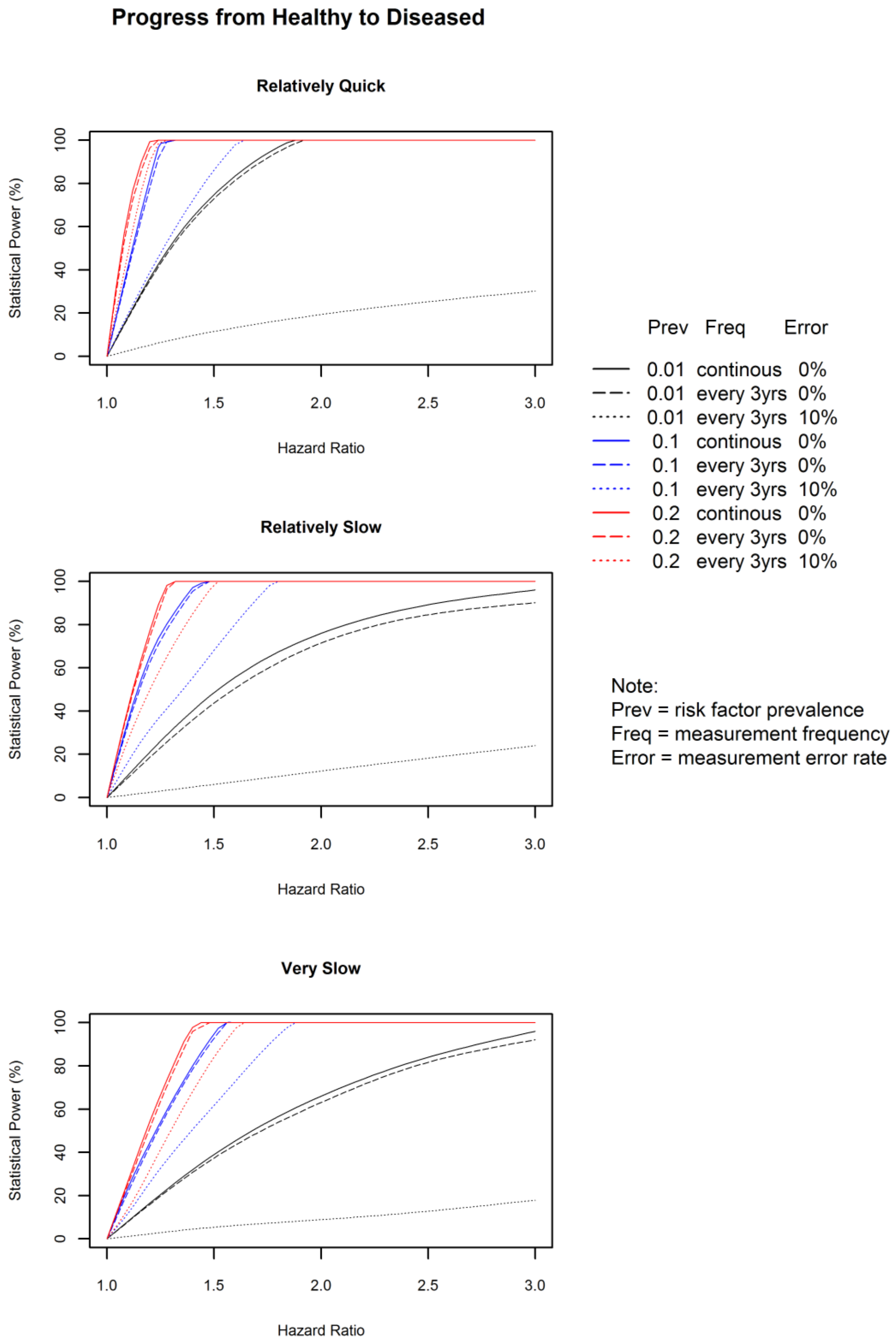


Figure 4 Power profile for genotype risk factor

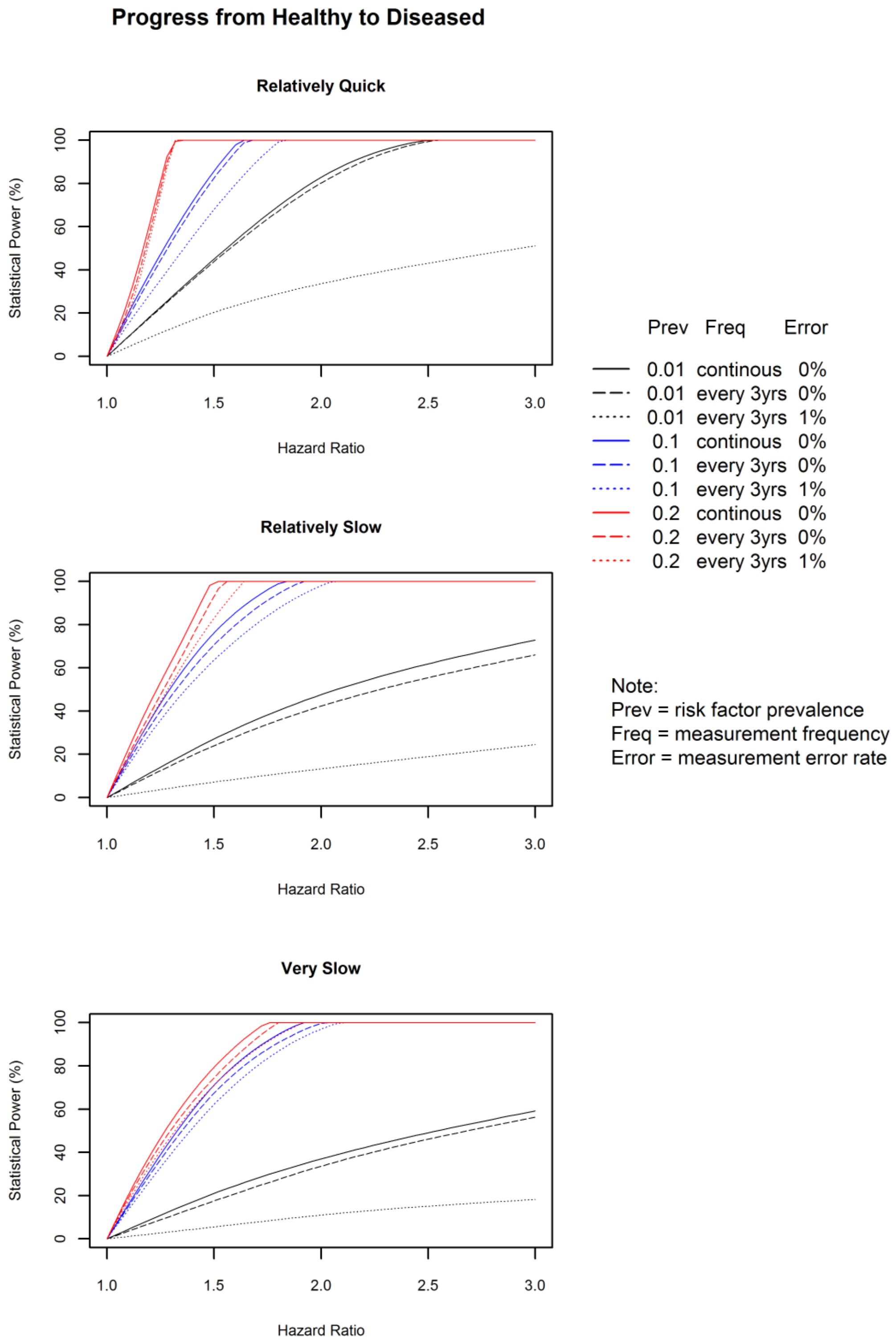
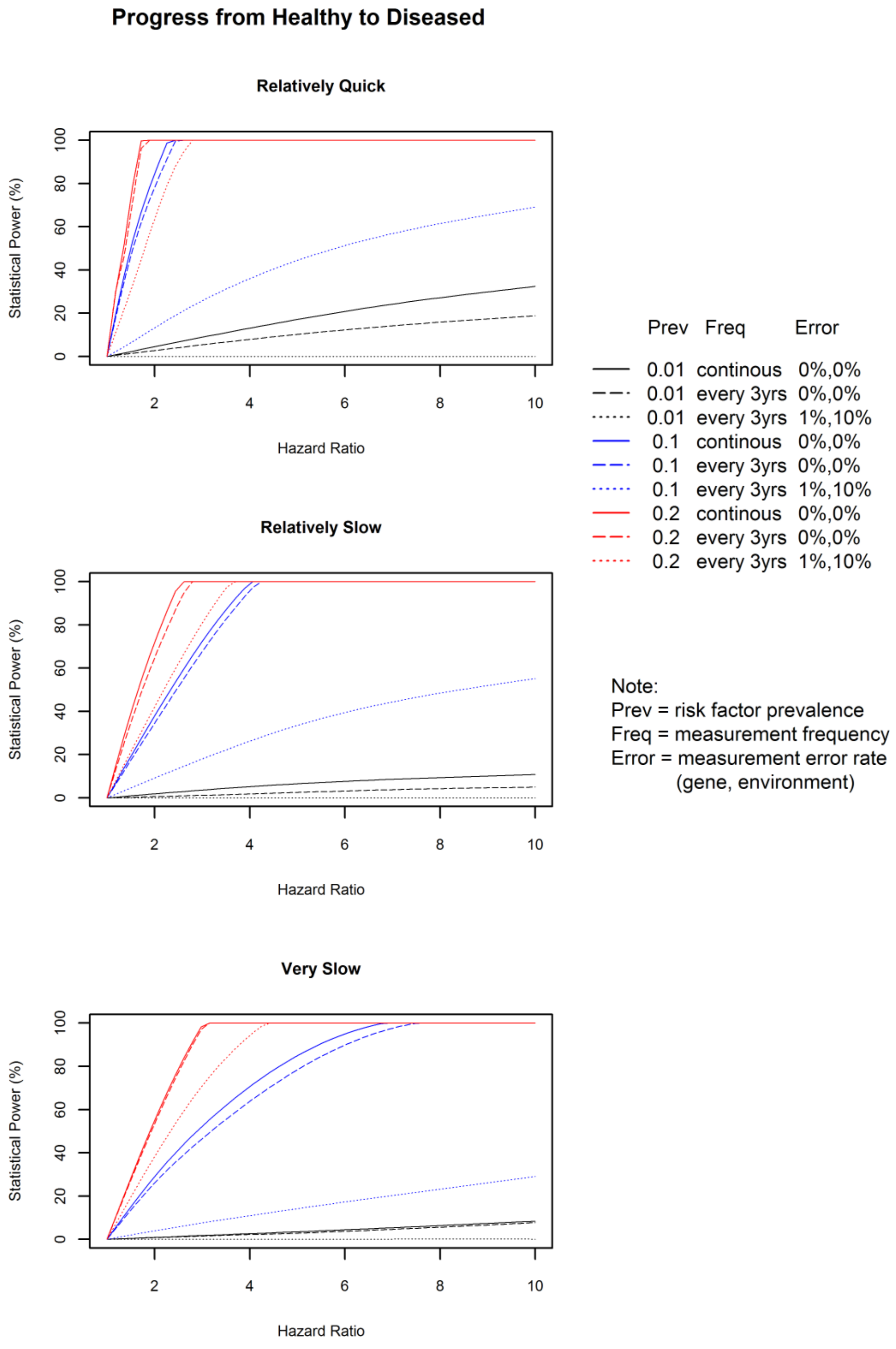


Figure 5. Power profile for gene-environment interaction



Note:
Prev = risk factor prevalence
Freq = measurement frequency
Error = measurement error rate
(gene, environment)

Appendix I Mathematical supplement

1. Illness-Death model

Let *t* denote the time since entry into a state and $Z(t)$ denote the state at time *t* for a subject. The movement between states is governed by transition intensities, which may depend on time *t* and a set of individual-level explanatory variables *X*, $q_{rs}(t \mid X)$ with $r, s \in \{1, 2, 3\}$, which represents the instantaneous risk of moving from state *r* to state $s \neq r$: $q_{rs}(t \mid X) = \lim_{\Delta t \to 0} \frac{P(Z(t+\Delta t) = s \mid Z(t) = r, X)}{\Delta t}$. The $q_{rs}(t \mid X)$ form an $n \times n$ matrix $Q(t \mid X)$ whose rows sum to zero, so that the diagonal entries are defined by $q_{rr}(t \mid X) = -\sum_{r \neq s} q_{rs}(t \mid X)$. The transition probabilities for a time intervals $(t_1, t_2]$ are given by the $n \times n$ matrix $P(t_1, t_2 \mid X) = \exp[(t_2 - t_1)Q(t_1 \mid X)]$, with entries $p_{rs}(t_1, t_2 \mid X) = \Pr(Z(t_2) = s \mid Z(t_1) = r, X)$.

The illness-death model is one form of multi-state model and is widely used in the medical literature to describe the disease progression over time between three states: Healthy, Diseased, and Dead (absorbing state). We focus on the irreversible illness-death model in this paper to model the progression process of incurable diseases. Let state 1, 2, 3 present states of Healthy, Diseased, and Dead respectively and *t* denote the time since the entry into a state and $Z(t)$ denote the state at time *t* for a subject. The movement between states is governed by transition intensities, which may depend on time *t* and a set of individual-level explanatory variables *X*, $q_{rs}(t \mid X)$ with $r, s \in \{1, 2, \ldots, n\}$, which represents the instantaneous risk of moving from state *r* to state $s \neq r$: $q_{rs}(t \mid X) = \lim_{\Delta t \to 0} \frac{P(Z(t+\Delta t) = s \mid Z(t) = r, X)}{\Delta t}$. The $q_{rs}(t \mid X)$ form an 3×3 transition intensity matrix expressed as

$$Q(t \mid X) = \begin{pmatrix} -q_{12}(t \mid X) - q_{13}(t \mid X) & q_{12}(t \mid X) & q_{13}(t \mid X) \\ 0 & -q_{23}(t \mid X) & q_{23}(t \mid X) \\ 0 & 0 & 0 \end{pmatrix}.$$

For time interval $[t_1,\ t_2]$, let $p_{rs}(t_1, t_2 \mid X) = \Pr(Z(t_2) = s \mid Z(t_1) = r, X)$, the transition probability matrix can be expressed as

$$P(t_1, t_2 \mid X) = \begin{pmatrix} p_{11}(t_1, t_2 \mid X) & p_{12}(t_1, t_2 \mid X) & 1 - p_{11}(t_1, t_2 \mid X) - p_{12}(t_1, t_2 \mid X) \\ 0 & p_{22}(t_1, t_2 \mid X) & 1 - p_{22}(t_1, t_2 \mid X) \\ 0 & 0 & 1 \end{pmatrix}$$

Where $p_{11}(t_1, t_2 \mid X) = \exp(-\mathrm{Q}_{12}(t_1, t_2 \mid X) - \mathrm{Q}_{13}(t_1, t_2 \mid X))$

$p_{22}(t_1, t_2 \mid X) = \exp(-\mathrm{Q}_{23}(t_1, t_2 \mid X))$

$$p_{12}(t_1,t_2 \mid X) = \int_{t_1}^{t_2} p_{11}(t_1,t \mid X) q_{12}(t \mid X) p_{22}(t,t_2 \mid X)dt$$

$\mathrm{Q}_{i\,\mathrm{j}}(t_1,t_2 \mid X) = \int_{t_1}^{t_2} q_{i\,\mathrm{j}}(t \mid X)dt$ for $1 \le i < j \le 3$ is the cumulative hazard function for transition from state *i* to state *j*.

2. Weibull distribution with left truncation

Suppose random variable $W \sim \mathrm{Weibull}(\lambda,\rho)$, where λ and ρ are scale and shape parameters respectively. Its probability density function (pdf) $f_W(\cdot)$ and cumulative distribution function (cdf) $F_W(\cdot)$ are:

$$f_W(w) = \frac{\rho}{\lambda}\left(\frac{w}{\lambda}\right)^{\rho-1} \exp\left(-\left(\frac{w}{\lambda}\right)^{\rho}\right) \text{ and } F_W(w) = 1-\exp\left(-\left(\frac{w}{\lambda}\right)^{\rho}\right).$$

Let *T* be a random variable obtained by left truncating *W* at $0 \le l < \infty$. Its pdf $f_T(\cdot)$ and cdf $F_T(\cdot)$ are given by

$$f_T(t) = \begin{cases} \dfrac{f_W(t)}{1-F_W(l)} & \text{if } t \ge l, \\ 0 & \text{otherwise,} \end{cases} \quad \text{and } F_T(t) = \frac{F_W(t)-F_W(l)}{1-F_W(l)}.$$

Both W and T have the same transition intensity $q(w;\lambda,\rho) = q(t;\lambda,\rho,l) = \frac{\rho}{\lambda}\left(\frac{w}{\lambda}\right)^{\rho-1}$.

The survival time being left truncated at *l* can then be simulated from $t = F_T^{-1}(\mu)$, where μ is randomly sampled from Uniform(0,1).

3. Simulation

Let λ_i^{rs} and ρ_i^{rs} be the Weibull scale and shape parameter for the transition from state *r* to state *s*. For subject *i*, whose age at baseline is $l_i + 45$, the transition time from state 1 (Healthy) to state 2 (Diseased) is generated by:

$$t_i^{12} = \lambda_i^{12}(-\ln(1-\mu_i(1-F_W(l_i;\lambda_i^{12},\rho_{12})) + F_W(l_i;\lambda_i^{12},\rho_{12})))^{1/\rho_{12}},$$

where

$$\lambda_i^{12} = \lambda_{(0)}^{12}\exp(\beta_{12}^{E}x_i^{E} + \beta_{12}^{G}x_i^{G} + \beta_{12}^{EG}x_i^{E}x_i^{G} + f_i),$$

$$\mu_i \sim \mathrm{Uniform}(0,1),$$

$$f_i \sim \mathrm{Normal}(0, 0.585^2),$$

$\lambda_{(0)}^{12}$ is the baseline scale parameter which are carefully chosen to ensure the expected value of the scale parameter in the sample equals to the Weibull scale parameter estimated based on the incidence of a disease for Canadian population (See Appendix II).

The transition time from state 1 to state 3 (Dead) is generated by:

$$t_i^{13} = \lambda_i^{13}(-\ln(1-\mu_i(1-F_W(l_i;\lambda_i^{13},\rho_{13})) + F_W(l_i;\lambda_i^{13},\rho_{13})))^{1/\rho_{13}},$$

where

$\mu_i \sim \text{Uniform}(0,1)$, and λ_i^{13} is the scale parameter estimated based on the mortality of the Canadian population.

The time from entering into the study to loss to follow-up (t_i^{LTFU}) is generated by: $t_i^{LTFU} \sim \text{Exponential}(0.005)$.

4. Analysis

Let t_i be the survival time (time since age 45 to age when disease being diagnosed, or time since age 45 to age when subject die or loss to follow-up) of subject *i*. If disease is observed (*C*=0), the contribution of this individual to the likelihood is

$$f_T(t_i) = \frac{f_W(t_i)}{1-F_W(l_i)}.$$

If a subject died or lost to follow-up (*C*=1), the contribution of this individual to the likelihood is

$$1-F_T(t_i) = \frac{F_W(t_i)-F_W(l_i)}{1-F_W(l_i)}.$$

The likelihood to be maximized for all subjects is

$$L = \sum_{i=1}^{30000} \log\left(\left(\frac{f_W(t_i)}{1-F_W(l_i)}\right)^{1-C}\left(\frac{F_W(t_i)-F_W(l_i)}{1-F_W(l_i)}\right)^{C}\right).$$

Appendix II Resources and procedures for determining Weibull parameters

Table 1. Approximate distribution of the 30,000 CLSA comprehensive cohort by age and sex

Age Band (at baseline)	Approximate number of male subjects in CLSA at Baseline	Approximate number of female subjects in CLSA at Baseline	Approximate number of subjects in CLSA at Baseline	Proportion of subjects among the 30,000 cohort
45-49	3204	3216	6420	0.21
50-54	2793	2850	5643	0.19
55-59	2439	2493	4932	0.16
60-64	1824	1890	3714	0.12
65-69	1407	1518	2925	0.10
70-74	1197	1359	2556	0.09
75-79	927	1194	2121	0.07
80-84	594	939	1533	0.05
85	51	105	156	0.01

Resources: Estimated according to the age-sex break down for Canadians at the 2005 Census

Table 2. Annual incidence of dementia and Parkinson's disease (‰)

Disease	Age Group								
	45-49	50-54	55-59	60-64	65-69	70-74	75-79	80-84	85+
Incidence									
Dementia (male) [1]	0	0	0	0	3.7	14.4	24.5	32.6	70.7
Dementia (female) [1]	0	0	0	0	6.8	7.6	17.7	36.7	69.5
Parkinson (male) [2]	0.11	0.11	0.11	0.54	0.54	1.33	1.33	2.13	2.13
Parkinson (Female) [2]	0.11	0.11	0.11	0.54	0.54	1.33	1.33	2.13	2.13
Diabetes (male) [3]	0.82	1.16	1.55	1.93	2.26	2.34	2.27	2.09	1.62
Diabetes (female) [3]	0.58	0.83	1.12	1.41	1.67	1.78	1.79	1.72	1.36
Prevalence									
Dementia (male) [4]	0	0	0	0	24	24	111	111	345
Dementia (female) [4]	0	0	0	0	24	24	111	111	345
Parkinson (male) [5]	0	0	16.39	15.9	15.9	39.51	39.51	90.91	90.91
Parkinson (Female) [5]	0	0	16.39	15.9	15.9	39.51	39.51	90.91	90.91
Diabetes (male) [3]	6.2	9.5	14.0	19.1	23.7	27.1	28.5	27.8	23.2
Diabetes (female) [3]	5.1	7.4	10.7	14.2	17.8	21.3	23.1	23.4	19.9

Sources:

1. Canadian Study of Health and Aging Working Group. The incidence of dementia in Canada. Neurology 2000; 55:66-73.
2. Morens DM, Davis JM, Grandinetti A, et al. Epidemiologic observations on Parkinson's disease: incidence and mortality in a prospective study of middle aged men. *Neurology* 1996;46:1044-50.
3. http://www.phac-aspc.gc.ca/cd-mc/publications/diabetes-diabete/facts-figures-faits-chiffres-2011/chap1-eng.php#Pre
4. Lindsay J, Sykes E, McDowell I, Verreault R, Laurin D. More than the epidemiology of Alzheimer's Disease: contributions of the Canadian Study of Health and Aging. Can J Psychiatry 2004;49(2):83-91.
5. BioBasics, Government of Canada http://www.biobasics.gc.ca/english/View.asp?x=771

Table 3. Estimated Weibull parameters for transition from Healthy to Diseased or Dead

Transitions	Scale parameter	Shape parameter
Healthy to Dementia	48	5.6
Healthy to Parkinson	130	3.3
Healthy to Diabetes	65	2.0
Health to Dead	42	4.3

R code for estimating Weibull shape and scale parameters

```
set.seed(5)

### generate data to mimic patients patients transition from health to diseased ###

gen_disease_data=function(para)
{
### age-gender Prevalence and incidence of diseases ###
### para==1  Dementia ###
### para==2  Parkinson ###
### para==3  Diabetes ###

        if(para==1)
        {
                prev_male<-c(0,0,0,0,24,24,111,111,345)/1000
                prev_female<-c(0,0,0,0,24,24,111,111,345)/1000

                inci_male<-c(0,0,0,0,3.7,14.4,24.5,32.6,70.7,70.7)/1000
                inci_female<-c(0,0,0,0,6.8,7.6,17.7,36.7,69.5,69.5)/1000
        }

        if(para==2)
        {
                prev_male<-c(0,0,16.39,15.9,15.9,39.51,39.51,90.91,90.91)/1000
                prev_female<-c(0,0,16.39,15.9,15.9,39.51,39.51,90.91,90.91)/1000

                inci_male<-c(0.11,0.11,0.11,0.54,0.54,1.33,1.33,2.13,2.13,2.13)/1000
                inci_female<-c(0.11,0.11,0.11,0.54,0.54,1.33,1.33,2.13,2.13,2.13)/1000
        }
        if(para==3)
        {
                prev_male<-c(6.2,9.5,14.0,19.1,23.7,27.1,28.5,27.8,23.2,23.2)/100
                prev_female<-c(5.1,7.4,10.7,14.2,17.8,21.3,23.1,23.4,19.9,19.9)/100

                inci_male<-c(8.2,11.6,15.5,19.3,22.6,23.4,22.7,20.9,16.2,16.2)/1000
                inci_female<-c(5.8,8.3,11.2,14.1,16.7,17.8,17.9,17.2,13.6,13.6)/1000
        }

### age-gender mortality rate ###

        death_male<-c(2.7,4.2,6.7,10.8,17.3,28.1,46,77.3,131.3,226.2)/1000
        death_female<-c(1.8,2.7,4.1,6.8,10.4,17.4,29.4,51.4,93.6,191.6)/1000

### subject ID ###
        pid<-c(1:30000)

### gender of subjects ###
        sex<-c(rep(0,3204), rep(1,3216), rep(0,2793), rep(1,2850), rep(0,2439),
rep(1,2493), rep(0,1824), rep(1,1890), rep(0,1407), rep(1,1518), rep(0,1197), rep(1,1359),
rep

(0,927), rep(1,1194), rep(0,594), rep(1,939), rep(0,51), rep(1,105))

### age and age-group of subjects at baseline ###
        age_group<-c(rep(1,3204), rep(1,3216), rep(2,2793), rep(2,2850), rep(3,2439),
rep(3,2493), rep(4,1824), rep(4,1890), rep(5,1407), rep(5,1518), rep(6,1197), rep(6,1359),

rep(7,927), rep(7,1194), rep(8,594), rep(8,939), rep(9,51), rep(9,105))

        age_base<-c(rep((45:49), 1284), rep((50:54), 1128), c(50,51,52), rep((55:59),986),
c(55,56), rep((60:64),742), c(60,61,62,63), rep((65:69), 585), rep((70:74), 511), c

(70), rep((75:79), 424), c(75), rep((80:84), 306), c(80,81,82), rep(85,156))

### whether subjects are diseased at baseline ###
        base_disease<-c(rbinom(3204,1,prev_male[1]), rbinom(3216,1,prev_female[1]),
rbinom(2793,1,prev_male[2]), rbinom(2850,1,prev_female[2]),
rbinom(2439,1,prev_male[3]),

rbinom(2493,1,prev_female[3]), rbinom(1824,1,prev_male[4]),
rbinom(1890,1,prev_female[4]), rbinom(1407,1,prev_male[5]),
rbinom(1518,1,prev_female[5]), rbinom(1197,1,prev_male
[6]), rbinom(1359,1,prev_female[6]), rbinom(927,1,prev_male[7]),
rbinom(1194,1,prev_female[7]), rbinom(594,1,prev_male[8]),
rbinom(939,1,prev_female[8]), rbinom(51,1,prev_male

[9]), rbinom(105,1,prev_female[9]))

### whether subjects are dead and time of death ###
    dead<-rep(0, 30000)
    time_dead<-rep(0, 30000)

### whether subjects are diseased and time when diseased during the follow-up period
###
    time_disease<-rep(0, 30000)
    disease<-base_disease


    age<-age_base


    for(year in 1:21)
    {
        age<-age+1
        age_group<-as.integer((age-45)/5+1)


        for(i in 1:30000)
        {
            if (age_group[i]>10)
            {
                age_group[i]=10
            }

            if (dead[i]==0)
            {
                if(sex[i]==0)
                {
                    dead[i]<-rbinom(1,1,death_male[age_group[i]])
                }
                else
                {
                    dead[i]<-rbinom(1,1,death_female[age_group[i]])
                }
                if (dead[i]==1)
                {
                    time_dead[i]<-year
                }
            }


            if(dead[i]!=1 && disease[i]!=1)
            {
                if(sex[i]==0)
                {
                    disease[i]<-rbinom(1,1,inci_male[age_group[i]])
                }
                else
                {
                    disease[i]<-rbinom(1,1,inci_female[age_group[i]])
                }

                if (disease[i]==1)
                {
                    time_disease[i]<-year
                }
            }

        }
    }

    for(i in 1:30000)
    {
        if(disease[i]==0 && dead[i]==1)
        {
            time_disease[i]=time_dead[i] ## if dead, time of diseased censored at time of dead ##
        }
        if(disease[i]==0 && dead[i]==0)
        {
            time_disease[i]=21 ## if not dead, time of diseased censored at the end of study ##
        }
        if(dead[i]==0)
        {
            time_dead[i]=21 ## if not dead, time of dead censored at the end of study ##
        }
    }



    data<-data.frame(cbind(pid,age_base,sex, base_disease, disease,time_disease,dead, time_dead))

    if(para==1)
    {
        write.table(data,"c:/dementia.txt", sep=" ", row.names=F)
    }
    else if(para==2)
    {
        write.table(data,"c:/parkinson.txt", sep=" ", row.names=F)
    }
    else if(para==3)
    {
        write.table(data,"c:/copd.txt", sep=" ", row.names=F)
    }

}

gen_disease_data(1)
gen_disease_data(2)
gen_disease_data(3)



### estimate Weibull shape and scale parameters for transition ###
### from Healthy to Diseased and Health to Dead ###

library(survival)

est_shape_scale=function(para)
{
    if(para==1)
    {
        data<-read.table("c:/dementia.txt", header=T, sep=" ")
    }
    else if(para==2)
    {
        data<-read.table("c:/parkinson.txt", header=T, sep=" ")
    }
    else if(para==3)
```

```
        {
                data<-read.table("c:/diabetes.txt", header=T, sep=" ")
        }

        healthy<-data[data$base_disease==0, ]

        time_healthy_diseased<-healthy$time_disease+healthy$age_base-45
        status_healthy_diseased<-healthy$disease
        reg_healthy_diseased<-
survreg(Surv(time_healthy_diseased,status_healthy_diseased)~1, dist="weibull")
        healthy_diseased_shape<-1/reg_healthy_diseased$scale
        healthy_diseased_scale<-exp(coef(reg_healthy_diseased))


        time_healthy_dead<-healthy$time_dead+healthy$age_base-45
        status_healthy_dead<-healthy$dead
        reg_healthy_dead<-survreg(Surv(time_healthy_dead,status_healthy_dead)~1,
dist="weibull")
        healthy_dead_shape<-1/reg_healthy_dead$scale
        healthy_dead_scale<-exp(coef(reg_healthy_dead))


        time_dead<-data$time_dead+data$age_base-45
        status_dead<-data$dead
        reg_dead<-survreg(Surv(time_dead,status_dead)~1, dist="weibull")
        dead_shape<-1/reg_dead$scale
        dead_scale<-exp(coef(reg_dead))

        return(data.frame(cbind(healthy_diseased_shape, healthy_diseased_scale,
healthy_dead_shape, healthy_dead_scale, dead_shape, dead_scale)))

}

est_shape_scale(1)
est_shape_scale(2)
est_shape_scale(3)
```

CHAPTER 3

Comparing the performance of different multiple imputation strategies for missing binary outcomes in cluster randomized trials: a simulation study

Jinhui Ma [1, 2, 3], Parminder Raina [1, 2], Joseph Beyene [1], Lehana Thabane [1, 3, 4, 5, *]

[1] Department of Clinical Epidemiology and Biostatistics, McMaster University, Hamilton, Ontario, Canada

[2] McMaster University Evidence-based Practice Center, Hamilton, Ontario, Canada

[3] Biostatistics Unit, St. Joseph's Healthcare Hamilton, Hamilton, Ontario, Canada

[4] Centre for Evaluation of Medicines, St Joseph’s Healthcare Hamilton, Ontario, Canada

[5] Population Health Research Institute, Hamilton Health Sciences, Hamilton, Ontario, Canada

Corresponding author:

Lehana Thabane

Biostatistics Unit/FSORC

3rd Floor Martha, Room H325

St. Joseph's Healthcare Hamilton

50 Charlton Avenue East, Hamilton, ON., L8N 4A6

Tel: 905.522.1155 x 33720

Fax: 905.308.7212

Email: thabanl@mcmaster.ca

Abstract

Introduction

Although researchers have proposed various strategies to handle missing outcomes in cluster randomized trials (CRTs), limited attention has been paid to the performance of these strategies. Under the assumption of covariate-dependent missingness, the objective of this simulation study is to compare the performance of various strategies in handling missing binary outcomes in CRTs under different design settings.

Methods

There are six missing data strategies investigated in this paper, which include complete case analysis, standard multiple imputation (MI) strategies using either logistic regression or Markov chain Monte Carlo (MCMC) method, within-cluster MI strategies using either logistic regression or MCMC method, and MI using logistic regression with cluster as a fixed effect. The performance of these strategies is evaluated through bias, empirical standard error, root mean squared error, and coverage probability.

Results

Under the assumption of covariate-dependent missingness and applying the generalized estimating equations approach for fitting the logistic regression, it was shown that complete case analysis yields valid inferences when the percentage of missing outcomes is not large (<20%) for all designs of CRTs considered in this paper. Standard MI strategies can be adopted when the design effect is small (variance inflation factor [VIF]≤3); however, they tend to underestimate the standard error of treatment effect when the design effect is large. Within-cluster MI strategy using logistic regression is valid for imputation of missing data from CRTs especially when the cluster size is large (>50) and the design effect is large (VIF>3). In contrast, within-cluster MI strategy using MCMC method may yield biased estimates of treatment effect for CRTs with small cluster size (≤50). MI using logistic regression with cluster as a fixed effect may substantially overestimate the standard error of the estimated treatment effect when the intracluster correlation coefficient is small. It may also lead to biased estimated treatment effect.

Conclusion

Findings from this simulation study provide researchers with quantitative evidence to guide selection of an appropriate strategy to deal with missing binary outcomes.

Keywords: cluster randomized trial, missing data, multiple imputation, design effect, variance inflation factor

1. INTRODUCTION

With the growing prominence of cluster randomized trials (CRTs) in health research, some attention has been paid to strategies for handling missing data in CRTs in the statistical community in recent years. Taljaard *et al* [1] evaluated imputation strategies for missing continuous outcomes in CRTs via simulation, assuming the data were missing completely at random (MCAR). They concluded that if the intracluster correlation coefficient (ICC) is small (<0.005), ignoring the clusters may yield acceptable Type I error; however, if the ICC is large, ignoring the clustering will lead to severe inflation of the Type I error. Andridge [2] investigated the impact of fixed-effects modeling of clusters in multiple imputation (MI) for CRTs with continuous outcomes assuming outcomes are MCAR or missing at random (MAR). She showed that incorporation of clustering using fixed effects for clusters can lead to severe overestimation of variance of group means, and the overestimation is more severe when cluster sizes and ICCs are small. A previous study [3] compared several strategies for handling missing binary outcomes in CRTs under the assumption of MCAR and covariate-dependent missingness (CDM) and found that within-cluster and across-cluster MI strategies, which take into account intracluster correlation, provide more conservative treatment effect estimates compared with MI strategies which ignore the clustering effect.

Though researchers have proposed various strategies, comprehensive guidelines on the selection of the most appropriate or optimal strategy for handling missing binary outcomes from CRTs are not available in the literature. The generalizability of the conclusions from a previous study [3] to other design settings may be limited since the simulation study was based on a real dataset which has a relatively large cluster size and ICC. Moreover, different imputation strategies were compared through the odds ratios (ORs) and corresponding 95% confidence intervals (CIs) for the estimated treatment effect, and the kappa statistics for agreement between imputed datasets and the real dataset. Other evaluation criteria, such as bias, root mean squared error (RMSE), and coverage probability, are considered more informative to assess the accuracy and efficiency of different imputation strategies.

This present paper extends earlier work [3] and evaluates the performance of various strategies for missing binary outcomes in CRTs under different design settings. Under the assumption of CDM, this present paper focuses on the following strategies: complete case analysis; two standard MI strategies, i.e. logistic regression and Markov chain Monte Carlo (MCMC) method; two within-cluster MI strategies, i.e. logistic regression and MCMC method; and MI strategy using logistic regression with cluster as a fixed effect. Using the generalized estimating equations (GEE) approach for fitting the population-averaged model for clustered binary data, the performance of these strategies was compared using bias, RMSE, coverage probability of nominal 95% CI, and empirical standard error of the estimated treatment effect. The ultimate aim of this project is to provide researchers with quantitative evidence to guide selection of an appropriate strategy to deal with missing binary outcomes based on the design settings of CRTs and percentage of missing data.

2. METHODS

Multiple imputation has been widely applied to missing data problems. Rubin [4] described MI as a three-step process: 1) replace each missing value with a set of plausible values that represent the uncertainty about the right value to impute; 2) analyze the multiple imputed datasets using complete-data methods; and 3) combine the results from the multiple analyses, which allows uncertainty regarding the imputation to be taken into account.

This paper investigates the performance of six strategies to handle missing binary outcomes from CRTs under the assumption of CDM, i.e. the probability of missing outcomes for CRTs depends only on the observed covariates. The six missing data strategies are complete case analysis, two standard MI strategies that ignore the clustering (logistic regression and MCMC method), two within-cluster MI strategies (logistic regression and MCMC method), and MI using logistic regression with cluster as a fixed effect. All programming and analyses are implemented in SAS 9.2 (Cary, NC) in the simulation. The *mi* procedure is used to implement the multiple imputation, the *genmod* procedure is used to obtain the intervention effect estimate and its standard error from the GEE approach, and the *mianalyze* procedure is used to obtain the pooled estimate and standard error across multiple imputed datasets.

This section is organized with an introduction of the strategies investigated in this paper, followed by an illustration of the statistical method used to analyze the binary outcomes from CRTs, and finally, description of how the results from multiple imputed datasets are combined to obtain pooled results.

2.1. Missing data strategies

2.1.1. Complete case analysis

A complete case analysis simply omits those for whom data are incomplete. This commonly used approach loses power and may introduce bias given that the incompleteness of data is not random.

2.1.2. Standard multiple imputation

2.1.2.1. Logistic regression method

The standard multiple imputation using logistic regression [5] is implemented through the following steps.

1) Fit a logistic regression using the observed outcome and covariates to obtain the posterior predictive distribution of the parameters:

$$\text{logit}(\Pr(y_{obs}=1)) = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k ,$$

where y_{obs} is the observed binary outcome of a subject, x_i, $i=1,\ldots,k$, denotes the i^{th} individual or cluster level covariates of the corresponding subject, $\beta = (\beta_0, \beta_1, \ldots, \beta_k)$ denotes the regression coefficients, and $\text{logit}(\Pr(y_{obs}=1)) = \log\left(\frac{\Pr(y_{obs}=1)}{1-\Pr(y_{obs}=1)}\right)$. In this project, we only include two covariates (treatment groups and another variable associated with the probability of missingness). The regression parameter estimates $\hat{\beta} = (\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_k)$ and the associated covariance matrix V are obtained to construct the posterior distribution of the parameters.

2) Draw new parameters $\tilde{\beta} = (\tilde{\beta}_{0,}\ \tilde{\beta}_1, \ldots, \tilde{\beta}_k)$ from the posterior distribution, where $\tilde{\beta} = \hat{\beta} + V_h^{'} Z$, $V_h^{'}$ is the upper triangular matrix in the Cholesky decomposition, $V = V_h^{'} V_h$, and Z is a vector of $k+1$ independent random normal variates.

3) For each subject with a missing outcome y_{mis} and observed covariates $x_1, \ldots, x_k$, compute $p = \dfrac{\exp(\tilde{\beta}_0 + \tilde{\beta}_1 x_1 \cdots + \tilde{\beta}_k x_k)}{1 + \exp(\tilde{\beta}_0 + \tilde{\beta}_1 x_1 \cdots + \tilde{\beta}_k x_k)}$ as the expected probability of $y_{mis} = 1$.

4) Draw a random uniform variate u, $0 \le u \le 1$. If $u < p$, then impute $y_{mis} = 1$, otherwise, impute $y_{mis} = 0$.

2.1.2.2. Markov chain Monte Carlo method

Assuming that the data are from a multivariate normal distribution, multiple imputation using MCMC method [6] constructs a Markov chain to simulate draws from the posterior distribution $\Pr(Y_{mis} \mid Y_{obs})$, where Y_{mis} and Y_{obs} denote the missing and observed values respectively. The missing data are imputed through repeating two steps: the imputation step and the posterior step. The i^{th} iteration of the steps can be defined as follows.

1) Imputation step: simulate the missing values for each observation independently given an estimated mean vector and covariance matrix denoted by *θ*, i.e. draw values for variables with missing data $Y_{mis}^{(t+1)}$ from a conditional

distribution $\Pr(Y_{mis} \mid Y_{obs}, \theta^{(t)})$ where Y_{mis} and Y_{obs} denote variables with missing and observed data, respectively.

2) Posterior step: simulate the posterior population mean vector and covariance matrix, which are then used in the imputation step, from the complete sample estimates, i.e. draw $\theta^{(t+1)}$ from $\Pr(\theta|Y_{obs}, Y_{mis}^{(t+1)})$.

The two steps are iterated long enough to generate a Markov chain $\{\theta^{(t)}, \mathrm{Y}_{mis}^{(t)} : t = 1, 2, \ldots\}$, which converges in distribution to the posterior distribution $\Pr(Y_{mis}, \theta|Y_{obs})$.

In this study, the observed data y_{obs} include the observed outcome, treatment exposure, and the values for another variable associated with the probability of missingness. We used a single chain and non-informative prior for the Bayesian simulations to derive posterior distributions. We then applied expectation-maximization (EM) algorithm to find maximum likelihood estimates to impute missing data. The iterations are considered to have converged when the change in the parameter estimates between iteration steps is less than 0.0001 for each parameter. Due to the assumption of multivariate normality, the imputed values from this method are continuous. They are rounded to 0 if less than 0.85, and to 1 otherwise, based on the prevalence of events in the simulated datasets.

2.1.3. Within-cluster multiple imputation

Within-cluster imputation refers to standard MI using either logistic regression or MCMC method being applied for each cluster independently, i.e. the missing values are imputed based on the observed data within the same cluster as the missing values;

therefore, the similarity of subjects from the same cluster is taken into account in within-cluster imputation methods.

Within-cluster MI strategies may not be applicable for CRTs with a small number of subjects within any cluster because the multiple imputation procedure (*mi* procedure) in SAS cannot handle the case when all subjects within a cluster are missing or when the non-missing binary outcomes within a cluster have identical observations (i.e. either all 0 or all 1), phenomena that happen very often for CRTs with a small number of subjects within a cluster. In this simulation study, only the situation when all the non-missing binary outcomes within a cluster are zero was encountered. In this case, the missing values in these clusters were replaced with zero to avoid imputing them. In addition, this strategy needs to be approached with caution, since there may be no missing data for some clusters, and in this case, the standard software programs cannot be directly applied to conduct within-cluster imputation. It will be necessary to separate clusters into two groups: clusters with missing data and clusters with complete data. For clusters with missing data, standard procedure can be used for imputing the missing data by clusters, at which point clusters with imputed data will need to be merged with the clusters with complete data for later analysis.

2.1.4. Multiple imputation using logistic regression with cluster as a fixed effect

This method is similar to the standard MI using logistic regression; however, as its name suggests, cluster is added as a fixed effect when fitting the logistic regression using observed data and the logistic regression for imputing missing data.

2.2. Statistical analysis method

As the statistical analysis model in this study, the GEE approach was used for fitting the logistic regression, which is a commonly used method for analyzing binary outcome in CRTs to estimate the marginal (or population-averaged) treatment effect. The GEE method, developed by Liang and Zeger [7], can be formulated as

$$\log\mathrm{it}(\Pr(y_{ijl}=1))=\beta_0+\beta_1 x_{ijl}^1+\cdots+\beta_k x_{ijl}^k,$$

where y_{ijl} denotes the binary outcome of patient *l* in cluster *j* in the intervention group *i*, x_{ijl}^k denotes the k^{th} individual-level or cluster level covariates of the corresponding subject, β_k denotes the regression coefficients, and $\mathrm{logit}\,(\Pr(y_{ijl}=1))=\log\left(\frac{\Pr(y_{ijl}=1)}{1-\Pr(y_{ijl}=1)}\right)$.

An exchangeable correlation matrix is specified to account for the potential within-cluster homogeneity in outcomes, and the robust standard error method is used to obtain the improved standard error for the estimated β coefficients. In this paper, only treatment exposure is included in the model fitting.

Another statistical analysis method, random-effects logistic regression, is also widely used to estimate the conditional (or cluster-specific) treatment effect. In this simulation study, the GEE method is adopted since the marginal treatment effect it tries to estimate is consistent with the effect used to generate the clustered binary data using beta-binomial distribution as described below. However, the GEE method underestimates the standard error of treatment effect when the number of clusters is small (<20). In this case, a small sample modification to GEE proposed by Mancl and DeRouen [8] was applied to

GEE, which corrects the downward bias of the sandwich standard error estimator by multiplying it by $\sqrt{J/(J-1)}$, where J is the number of clusters in each arm.

2.3. Combining the results from different imputed data sets

For multiple imputed CRT data, the estimate of treatment effect (logarithm of the odds ratio) and its variance are obtained in the same fashion as for the independent data. Suppose *M* sets of imputed values are generated, and the *M* estimates of the treatment effects are $\beta^{(1)}, \beta^{(2)}, \ldots, \text{and}\, \beta^{(M)}$ with corresponding variance estimates $V^{(1)}, V^{(2)}, \ldots, \text{and}\, V^{(M)}$, these estimates can be combined as described by Rubin [5]. The point estimate for the treatment effect estimate from MI is $\bar{\beta} = \frac{1}{M}\sum_{m=1}^{M}\beta^{(m)}$, its variance estimate is $V = W + \left(1 + \frac{1}{M}\right)B$, where $W = \frac{1}{M}\sum_{m=1}^{M}V^{(m)}$ is the average within-imputation variance, and $B = \frac{1}{M-1}\sum_{m=1}^{M}\left(\beta^{(m)} - \bar{\beta}\right)^2$ is the between imputation variance. The adjusted *t*-test under the MI is then given by $T = \frac{\bar{\beta} - \beta}{\sqrt{V}} \sim t_{v_M}$. The degree of freedom is calculated as $v_M = (M-1)\left(1 + \frac{M}{M+1}\frac{W}{B}\right)^2$. For CRTs, complete data degrees of freedom are small since they are based on the number of clusters, rather than the total number of subjects. In this study, the adjusted degree of freedom recommended by Barnard and Rubin [9, 10] was used, i.e. $v_{adj} = \left(\frac{1}{v_M} + \frac{V}{W}\frac{v_{com}+3}{v_{com}+1}\frac{1}{v_{com}}\right)^{-1}$, where v_{com} is the degree of freedom for the complete data test: if, for example, there are k (k>2) clusters in each of the two study groups, $v_{com} = 2(k-1)$.

3. SIMULATION STUDY

this section describes considerations for selection of design parameters for CRTs, the data generation process, and measures of performance for the missing data strategies.

3.1. Choices of design parameters for the simulation

For simplicity, only two-arm CRTs which are completely randomized, have an equal number of subjects within each cluster, and an equal number of clusters within each arm are considered. The number of clusters, the number of subjects per cluster, and the ICC are allowed to vary.

In accordance with the review of CRTs in primary care by Eldrige *et al* [11], the CRTs were categorized by sample size as either trials with a small number of clusters and a large number of subjects in each cluster or trials with a large number of clusters and a small number of subjects in each cluster. The empirical findings suggest that larger values of ICC tend to be associated with studies having a small number of participants within each cluster [12]. Guided by this information, the choices of combinations of design parameters used in this simulation study are as follows.

(1) CRTs with n=6 clusters per arm, m=500 subjects per cluster, and ICC is ρ=0.001, 0.01, and 0.05.

(2) CRTs with n=20 clusters per arm, m=50 subjects per cluster, and ICC is ρ=0.01, 0.05, and 0.1.

(3) CRTs with *n*=30 clusters per arm, *m*=30 subjects per cluster, and ICC is ρ= 0.05, 0.1, and 0.2.

(4) The choice of the percentage of missing binary outcome is 20%; the outcome prevalence for the intervention and control arms are 10% and 20% respectively.

In addition, five replacements are generated for each of the missing data, i.e. generate five datasets when applying the above MI strategies to achieve relative efficiency of more than 90% [5].

3.2. Data generation

According to Ridout *et al*, clustered binomial responses can be generated using a beta-binomial distribution [13]. The prevalence of outcome π ($0<\pi<1$) varies from cluster to cluster according to a beta distribution with parameters $\alpha>0$ and $\beta>0$. The binary outcomes for each cluster are generated from the binomial distribution conditional on this prevalence or probability. To generate data with an intracluster correlation coefficient $0<\rho<1$ and marginal prevalence of outcome π, the parameters of the beta distribution are chosen such that $\alpha = \pi \frac{1-\rho}{\rho}$ and $\beta = (1-\pi)\left(\frac{1-\rho}{\rho}\right)$.

Besides the variable of intervention group, another binary covariate is generated which is associated with the probability of missingness. It is assumed that this binary covariate has equal chance to take the value of 0 or 1, and is independent of the intervention and the outcome. For any percentage of missing data, it is considered that subjects with value of 1 for this binary covariate are 1.3 times more likely to have missing outcome than subjects with value of 0. This variable is incorporated into the imputation model as a covariate. Moreover, for each combination of the design parameters, 1200 replications are generated to achieve enough precision for estimating treatment effect (within 5% accuracy of the true effect for all designs of CRTs investigated in this paper with a 5% significance level).

3.3. Measures of performance

Quantities used to assess the performance of various missing data strategies include bias, RMSE, coverage probability, and standard error of the treatment effect. Details of these measures are presented below.

3.3.1. Bias

Bias is defined as the difference between the average value of estimated treatment effects over the simulation repetitions and the true parameter set for treatment effect when generating data.

3.3.2. Root mean squared error

The mean square error (MSE) is defined as the average squared difference between the estimated treatment effects $\hat{\beta}$ and true parameter β , which is set for treatment effect when generating the data. MSE is equal to the sum of the variance and the squared bias of the estimated treatment effect. RMSE is $\sqrt{E_{\beta}[(\hat{\beta}-\beta)^2]}$, which is the square root of the MSE. The RMSE is a useful measure of overall precision or accuracy.

3.3.3. Coverage

The actual coverage of nominal 95% CIs of the estimated treatment effect is the proportion of time that the nominal interval contains the true treatment effect across all simulation replications. Since the 95% CI aims to contain the true treatment effect with probability of 0.95, nominal coverage should be approximately equal to coverage probability if the missing data strategy works well.

3.3.4. Empirical Standard error of the treatment effect

Empirical standard error (ESE) of the treatment effect is calculated as the average of standard errors of estimated treatment effects across all simulation replications. It has been well established that analytical methods failing to account for the correlation between responses within cluster, i.e. the clustering effect, result in underestimation of standard error for the intervention effect. The appropriate imputation model should also reflect this data structure, as pointed out by Kenward *et al* [14]; therefore, the ESE is considered to be the primary criterion in this study.

4. RESULTS

Since subjects within the same cluster are more likely to be similar to each other than those from different clusters, an additional subject from the same cluster adds less new information than would a completely independent subject. The design effect or the variance inflation factor (VIF) is commonly used to measure the clustering effect due to lack of independence in the data from a CRT design. The main components of the design effect are the intracluster correlation coefficient ρ and the size of cluster m, and VIF=1+(m-1) ρ. For each design setting of CRT investigated in this simulation study, the empirical standard error of estimated treatment effect, bias, RMSE, and coverage probability for analyzing the complete data (no missing data) are considered as references and compared with those for each missing data strategy. These results are presented in Table 1-6 and discussed in detail below.

4.1. Complete case analysis

For all the design settings of CRTs, the ESEs are inflated slightly; the biases are close to zero; and the RMSEs and coverage probabilities are very similar to their references (see Table 1). This result is not surprising since the complete case analysis is analogous to analyzing a size-reduced dataset in which all variables are fully observed under the assumption of CDM. It can yield an unbiased estimate of the intervention effect, but with a larger empirical standard error for CRTs with a small design effect compared to those with large design effect; this is due to loss of efficiency.

These findings suggest that for covariate-dependent missingness, complete case analysis may be an acceptable strategy as long as the percentage of missing data is not large (≤20%). The advantage of complete case analysis lies in its simplicity, and it is the default method applied to handling missing data in the standard software.

4.2. Standard multiple imputation strategies

When standard MI using logistic regression is used to handle the missing data, the ESEs are substantially underestimated for CRTs with large design effect (VIF>3) and similar for CRTs with small design effect; the biases are close to zero; biases and RMSEs are similar to their references; and the coverage for CRTs with large design effects is smaller than their references (see Table 2). The performance of the standard MI using MCMC method is similar to that of the standard MI using logistic regression, except that the underestimation of the standard error for imputation using MCMC method is not substantial (see Table 3).

Two reasons can help to interpret why the biases for standard MI using logistic regression are close to zero: first, CDM is assumed in this simulation study; and second, both the imputation strategy and statistical analysis model (GEE approach for fitting logistic regression) estimate the population-averaged treatment effect, which is consistent with the treatment effect used to generate the clustering data.

For CRTs with very small design effect, the information contributed from a subject within a cluster is quite similar to that from a completely independent subject and therefore the standard MI using logistic regression, which accounts for the uncertainty of the missing data through both within- and between-imputation variances, may provide larger standard error compare with that when there are no missing data. However, for CRTs with large design effects (VIF>3), this strategy underestimates standard error of the intervention effect since it ignores the clustering of data.

These findings suggest that the standard MI strategies are acceptable when the VIF is small (<3); otherwise, they tend to underestimate the standard error of the treatment effect. In addition, an MI strategy using MCMC can be applied for arbitrary missing data pattern, whereas an MI strategy using logistic regression can only be applied for monotone missing pattern. MI strategy using the MCMC method presents a convergence problem and may produce biased results when the prevalence of cases (i.e. the prevalence of outcome) is close to 0 or to 1.

4.3. Within-cluster multiple imputation strategies

The ESEs for within-cluster MI using logistic regression are larger than their references and the increased amount is ignorable for CRTs with large design effect and large cluster size; the biases and the RMSEs are quite similar to their references; the coverage probabilities are slightly larger than their references (see Table 4). This strategy imputes each incident of missing data based on only the observed data within the same cluster, which leads to increasing within- and between-imputation variance and thus affects overall variance of estimated treatment effect when compared with imputation based on all observed data across the different clusters.

The ESEs for within-cluster MI using MCMC method are similar to their references; for CRTs with small cluster size, the biases are not ignorable, which lead to larger RMSEs and smaller coverage probabilities compared with their references (see Table 5).

These findings suggest that within-cluster MI strategies are appropriate imputation strategies for CRTs, especially for CRTs with large cluster size and large design effect (VIF>3).

4.4. Multiple imputation using logistic regression with cluster as a fixed effect

When the MI using logistic regression with cluster as a fixed effect is applied to impute the missing data, the ESEs are substantially overestimated for CRTs with small ICC ($\rho<0.1$), which lead to larger coverage probabilities compared with their references. The biases are large especially for CRTs with small cluster size (≤ 50), which lead to smaller coverage probabilities. The large biases and overestimated SEs lead to increased RMSEs compared with their references (see Table 6).

These findings suggest that application of this strategy may result in a biased estimate of treatment effect and may substantially overestimate the standard error of the estimated treatment effect when ICC is small (ρ <0.1) and the cluster size is small (≤50).

5. DISCUSSION

Missing data is a common issue in CRTs which may lead to spurious conclusions if handled inappropriately. This study used an extensive set of simulations to assess the performance of different strategies for handling missing binary outcomes in CRTs under different design settings. Results from the present study demonstrate that the design of CRTs, including factors such as the number of clusters in each intervention group, the number of subjects within each cluster, the ICC and the VIF, are important determinants for selecting an appropriate missing data strategy. Under the assumption of CDM and application of the GEE approach for statistical analysis, complete case analysis can be used to obtain valid inference when the percentage of missing binary outcomes is small (<20%). Standard MI using logistic regression or MCMC method can be used to impute the missing values when the design effect is small (VIF≤3); however, they tend to underestimate the standard error of the treatment effect when the design effect is large, though the underestimation of the standard MI using MCMC method is not substantial. Within-cluster MI using logistic regression may be an appropriate strategy to impute missing binary outcomes in CRTs, especially for CRTs with large cluster size and design effect. The performance of within-cluster MI using MCMC method is good for CRTs with large cluster size and design effect (VIF>3); however, may yield biased estimation of the treatment effect for CRTs with small cluster size. The MI using logistic regression with cluster as a fixed effect substantially overestimates the standard error of the treatment effect for CRTs with small ICC (<0.05) and may result in a biased estimated treatment effect for CRTs with small cluster size.

The finding for the MI using logistic regression with cluster as a fixed effect in this paper parallels previous results by Andridge [2] who demonstrated that incorporating clusters as fixed effects to handle missing continuous outcomes can lead to serious overestimation of variance of group means, and this overestimation is more severe for small cluster sizes and small ICCs. The findings for complete case analysis, standard MI strategies which ignore the clustering effect, and within-cluster MI strategies are similar to those from Taljaard *et al* [1]; although, they evaluated imputation strategies for missing continuous outcomes in CRTs assuming the missingness was completely at random and used Type I error rate and statistical power as the main evaluation criteria. This present study adopted the design effect VIF and the ICC, rather than the ICC alone, to interpret simulation results, since VIF is determined by both the number of subjects within each cluster and the magnitude of ICC, and is more appropriate for capturing the pattern of the performance for different missing data imputation strategies.

It should be emphasized that complete case analysis may not be an appropriate strategy in practice though it shows good performance in this simulation study. In fact, the good performance of complete case analysis is highly dependent on the CDM assumption. In realistic scenarios, it is more likely for a CRT to have mixed missing data mechanisms, i.e. combination of missing completely at random (a participant accidentally missed the medical appointment for assessing his health outcome), CDM (older participants are more likely to be lost to follow-up), or missing not at random (participants with poor health outcome are more likely to be lost to follow-up). A simulation study by Allison [15] showed that MI is robust to model violations while complete case analysis is not. King *et al* [16] further showed that multiple imputation works well even when the assumptions of MI are violated.

There are certain limitations to the current study. First, the performance of different missing data strategies is only assessed in the setting of a completely randomized study design. Other designs such as matched pairs design and stratified randomized design are also used for CRTs but were not considered in this study. Second, only balanced design of CRTs (with equal number of subjects per cluster, equal number of clusters in each arm) was considered. These design restrictions were made in order to understand the performance of the methods in simple scenarios. However, the findings are relevant to more general settings, such as the unbalanced design of CRTs. Third, the scenario of missing an entire cluster was not investigated. Even though the complete case analysis, standard MI strategies, and MI using logistic regression with cluster as fixed effect can manage conditions when an entire cluster is missing, their performance in this scenario needs further investigation. Finally, only the GEE approach is considered as the analysis method for the present study; however, in practice, random-effects logistic regression is also commonly adopted for analyzing binary outcomes in CRTs. Further study could include investigation of the performance of these missing data strategies when random-effects logistic regression is used as the analysis model.

The strengths of this study include comparison to a previous simulation study [3] which focused on the estimate of the treatment effect under one particular design setting while emphasis of the present study has been on the accuracy and effectiveness of different missing data strategies, and should provide more informative criteria to assess performance of different imputation strategies. As well, this simulation study is designed to cover a wide range of design settings for CRTs and applies an amount of missing data commonly encountered in epidemiological research. All the above strengths enhance the generalizability of these findings.

6. CONCLUSIONS

The current study is the most comprehensive to date to examine performance of different strategies for handling missing binary outcomes in CRTs. When the percentage of missing data is large and the design effect of the CRT varies, different strategies may lead to varying results, and therefore the appropriate strategy needs to be chosen carefully to obtain valid inferences and mitigate design issues. Findings from this simulation study provide researchers with quantitative evidence to guide selection of appropriate strategies for handling missing binary outcomes based on the design settings of CRTs and the percentage of missing data.

7. ACKNOWLEDGEMENT

This study was supported in part by funds from the Canadian Network and Centre for Trials Internationally (CANNeCTIN) program and the Drug Safety and Effectiveness Cross-disciplinary Training (DSECT) program in the form of studentship on training awards for the first author. No additional external funding received for this study. Dr. Lehana Thabane is a clinical trials mentor for the Canadian Institutes of Health Research (CIHR). Dr. Parminder Raina holds a Raymond and Margaret Labarge Chair in Research and Knowledge Application for Optimal Aging, and the Canada Research Chair in GeroScience.

References

1. Taljaard M, Donner A, Klar N. (2008) Imputation strategies for missing continuous outcomes in cluster randomized trials. *Biom J* 50(3): 329-345.

2. Andridge RR. (2011) Quantifying the impact of fixed effects modeling of clusters in multiple imputation for cluster randomized trials. *Biom J* 53(1): 57-74.

3. Ma J, Akhtar-Danesh N, Dolovich L, Thabane L, CHAT investigators. (2011) Imputation strategies for missing binary outcomes in cluster randomized trials. *BMC Med Res Methodol* 11: 18.

4. Rubin DB. (1996) Multiple imputation after 18+ years. *Journal of the American Statistical Association* (91): 473-489.

5. Rubin DB. (1987) *Multiple imputation for nonresponse in surveys*. New York: John Wiley & Sons.

6. Schafer, J.L. (1997) *Analysis of Incomplete Multivariate Data*. Chapman & Hall, London.

7. Liang K, Zeger S. (1986) Longitudinal data analysis using generalized linear models. *Biometrika* 73(1): 13-22.

8. Mancl LA, DeRouen TA. (2001) A covariance estimator for GEE with improved small-sample properties. *Biometrics* **57**:126–134.

9. Barnard J, Rubin DB. (1999) Small-sample degrees of freedom with multiple imputation. *Biometrika* (86): 949-955.

10. Little RJ, Rubin DB. (2002) Statistical analysis with missing data (2 edition.). New York: John Wiley & Sons.

11. Eldridge SM, Ashby D, Feder GS, Rudnicka AR, Ukoumunne OC. (2004) Lessons for cluster randomized trials in the twenty-first century: A systematic review of trials in primary care. *Clin Trials* 1(1): 80-90.

12. Donner A. (1982) An empirical study of cluster randomization. *Int J Epidemiol* 11(3): 283-286.

13. Ridout MS, Demetrio CG, Firth D. (1999) Estimating intraclass correlation for binary data. *Biometrics* 55(1): 137-148.

14. Kenward MG, Carpenter J. (2007) Multiple imputation: Current perspectives. *Stat Methods Med Res* 16(3): 199-218.

15. Allison, P. (2000). Multiple imputation for missing data: A cautionary tale. Sociological methods and research, 28(3), 301-309.

16. King, G., Honaker, J., Joseph, A. & Scheve, K. (2001). Analyzing incomplete political science data: An alternative algorithm for multiple imputation. *American Political Science Review*, 95(1), 49-69.

Table 1. Performance of complete case analysis

Design parameters of CRTs			Design effect (variance inflation factor)	Empirical Standard Error		Bias		RMSE[1]		Coverage	
Num. of clusters per arm (m)	Num. of subjects per cluster (n)	Intra-cluster correlation coefficient (ρ)		No missing data (Ref.)[3]	Complete case analysis	No missing data (Ref.)	Complete case analysis	No missing data (Ref.)	Complete case analysis	No missing data (Ref.)	Complete case analysis
5[2]	500	0.001	1.499	0.08/0.09	0.09/0.10	0.00	0.00	0.10	0.10	0.88/0.91	0.89/0.92
		0.01	5.99	0.17/0.18	0.17/0.19	-0.01	-0.01	0.18	0.18	0.92/0.94	0.91/0.94
		0.05	25.95	0.34/0.38	0.34/0.38	-0.03	-0.03	0.38	0.38	0.91/0.94	0.90/0.94
20	50	0.01	1.49	0.16	0.17	0.01	0.01	0.16	0.18	0.92	0.93
		0.05	3.45	0.24	0.25	-0.02	-0.02	0.24	0.24	0.95	0.96
		0.1	5.9	0.31	0.32	-0.04	-0.04	0.32	0.33	0.94	0.94
30	30	0.05	2.45	0.21	0.22	-0.01	-0.02	0.22	0.23	0.94	0.94
		0.1	3.9	0.27	0.28	-0.00	-0.01	0.27	0.28	0.94	0.94
		0.2	6.8	0.35	0.36	-0.03	-0.03	0.37	0.38	0.93	0.93

Note:
1. RMSE: root mean squared error.
2. For CRTs with 6 clusters per arm, standard errors/modified standard errors and coverage/modified coverage are provided.
3. Ref.: reference

Table 2. Performance of standard multiple imputation using logistic regression

Design parameters of CRTs			**Design effect (variance inflation factor)**	**Empirical Standard Error**		**Bias**		**RMSE[1]**		**Coverage**	
Num. of clusters per arm (m)	**Num. of subjects per cluster (n)**	**Intra-cluster correlation coefficient (ρ)**		**No missing data (Ref.)[4]**	**Standard MI using LR[2]**	**No missing data (Ref.)**	**Standard MI using LR[2]**	**No missing data (Ref.)**	**Standard MI using LR[2]**	**No missing data (Ref.)**	**Standard MI using LR[2]**
5[3]	500	0.001	1.499	0.08/0.09	0.09/0.10	0.00	0.00	0.10	0.10	0.88/0.91	0.89/0.93
		0.01	5.99	0.17/0.18	0.14/0.16	-0.01	-0.01	0.18	0.18	0.92/0.94	0.86/0.91
		0.05	25.95	0.34/0.38	0.28/0.31	-0.03	-0.03	0.38	0.38	0.91/0.94	0.83/0.87
20	50	0.01	1.49	0.16	0.16	0.01	0.01	0.16	0.18	0.92	0.92
		0.05	3.45	0.24	0.22	-0.02	-0.02	0.24	0.25	0.95	0.92
		0.1	5.9	0.31	0.27	-0.04	-0.04	0.32	0.33	0.94	0.89
30	30	0.05	2.45	0.21	0.20	-0.01	-0.02	0.22	0.23	0.94	0.91
		0.1	3.9	0.27	0.24	-0.00	-0.01	0.27	0.28	0.94	0.90
		0.2	6.8	0.35	0.31	-0.03	-0.03	0.37	0.38	0.93	0.89

Note:
1. RMSE: root mean squared error.
2. Standard MI using LR: Standard multiple imputation (MI) using logistic regression.
3. For CRTs with 6 clusters per arm, standard errors/modified standard errors and coverage/modified coverage are provided.
4. Ref.: reference

Table 3. Performance of standard multiple imputation using Markov chain Monte Carlo method

Design parameters of CRTs			Design effect (variance inflation factor)	Empirical Standard Error		Bias		RMSE[1]		Coverage	
Num. of clusters per arm (m)	Num. of subjects per cluster (n)	Intra-cluster correlation coefficient (ρ)		No missing data (Ref.) [4]	Standard MI using MCMC[1]	No missing data (Ref.)	Standard MI using MCMC[1]	No missing data (Ref.)	Standard MI using MCMC[1]	No missing data (Ref.)	Standard MI using MCMC[1]
5[3]	500	0.001	1.499	0.08/0.09	0.09/0.10	0.00	0.02	0.10	0.10	0.88/0.91	0.88/0.91
		0.01	5.99	0.17/0.18	0.16/0.18	-0.01	0.01	0.18	0.18	0.92/0.94	0.90/0.93
		0.05	25.95	0.34/0.38	0.32/0.36	-0.03	-0.01	0.38	0.37	0.91/0.94	0.89/0.93
20	50	0.01	1.49	0.16	0.16	0.01	0.03	0.16	0.18	0.92	0.93
		0.05	3.45	0.24	0.24	-0.02	0.00	0.24	0.24	0.95	0.95
		0.1	5.9	0.31	0.30	-0.04	-0.01	0.32	0.32	0.94	0.93
30	30	0.05	2.45	0.21	0.22	-0.01	0.00	0.22	0.22	0.94	0.94
		0.1	3.9	0.27	0.27	-0.00	0.02	0.27	0.28	0.94	0.93
		0.2	6.8	0.35	0.34	-0.03	-0.00	0.37	0.37	0.93	0.92

Note:
1. RMSE: root mean squared error.
2. Standard MI using MCMC: Standard multiple imputation (MI) using Markov chain Monte Carlo method.
3. For CRTs with 6 clusters per arm, standard errors/modified standard errors and coverage/modified coverage are provided.
4. Ref.: reference

Table 4. Performance of within-cluster multiple imputation using logistic regression

Design parameters of CRTs			Design effect (variance inflation factor)	Empirical Standard Error		Bias		RMSE[1]		Coverage	
Num. of clusters per arm (m)	Num. of subjects per cluster (n)	Intra-cluster correlation coefficient (ρ)		No missing data (Ref.) [4]	Within-cluster MI using LR[2]	No missing data (Ref.)	Within-cluster MI using LR[2]	No missing data (Ref.)	Within-cluster MI using LR[2]	No missing data (Ref.)	Within-cluster MI using LR[2]
5[3]	500	0.001	1.499	0.08/0.09	0.10/0.12	0.00	0.00	0.10	0.11	0.88/0.91	0.94/0.96
		0.01	5.99	0.17/0.18	0.18/0.20	-0.01	-0.01	0.18	0.18	0.92/0.94	0.92/0.95
		0.05	25.95	0.34/0.38	0.35/0.39	-0.03	-0.03	0.38	0.38	0.91/0.94	0.90/0.94
20	50	0.01	1.49	0.16	0.19	0.01	0.05	0.16	0.18	0.92	0.96
		0.05	3.45	0.24	0.26	-0.02	0.02	0.24	0.24	0.95	0.97
		0.1	5.9	0.31	0.33	-0.04	-0.01	0.32	0.32	0.94	0.95
30	30	0.05	2.45	0.21	0.24	-0.01	0.02	0.22	0.22	0.94	0.97
		0.1	3.9	0.27	0.29	-0.00	0.02	0.27	0.27	0.94	0.96
		0.2	6.8	0.35	0.37	-0.03	-0.01	0.37	0.37	0.93	0.94

Note: 1. RMSE: root mean squared error.
2. Within-cluster MI using LR: Within-cluster multiple imputation (MI) using logistic regression.
3. For CRTs with 6 clusters per arm, standard errors/modified standard errors and coverage/modified coverage are provided.
4. Ref.: reference

Table 5. Performance of within-cluster multiple imputation using Markov chain Monte Carlo method

Design parameters of CRTs			**Design effect (variance inflation factor)**	**Empirical Standard Error**		**Bias**		**RMSE**[1]		**Coverage**	
Num. of clusters per arm (m)	**Num. of subjects per cluster (n)**	**Intra-cluster correlation coefficient (ρ)**		**No missing data (Ref.)** [4]	**Within-cluster MI using MCMC**[2]	**No missing data (Ref.)**	**Within-cluster MI using MCMC**[2]	**No missing data (Ref.)**	**Within-cluster MI using MCMC**[2]	**No missing data (Ref.)**	**Within-cluster MI using MCMC**[2]
5[3]	500	0.001	1.499	0.08/0.09	0.10/0.11	0.00	-0.03	0.10	0.11	0.88/0.91	0.88/0.92
		0.01	5.99	0.17/0.18	0.18/0.20	-0.01	-0.04	0.18	0.20	0.92/0.94	0.91/0.94
		0.05	25.95	0.34/0.38	0.35/0.39	-0.03	-0.05	0.38	0.40	0.91/0.94	0.90/0.93
20	50	0.01	1.49	0.16	0.18	0.01	0.02	0.16	0.18	0.92	0.94
		0.05	3.45	0.24	0.24	-0.02	0.03	0.24	0.25	0.95	0.93
		0.1	5.9	0.31	0.30	-0.04	0.16	0.32	0.35	0.94	0.90
30	30	0.05	2.45	0.21	0.21	-0.01	0.17	0.22	0.27	0.94	0.89
		0.1	3.9	0.27	0.26	-0.00	0.27	0.27	0.37	0.94	0.82
		0.2	6.8	0.35	0.33	-0.03	0.37	0.37	0.51	0.93	0.77

Note:
1. RMSE: root mean squared error.
2. Within-cluster MI using MCMC: Within-cluster multiple imputation (MI) using Markov chain Monte Carlo method.
3. For CRTs with 6 clusters per arm, standard errors/modified standard errors and coverage/modified coverage are provided.
4. Ref.: reference

Table 6. Performance of multiple imputation using logistic regression with cluster as a fixed effect

Design parameters of CRTs			Design effect (variance inflation factor)	Empirical Standard Error		Bias		RMSE[1]		Coverage	
Num. of clusters per arm (m)	Num. of subjects per cluster (n)	Intra-cluster correlation coefficient (ρ)		No missing data (Ref.)[4]	MI using LR with cluster as a fixed effect[2]	No missing data (Ref.)	MI using LR with cluster as a fixed effect[2]	No missing data (Ref.)	MI using LR with cluster as a fixed effect[2]	No missing data (Ref.)	MI using LR with cluster as a fixed effect[2]
5[3]	500	0.001	1.499	0.08/0.09	0.33/0.37	0.00	0.06	0.10	0.15	0.88/0.91	1.00/1.00
		0.01	5.99	0.17/0.18	0.36/0.40	-0.01	0.05	0.18	0.21	0.92/0.94	0.99/1.00
		0.05	25.95	0.34/0.38	0.45/0.51	-0.03	0.03	0.38	0.38	0.91/0.94	0.97/0.98
20	50	0.01	1.49	0.16	0.21	0.01	0.05	0.16	0.18	0.92	0.96
		0.05	3.45	0.24	0.27	-0.02	0.08	0.24	0.23	0.95	0.97
		0.1	5.9	0.31	0.32	-0.04	0.12	0.32	0.30	0.94	0.95
30	30	0.05	2.45	0.21	0.24	-0.01	0.15	0.22	0.24	0.94	0.94
		0.1	3.9	0.27	0.28	-0.00	0.21	0.27	0.31	0.94	0.92
		0.2	6.8	0.35	0.33	-0.03	0.25	0.37	0.38	0.93	0.91

Note: 1. RMSE: root mean squared error.
2. MI using LR with cluster as a fixed effect: Multiple imputation (MI) using logistic regression with cluster as a fixed effect.
3. For CRTs with 6 clusters per arm, standard errors/modified standard errors and coverage/modified coverage are provided.
4. Ref.: reference

CHAPTER 4

Comparison of population-averaged and cluster-specific models for the analysis of cluster randomized trials with missing binary outcomes: a simulation study

Jinhui Ma [1, 2, 3], Parminder Raina [1, 2], Joseph Beyene [1], Lehana Thabane [1, 3, 4, 5, *]

[1] Department of Clinical Epidemiology and Biostatistics, McMaster University, Hamilton, Ontario, Canada

[2] McMaster University Evidence-based Practice Center, Hamilton, Ontario, Canada

[3] Biostatistics Unit, St. Joseph's Healthcare Hamilton, Hamilton, Ontario, Canada

[4] Centre for Evaluation of Medicines, St Joseph’s Healthcare Hamilton, Ontario, Canada

[5] Population Health Research Institute, Hamilton Health Sciences, Hamilton, Ontario, Canada

Corresponding author:

Lehana Thabane

Biostatistics Unit/FSORC

3rd Floor Martha, Room H325

St. Joseph's Healthcare Hamilton

50 Charlton Avenue East, Hamilton, ON., L8N 4A6

Tel: 905.522.1155 x 33720

Fax: 905.308.7212

Email: thabanl@mcmaster.ca

Abstract

Objective: To compare the accuracy and efficiency of population-averaged (i.e. generalized estimating equations (GEE)) and cluster-specific (i.e. random-effects logistic regression (RELR)) models for analyzing data from cluster randomized trials (CRTs) with missing binary responses.

Methods: Clustered responses were generated from a Beta-binomial distribution. Under the assumption of covariate dependent missingness, missing outcomes were handled by complete case, standard multiple imputation (MI) and within-cluster MI strategies. Data were analyzed using GEE and RELR. Performance of the methods was assessed using standardized bias (SB), empirical standard error (ESE), root mean squared error (RMSE), and coverage probability.

Results:

GEE performs well on all four measures — provided the downward bias of the standard error (when the number of clusters per arm is small) is adjusted appropriately — under the following scenarios: complete case analysis for CRTs with small amount of missing data; standard MI for CRTs with variance inflation factor (VIF) <3; within-cluster MI for CRTs with VIF≥3 and cluster size>50. RELR performs well only when small amount of data were missing and complete case analysis was applied.

Conclusion: GEE performs well as long as appropriate missing data strategies are adopted based on the design of CRTs and the percentage of missing data. In contrast, RELR does not perform well when either standard or within-cluster MI strategy is applied prior to the analysis.

Keywords: marginal model; population-averaged model; cluster-specific model; multiple imputation; cluster randomized trial; covariate dependent missingness; generalized estimating equations; random-effects logistic regression

1. INTRODUCTION

Cluster randomized trials (CRTs) are randomized controlled trials in which clusters of subjects rather than independent subjects are randomly allocated to trial arms and outcomes are measured for individual subjects or clusters. CRTs increasingly are being used in health services research and primary care. Reasons for adopting cluster randomization as a more appropriate design include: 1) administrative convenience; 2) ethical considerations; 3) intervention is naturally applied at the cluster level; 4) to enhance the subject compliance; and 5) to minimize the potential treatment "contamination" between the intervention and control subjects [1]. In CRTs, outcomes from subjects within the same cluster may exhibit a greater correlation than do outcomes from subjects in different clusters. The correlation within clusters, which is quantified by the intracluster correlation coefficient (ICC) ρ, may result in substantially reduced statistical efficiency relative to trials that randomize the same number of individuals. The overall outcome variance σ^2 in a CRT can be expressed as the sum of between-cluster variance σ_B^2 and within-cluster variance σ_W^2. Correspondingly, the ICC is defined as $\rho = \sigma_B^2 / (\sigma_B^2 + \sigma_W^2)$, which is interpreted as the amount of variation that can be explained by variation between clusters. The reduction in efficiency is a function of the variance inflation due to clustering, also known as the design effect or variance inflation factor (VIF), given by $\text{VIF} = 1 + (\overline{m} - 1)\rho$, where $\overline{m}$ denotes the average cluster size.

A key property of CRTs is that inferences or analyses are frequently done to apply at the individual level while randomization is at the cluster level, thus the unit of randomization may be different from the unit of inference or analysis. In this case, the lack of independence among individuals in the same cluster, i.e. the between-cluster variation, presents special methodological challenges that affect both design and analysis of CRTs. Consequently, standard approaches for statistical analysis do not apply because they may result in severely underpowered studies and spuriously elevated Type I error rates [1]. Different statistical methods that account for the clustering effect have been proposed in the literature and they are categorized into individual-level and cluster-level data analysis methods. Individual-level analysis models, such as population-averaged (PA) models (also called marginal models) and cluster-specific (CS) models (also called conditional models), have been advocated for the analysis of CRTs with binary outcomes since they allow for the possible imbalance of both cluster-level and individual-level characteristics to be incorporated into the analysis. The ability to adjust for imbalanced characteristics between trial arms is very important when the number of clusters is not large enough to keep the cluster- or individual-level characteristics balanced between the trial arms. The generalized estimating equations (GEE) approach [2] and the random-effects logistic regression (RELR) are two commonly used individual-level analysis methods for estimating the PA and CS intervention effect for CRTs with binary outcomes, respectively.

Some attention has been paid in the literature to the performance of GEE approach and RELR in the analysis of binary outcomes in CRTs. Austin [3] compared their statistical powers through a simulation study in which the minimum number of clusters examined was 26 (13 clusters per trial arm). The results showed that the differences between the two methods were negligible in most settings. Bellamy *et al*. also conducted a series of simulation studies comparing their statistical power [4]. They examined settings in which the total number of clusters was 10, 20, 30 or 50, the mean number of subjects per cluster was either 10 or 100, the ICC was 0.1, the response proportion in the control arm was 0.23 and the response proportions in the intervention arm were: 0.09, 0.13, 0.18, 0.23 or 0.28. The study showed that the difference between the two models diminished as the number of clusters increased. In particular, the difference was negligible if the total number of clusters was at least 30. However, if the total number of clusters was 10 or 20, RELR had moderately lower power than GEE method. Ukoumunne *et al*. [5] compared the accuracy of estimated treatment effect and confidence interval coverage of several methods for analyzing binary outcomes in CRTs through a simulation study. They showed that the GEE method had generally acceptable properties as long as the bias of the standard error was corrected when the number of clusters was small. The RELR was not assessed in their simulation study.

The risk of attrition may be high in some CRTs due to lack of direct contact with individual subjects and lengthy follow-up [6]. The impact of missing data on estimating treatment effect and its confidence interval depends on the mechanism which caused the data to be missing, the strategy to handle missing data, and the statistical model used for analysis. The objective of the present paper is to compare the accuracy and efficiency of PA and CS models through a simulation study, in particular, the GEE method and the RELR respectively, for analyzing binary outcomes in CRTs with missing data. The performance of the methods is compared in terms of standardized bias (SB), empirical standard error (ESE), root mean squared error (RMSE), and coverage probability. The simulation is designed under the assumptions of covariate dependent missingness (CDM) and CRTs with a balanced completely randomized design.

2. METHODS

The rest of this section is organized as follows: First, the statistical analysis methods (i.e. GEE and RELR) used to analyze binary outcomes in CRTs are described. Second, the missing data strategies used in this study for handling missing binary outcomes are briefly introduced. Third, the method for combining the results across multiply imputed datasets is described.

2.1. Statistical analysis methods

2.1.1. Generalized estimating equations

The GEE approach for fitting the logistic regression developed by Liang and Zeger [7] can be formulated as

$$\text{logit}\,(\Pr(y_{ijl} = 1) = X_{ijl}\beta_{\text{marginal}},$$

where y_{ijl} denotes the binary outcome of patient *l* in cluster *j* in the intervention group *i*, $\Pr(y_{ijl} = 1)$ denotes the corresponding probability of success, X_{ijl} denotes the corresponding vector of individual-level or cluster level covariates. $\beta_{m\arg inal}$ denotes the marginal regression coefficients, and $\text{logit}\,(\Pr(y_{ijl} = 1)) = \log\left(\dfrac{\Pr(y_{ijl} = 1)}{1 - \Pr(y_{ijl} = 1)}\right)$.

To analyze the data from CRTs, an exchangeable correlation matrix is usually specified to account for potential within-cluster homogeneity in outcomes, and the robust standard error method is used to obtain the improved standard error for estimation of β_{marginal}. In this paper, we only include one covariate and treatment group in the model fitting.

It has been recommended that at least 40 clusters need to be included in a study in order for the GEE method to produce reliable standard errors [8]. This is because, firstly, the method tends to underestimate the covariance of observations leading to downward biased estimate of standard error and, secondly, the estimate of standard error is highly variable when the number of clusters is too small [9]. A number of methods have been proposed for dealing with the shortcomings of the robust standard error estimator [8]. In this paper, the downward bias of the sandwich standard error estimator is adjusted by multiplying it by $\sqrt{J/(J-1)}$, where J is the number of clusters in each arm.

2.1.2. Random-effects logistic regression

RELR incorporates cluster-specific random effects into the logistic regression and assumes that the random effects follow a normal distribution. The model can be formulated as $\text{logit}\,(\Pr(y_{ijl}=1)=X_{ijl}\beta_{\text{conditional}}+U_{ij}$, where $U_{ij}\sim N(0,\ \sigma_B^2)$ represent the random effects, which vary independently from one cluster to another according to a common Normal distribution with mean of zero and variance of σ_B^2, which represents the between-cluster variance. $\beta_{\text{conditional}}$ denotes the conditional regression coefficients. Model parameters can be estimated using maximum likelihood [10].

Both GEE and RELR are commonly used statistical analysis methods for analyzing binary outcomes in CRTs [1]; however, the two methods do not estimate the same parameter. As described above, the GEE method allows one to estimate the marginal or PA intervention effect, whereas RELR allows one to estimate the conditional or CS intervention effect [4, 11, 12]. Neuhaus has suggested that marginal models are preferable for testing the effects of cluster-level covariates [12]. In cluster randomization trials, the intervention is a cluster-level exposure variable and, thus, GEE approach may be preferable to RELR. Nevertheless, RELR may remain relevant for the analysis of CRTs since Neuhaus has demonstrated that for a binary outcome, marginal treatment effect tends to be smaller than conditional treatment effect: $\beta_{m\arg inal} = \beta_{conditinal}(1-\rho)$, where ρ is the intracluster correlation coefficient. In addition, different assumptions are required for the two models regarding missing data. The marginal model using GEE method requires data to be missing completely at random (MCAR), whereas the cluster-specific model using RELR requires data to be missing at random (MAR). MCAR means the probability of an observation being missing does not depend on either observed or unobserved data. MAR means the probability of an observation being missing depends only on observed data (covariates or previous outcomes) [13]. In this paper, we assume a less stringent case of MCAR referred to by Little as CDM, i.e. the probability of missing binary outcomes in CRTs depends only on observed covariates. Under the assumption of CDM, both GEE and RELR are valid if the known covariates associated with missingness are adjusted for.

2.2. Missing data strategies

In this paper, we consider three strategies to handle missing binary outcomes in CRTs: 1) complete case analysis, 2) standard MI using logistic regression, and 3) within-cluster MI using logistic regression. The performance of GEE method and RELR is compared after missing data are handled by the above strategies.

Complete case analysis has been an attractive method to handle the missing data due to its simplicity. In adopting this strategy, only subjects with complete data are included for analysis, while subjects with missing data are excluded.

MI is widely applied to missing data problems. Rubin [14] described MI as a three-step process: 1) replace each missing value with a set of plausible values that represent the uncertainty about the right value to impute; 2) analyze the multiple imputed datasets independently using complete-data methods; and 3) combine the results from the multiple analyses, which allows the uncertainty regarding the imputation to be taken into account.

The standard MI using logistic regression method is now described in detail. The Within-cluster MI strategy is consists of applying the standard MI method to impute missing data for each cluster independently.

Standard multiple imputation using logistic regression is implemented through the following steps:

First, fit a logistic regression using the observed outcome and covariates to obtain the posterior predictive distribution of the parameters:

$$\operatorname{logit}(\Pr(y_{obs}=1)) = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k \,,$$

where y_{obs} is the observed binary outcome of a subject, x_i, $i = 1,\ldots,k$, denotes the i^{th} individual or cluster-level covariate of the corresponding subject (two covariates are included in this study: treatment group and the variable associated with the missingness), $\beta = (\beta_0, \beta_1, \ldots, \beta_k)$ denotes the regression coefficients. The regression parameter estimates $\hat{\beta} = (\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_k)$ and the associated covariance matrix V are obtained to construct the posterior distribution of the parameters.

Second, draw new parameters $\tilde{\beta} = (\tilde{\beta}_{0,}\ \tilde{\beta}_1, \ldots, \tilde{\beta}_k)$ from the posterior distribution, where $\tilde{\beta} = \hat{\beta} + V_h^{'} Z$, $V_h^{'}$ is the upper triangular matrix in the Cholesky decomposition, $V = V_h^{'} V_h$, and Z is a vector of $k+1$ independent random Normal variates.

Third, for each subject with a missing outcome y_{mis} and observed covariates $x_1, \ldots, x_k$, compute $p = \dfrac{\exp(\tilde{\beta}_0 + \tilde{\beta}_1 x_1 \cdots + \tilde{\beta}_k x_k)}{1 + \exp(\tilde{\beta}_0 + \tilde{\beta}_1 x_1 \cdots + \tilde{\beta}_k x_k)}$ as the expected probability of $y_{mis} = 1$.

Fourth, draw a random Uniform variate u, $0 \le u \le 1$. If $u < p$, then impute $y_{mis} = 1$, otherwise, impute $y_{mis} = 0$.

The above steps imply two assumptions: first, subjects are independent, which essentially ignores the similarity of subjects from the same cluster and, second, the missing data are imputed based on the PA treatment effect.

2.3. Combination of results from different imputed data sets

Suppose *M* sets of imputed values are generated. *M* estimates of the treatment effects $\beta^{(1)}, \beta^{(2)}, \ldots, \text{and}\, \beta^{(M)}$ with corresponding variance estimates $V^{(1)}, V^{(2)}, \ldots, \text{and}\, V^{(M)}$ are obtained after GEE or RELR are applied to the multiple imputed datasets. The pooled treatment effect estimate from MI is calculated as $\bar{\beta} = \frac{1}{M}\sum_{m=1}^{M}\beta^{(m)}$. Its variance estimate is calculated as $V = W + \left(1 + \frac{1}{M}\right)B$, where $W = \frac{1}{M}\sum_{m=1}^{M}V^{(m)}$ is the average within-imputation variance, and $B = \frac{1}{M-1}\sum_{m=1}^{M}\left(\beta^{(m)} - \bar{\beta}\right)^2$ is the between-imputation variance. As recommended by Barnard and Rubin [13, 15], the adjusted degree of freedom is calculated for CRTs as $v_{adj} = \left(\frac{1}{v_M} + \frac{V}{W}\frac{v_{com}+3}{v_{com}+1}\frac{1}{v_{com}}\right)^{-1}$, where $v_M = (M-1)\left(1 + \frac{M}{M+1}\frac{W}{B}\right)^2$ is the degree of freedom when subjects are assumed to be independent, and v_{com} is the degree of freedom for the complete data test; for example, if there are *k* (*k*>2) clusters in each of the two study groups, $v_{com} = 2(k-1)$.

3. SIMULATION STUDY

The schematic overview of the simulation study is illustrated in Figure 1. This simulation study is implemented in SAS 9.2 (Cary, NC). The *mi* procedure is used to implement the MI, *genmod* and *nlmixed* procedures are used to estimate the intervention effect and its standard error from GEE approach and RELR respectively, and the *mianalyze* procedure is used to obtain the pooled estimate and its standard error across multiple imputed datasets.

According to the review of CRTs in primary care by Eldrige *et al* [16], CRTs can be categorized into two types: S-design and L-design, which refer to design settings of CRTs with a small and large number of clusters per arm, respectively. Design parameters for CRTs in this simulation study are guided by the empirical findings that larger values of intracluster correlation coefficient tend to be associated with studies having a small number of participants within each cluster [17]. The choices of these parameters are:

(1) For CRTs with 5 clusters per arm (S-design) and 500 subjects per cluster, ICC was set to be 0.001, 0.01 or 0.05.

(2) For CRTs with 20 clusters per arm (L-design) and 50 subjects per cluster, ICC was set to be 0.01, 0.05, or 0.1.

(3) For CRTs with 30 clusters per arm (L-design) and 30 subjects per cluster, ICC was set to be 0.05, 0.1, or 0.2.

Only two-arm, balanced, and completely randomized CRTs are considered in this study. The clustered binomial responses are generated using a beta-binomial distribution [18]. The prevalence of outcome for intervention and control arms is assumed to be 30% and 40% respectively. In addition, another binary covariate is generated, which has an equal chance of taking the value of 0 or 1 and is independent of the intervention and the outcome. For any percentage of missing data, we consider that subjects with value of 1 for this binary covariate are 1.3 times more likely to have missing outcome than subjects with a value of 0 for this covariate. For each combination of design parameters, we generate 1000 replications to achieve enough precision for estimating treatment effect

[19]. Choices of the percentage of missing binary outcome are 0% (complete data), 15%, and 30%. We generate 5 replacements for each of the missing data.

Four quantities are chosen to evaluate the performance of GEE method and RELR: 1) standardized bias (SB) calculated as $\left|\frac{\text{Average of estimates - parameter}}{\text{standard deviation of estimates}}\right|$; 2) root mean squared error (RMSE) defined as $\sqrt{E_{\beta}[(\hat{\beta}-\beta)^2]}$, where $\hat{\beta}$ and β are the estimated treatment effect and its true value respectively; 3) coverage probability, which is the proportion of times that the nominal 95% confidence interval contains the true treatment effect across all simulation replications; and 4) empirical standard error (ESE) of the treatment effect calculated as the average of standard errors of the estimated treatment effects across all simulation replications.

4. RESULTS

4.1. Empirical standard error

The ESEs from GEE method and RELR for different design scenarios are presented in Table 1 and Figure 2. When complete case analysis was used to handle missing data, ESEs from GEE and RELR for all designs of CRTs increased with the increasing percentage of missing data. The magnitude of increase for the GEE method depended on the VIF of CRTs: the larger the VIF, the smaller amount of increase. In contrast, the magnitude of increase for the RELR depended on the cluster size: the smaller the cluster size, the larger amount of increase.

When standard MI was used to impute missing data, ESEs from the GEE method were acceptable for CRTs with VIF<3 in terms of yielding similar or slightly larger ESEs compared to those obtained for analyzing complete data but underestimated for CRTs with VIF≥3. This is because standard MI strategy assumes data are independent and cluster effect may be safely ignored for CRTs with VIF<3 when imputing missing data. In contrast, ESEs from RELR were not similar as those obtained from analyzing complete data. This is because that the imputed datasets were obtained based on the estimated PA treatment effect and corresponding underestimated standard error, which led to a difference between the standard error estimated from RELR based on the imputed datasets and that based on the complete data.

Within-cluster MI was not applicable for L-design of CRTs, which usually had small cluster size, since all outcomes in a cluster were missing or all observed outcomes had identical values, which caused the imputation procedure to fail. In the cases when within-cluster MI was applicable and used to impute the missing data, ESEs from GEE method were acceptable for CRTs with VIF≥3; however, for CRTs with VIF<3, ESEs were inflated. This is because when within-cluster MI were used to impute the missing data, the clustering effects were accounted for by imputing missing data based on the observed information within the same cluster as the missing data, therefore, the ESEs for GEE were acceptable for CRTs with VIF≥3. The ESEs from RELR were acceptable only when the cluster size is large (>50) and the ICC is small (≤0.01).

4.2. Standardized bias

The SBs from GEE method and RELR for different design scenarios are presented in Table 2 and Figure 3. SBs from GEE method were close to zero for any design settings and percentage of missing data, no matter which missing data strategy was used. In contrast, SBs for RELR were relatively larger. When complete case analysis was used to handle missing data, SBs for RELR did not change substantially with increasing percentage of missing data for S-design with larger design effect (VIF>3) and L-design with larger ICC (ICC≥0.1); however, SBs changed largely with an increasing percentage of missing data for other scenarios. When missing data were imputed by standard MI or within-cluster MI prior to statistical analysis, SBs for RELR were much smaller than those obtained by analyzing complete data (i.e. 0% missing data) using the same statistical method.

The magnitude of SB is dependent on the original data structure, i.e. how the data were generated, how the missing data were handled, and which statistical model is used for analysis. As described in the previous section, the clustered binary data were generated using Beta-binomial distribution, which assumes a PA treatment effect. Since complete case analysis does not change the original data structure under the assumption of CDM, the PA and CS treatment effects estimated from the GEE and RELR are quite consistent with those estimated based on complete data (i.e. datasets without missing values). The relationship between the PA and the CS treatment effects estimated from GEE method and RELR respectively still held; however, when either standard MI or within-cluster MI was used, the imputed values were obtained based on the estimated PA treatment effect and corresponding underestimated standard error, which largely distorted the CS treatment effects estimated from RELR compared with those estimated based on complete data.

4.3. Root mean squared error

The RMSE incorporates both the variance of the estimator and its bias, and measures the overall accuracy of the point estimator. RMSEs from GEE method and RELR for different design scenarios are presented in Table 3 and Figure 4. When complete case analysis was used to handle missing data, RMSEs from GEE method were very similar to those obtained based on complete data for all designs of CRTs with no larger than 15% missing data. With 30% missing values, the RMSEs from GEE were larger than those obtained based on analyzing complete data for the design of CRTs with small design effect (VIF<3). Similarly, RMSEs from RELR were very similar to those obtained based on complete data for all designs of CRTs with no larger than 15% missing data; however, with 30% missing values, RMSEs from RELR were much larger than those obtained based on complete data for the design of CRTs with small design effect (VIF<3) and small cluster size (<50).

When standard MI was used to impute missing data, RMSEs from GEE method increased with the percentage of missing data. With no larger than 15% missing data, the increase of RMSEs from GEE compared to those obtained based on complete data was not substantial. When the amount of missing values increased to 30%, RMSEs from the GEE method increased substantially for CRTs with small design effects (VIF<3). In contrast, RMSEs from RELR method were much smaller than those obtained from analyzing complete data for most of the design scenarios. We should note that the small RMSE for RELR here was not an indication of more accurate or precise estimate for the treatment effect, but rather a result of biased CS treatment effects and the corresponding underestimated standard error.

When within-cluster MI was used to impute missing data, the same pattern for RMSEs from both GEE and RELR was observed as when standard MI was used to impute missing data.

4.4. Coverage probability

Table 4 and Figure 5 show the coverage probabilities fromGEE method and RELR for different designs of CRTs. When complete case analysis was used to handle missing data, the coverage probabilities from GEE method were at least 0.90 for all the scenarios considered in this paper. The coverage probabilities from RELR were at least 0.95 for design of CRTs with small design effect (VIF<3) but were very low for CRTs with large design effect (VIF≥3).

When standard MI was used to impute missing data, coverage probabilities from GEE method increased for CRTs with small design effects but decreased for CRTs with large design effects. Coverage probabilities from RELR increased for almost all designs of CRTs compared to those obtained by analyzing complete data using the same statistical analysis method. When within-cluster MI was used to impute missing data, the same pattern for the coverage probabilities from both GEE and RELR was observed as when standard MI was used to impute missing data. It should be noted that the higher coverage from RELR when either standard or within-cluster MI strategy was applied prior to the analysis was not an indication of high efficiency, but rather a result of biased CS treatment effects and the corresponding underestimated standard effort.

4.3. Convergence problems

For the GEE method, at most 1 out of 1000 simulated datasets with S-design could not converge to a solution because they either encountered a non-positive definite matrix in the iterations or because there was no variation between the clusters in each arm. No convergence problems occurred for the simulated datasets based on the L-design. Lack of convergence was encountered more often for RELR than GEE. About 10 out of 1000 simulated datasets for some designs of CRTs could not converge for RELR due to negative estimates of between-cluster variance component during iteration.

5. DISCUSSION

In this paper, we compared the accuracy and efficiency of PA and CS models through a simulation study, in particular, the GEE method and the RELR respectively, for analyzing binary outcomes in CRTs with missing data. Results from the present simulation study, summarized in Table 5, show that under the assumption of CDM, the GEE method performs well as long as an appropriate strategy is applied to handle missing data based on the percentage of missing data and the design of CRTs. The appropriate strategy in this instance is using complete case analysis for any CRTs with small percentage of missing outcomes (<15%), using standard MI to impute missing outcomes for CRTs with small design effect (VIF<3), or within-cluster MI to impute missing outcomes for CRTs with large design effect (VIF≥3) and cluster size (>50). In contrast, the RELR performs poorly when either standard or within-cluster MI strategy is used to impute missing data prior to the analysis.

Results from the present comprehensive simulation study also imply that MI using random-effects logistic regression may not appropriate for imputing binary outcomes in CRTs. This is because that if the underlying data structure assumes a PA treatment effect, the MI using random-effects logistic regression, which impute missing data based on the CS treatment effect, may distort the original data structure and lead to invalid inference. Moreover, the convergence problems will greatly hinder the application of this method for imputing missing binary data. This implication seems to be in contradiction with current literature: for example, Taljaard *et al* [20] proposed mixed-effects regression imputation strategies to handle missing continuous outcomes in CRTs. Results from that study showed that the mixed-effects regression imputation strategy takes into account the between-cluster variance and therefore provides valid inferences for the treatment effect. In a previous study [21], we proposed MI using random-effects logistic regression to impute missing binary outcomes in CRTs and found that this strategy may be valid for imputing binary outcomes in CRTs. These two studies reached a different conclusion from the present simulation study since the mixed-effects regression imputation strategy by Taljaard *et al* is used to handle the missing continuous outcome, and the MI using random-effects logistic regression by Ma *et al* is based on a real dataset which has relatively large ICC, number of clusters per arm, and number of subjects per cluster, which limited the generalizability of their conclusions to more general settings.

MI has been accepted as a solution for missing data problems in many settings. Both GEE and RELR are commonly used for analyzing binary data in CRTs [1]. Results from this paper also imply that the choice of statistical analysis method and imputation method should reflect the same data structure as the inherent structure of the original data, otherwise, valid or improved inferences will not be achieved. For researchers with thorough understandings of the GEE method, RELR, CRTs, and the MI, results from this present study may not entirely surprising; however, the application of imputation and analysis methods in practice for CRTs does not reflect this finding. Some CRTs used mixed effects models for statistical analysis, but fixed-effects for clusters in imputation [22, 23, 24, 25]. In some other CRTs [26, 27, 28], no details were provided on which imputation procedure was applied. Findings from this simulation study urge caution on the use of RELR in the analysis of data from CRTs when missing binary outcomes are imputed by either standard or within-cluster MI strategy, thus improve the statistical practice in epidemiological research.

There are certain limitations to the current study. First, the performance of the marginal model and cluster-specific model are assessed only for CRTs with a completely randomized design. Other designs such as the matched pairs design and stratified randomized design are also used for CRTs but were not considered in this study. Second, only CRTs with balanced design were considered; however, settings found more often in empirical situations, such as unequal numbers of subjects per cluster, or unequal number of clusters in each trial arm, are not considered in this study. These design restrictions were made to understand the performance of the methods in simple scenarios. Further research is required to assess the extent to which our findings are relevant to more general settings. Third, there are two main approaches in handling missing data: likelihood based analyses and imputation [13]. In this paper, only complete case analysis, standard and within-cluster MI using logistic regression method to handle the missing data are considered; therefore, the conclusion from this paper regarding to the performance of RELR may not be applicable when missing data are handled using likelihood based analyses or other imputation methods. Further research may investigate the scenarios when missing data are handled by likelihood based analysis. Finally, the difference between the results from GEE and RELR is because the two models estimate different parameters, as outlined in the previous section. The intervention effect in the simulation has a population-average interpretation since the beta-binomial model is used to specify an overall unconditional probability within each trial arm, which gives preference to the GEE model.

6. CONCLUSIONS

Under the assumption of CDM, GEE method performs well as long as an appropriate missing data strategy is adopted based on the design of CRTs and the percentage of missing data. In contrast, RELR dose not perform well when either standard or within-cluster MI strategy is applied to impute missing data prior to the analysis.

7. ACKNOWLEDGEMENT

This study was supported in part by funds from the Canadian Network and Centre for Trials Internationally (CANNeCTIN) program and the Drug Safety and Effectiveness Cross-Disciplinary Training (DSECT) program in the form of training awards for the first author. No additional external funding received for this study. Dr. Lehana Thabane is a clinical trials mentor for the Canadian Institutes of Health Research (CIHR). Dr. Parminder Raina holds a Raymond and Margaret Labarge Chair in Research and Knowledge Application for Optimal Aging, and the Canada Research Chair in GeroScience.

Reference

[1] Donner A, Klar N. Design and Analysis of Cluster Randomization Trials in Health Research. New York: John Wiley & Sons; 2000

[2] Zeger SL, Liang KY, Albert PS. Models for longitudinal data: a generalized estimating equation approach. Biometrics 1988; 44:1049-60

[3] Austin PC. A comparison of the statistical power of different methods for the analysis of cluster randomization trials with binary outcomes. Statistics in Medicine 2007; 26: 3550-3565

[4] Bellamy SL, Gibbard R, Hancock L, Howley P, Kennedy B, Klar N, Lipsitz S, Ryan L. Analysis of dichotomous outcome data for community intervention studies. Statistical Methods for Medical Research 2000; **9**:135–59

[5] Ukoumunne OC, Carlin JB, Gulliford MC. A simulation study of odds ratio estimation for binary outcomes from cluster randomized trials. Stat Med. 2007; 26(18):3415-28

[6] Donner A, Brown KS, Brasher P. A methodological review of non-therapeutic intervention trials employing cluster randomization, 1979-1989. Int J Epidemiol. 1990 19(4):795-800

[7] Liang K, Zeger S. Longitudinal data analysis using generalized linear models. Biometrika 1986; 73(1): 13-22

[8] Murray DM, Varnell SP, Blitstein JL. Design and analysis of group-randomized trials: a review of recent methodological developments. American Journal of Public Health 2004; 94:423–432

[9] Mancl LA, DeRouen TA. A covariance estimator for GEE with improved small-sample properties. Biometrics 2001; 57:126–134]

[10] McCulloch CE, Searle SR. Generalized, Linear and Mixed Models. New York: John Wiley & Sons Inc; 2001

[11] Feng Z, Diehr P, Peterson A, McLerran D. Selected statistical issues in group randomized trials. Annual Review of Public Health 2001; 22:167–187

[12] Neuhaus JM. Statistical methods for longitudinal and clustered designs with binary responses. Statistical Methods in Medical Research 1992; 1:249–73

[13] Little RJA, Rubin DB. Statistical analysis with missing data. 2nd ed. New-York: John Wiley & Sons; 2002

[14] Rubin DB. Multiple imputation after 18+ years. Journal of the American Statistical Association 1996; 91: 473-89

[15] Barnard J, Rubin DB. Small-sample degrees of freedom with multiple imputation. Biometrika 1999; 86: 949-55

[16] Eldridge SM, Ashby D, Feder GS, Rudnicka AR, Ukoumunne OC. (2004) Lessons for cluster randomized trials in the twenty-first century: A systematic review of trials in primary care. Clin Trials 1(1): 80-90.

[17] Donner A. (1982) An empirical study of cluster randomization. Int J Epidemiol 11(3): 283-286.

[18] Lee EW, Bubin N. Estimation and sample size considerations for clustered binary responses. Statistics in Medicine 1994; 13:1241-52

[19] Burton A, Altman DG, Royston P, Holder RL. The design of simulation studies in medical statistics. Stat Med. 2006; 25(24):4279-92

[20] Taljaard M, Donner A, Klar N. Imputation strategies for missing continuous outcomes in cluster randomized trials. Biom J 2008; 50(3): 329-45

[21] Ma J, Akhtar-Danesh N, Dolovich L, Thabane L, CHAT investigators. Imputation strategies for missing binary outcomes in cluster randomized trials. BMC Med Res Methodol 2011; 11: 18

[22] Brown EC, Graham JW, Hawkins JD, Arthur MW, Baldwin MM, Oesterle S, Briney JS, Catalano RF, Abbott RD. Design and analysis of the community youth development study longitudinal cohort sample. Evaluation Review 2009; 33:311–24

[23] Hawkins JD, Brown EC, Oesterle S, Arthur MW, Abbott RD, Catalano R F. Early effects of communities that care on targeted risks and initiation of delinquent behavior and substance use. Journal of Adolescent Health 2008; 43:15–22

[24] Hawkins JD, Oesterle S, Brown EC, Arthur MW, Abbott RD, Fagan AA, Catalano RF. Results of a type 2 translational research trial to prevent adolescent drug use and delinquency. Archives of Pediatrics and Adolescent Medicine 2009; 163:789–98

[25] Clark NM, Shah S, Dodge JA, Thomas LJ, Andridge RR, Awad D, Little RJA. An evaluation of asthma interventions for preteen students. Journal of School Health 2010; 80:80–87

[26] French SA, Story M, Fulkerson JA, Himes JH, Hannan P, Neumark-Sztainer D, Ensrud K. Increasing weight-bearing physical activity and calcium-rich foods to promote bone mass gains among 9–11 year old girls: outcomes of the cal-girls study. International Journal of Behavioral Nutrition and Physical Activity 2005; 2: 8

[27] Pate RR, Ward DS, Saunders RP, Felton G, Dishman RK, Dowda M. Promotion of physical activity among high-school girls: a randomized controlled trial. American Journal of Public Health 2005; 95:1582–87

[28] Ganz PA, Farmer MM, Belman MJ, Garcia CA, Streja L, Dietrich AJ, Winchell C, Bastani R, Kahn KL. Results of a randomized controlled trial to increase colorectal cancer screening in a managed care health plan. Cancer 2005; 104:2072–83

Table 1. Comparison of empirical standard error

Design of CRTs			VIF [4]	% of missing data	Complete case analysis		Standard MI[5]		Within-cluster MI[6]	
m [1]	n [2]	ρ [3]			GEE[7]	RELR[8]	GEE	RELR	GEE	RELR
5 [9] (S-Design)	500	0.001	1.499	0%	0.07	0.10				
				15%	0.08	0.11	0.08	0.07	0.08	0.08
				30%	0.08	0.12	0.08	0.08	0.10	0.09
		0.01	5.99	0%	0.15	0.12				
				15%	0.15	0.13	0.13	0.12	0.16	0.14
				30%	0.15	0.15	0.12	0.11	0.16	0.15
		0.05	25.95	0%	0.30	0.15				
				15%	0.30	0.16	0.26	0.24	0.30	0.28
				30%	0.30	0.16	0.22	0.20	0.30	0.29
20 (L-Design)	50	0.01	1.49	0%	0.11	0.17				
				15%	0.11	0.17	0.12	0.12	0.13	0.13
				30%	0.12	0.19	0.12	0.13	0.15	0.16
		0.05	3.45	0%	0.17	0.31				
				15%	0.17	0.34	0.16	0.16	0.18	0.19
				30%	0.18	0.39	0.15	0.16	0.20	0.21
		0.1	5.90	0%	0.22	0.18				
				15%	0.23	0.21	0.20	0.22	0.23	0.26
				30%	0.23	0.22	0.18	0.19	NA	NA
30 (L-Design)	30	0.05	2.45	0%	0.15	0.28				
				15%	0.16	0.33	0.15	0.15	0.17	0.18
				30%	0.17	0.37	0.15	0.15	NA	NA
		0.1	3.90	0%	0.19	0.33				
				15%	0.20	0.38	0.18	0.19	NA	NA
				30%	0.20	0.42	0.17	0.18	NA	NA
		0.2	6.80	0%	0.26	0.38				
				15%	0.26	0.40	0.23	0.27	NA	NA
				30%	0.26	0.44	0.21	0.23	NA	NA

Note:
1. m: Number of clusters per trial arm;
2. n: Number of subjects per cluster;
3. ρ: Intracluster correlation coefficient;
4. VIF: Variance inflation factor, i.e. 1+(m-1) x ρ;
5. Standard MI: Standard multiple imputation using logistic regression method;
6. Within-cluster MI: Within-cluster multiple imputation using logistic regression method, which is not applicable (NA) for some L-design of cluster randomized trials;
7. GEE: Generalized estimating equations;
8. RELR: Random-effects logistic regression;
9. For CRTs with 5 clusters per arm, modified standard errors are provided

Table 2. Comparison of standardized bias

Design of CRTs			VIF [4]	% of missing data	Complete case analysis		Standard MI[5]		Within-cluster MI[6]	
m [1]	n [2]	ρ [3]			GEE[7]	RELR[8]	GEE	RELR	GEE	RELR
5 [9] (S-Design)	500	0.001	1.499	0%	0.02	0.73				
				15%	0.03	0.71	0.02	0.17	0.03	0.15
				30%	0.01	0.63	0.00	0.18	0.00	0.08
		0.01	5.99	0%	0.01	0.34				
				15%	0.00	0.33	0.00	0.02	0.00	0.03
				30%	0.00	0.32	0.00	0.01	0.00	0.03
		0.05	25.95	0%	0.02	0.15				
				15%	0.02	0.15	0.02	0.08	0.03	0.10
				30%	0.02	0.14	0.01	0.05	0.02	0.09
20 (L-Design)	50	0.01	1.49	0%	0.04	0.38				
				15%	0.04	0.37	0.03	0.04	0.06	0.01
				30%	0.04	0.36	0.03	0.05	0.08	0.01
		0.05	3.45	0%	0.01	0.26				
				15%	0.00	0.24	0.00	0.09	0.03	0.11
				30%	0.02	0.13	0.01	0.06	0.03	0.12
		0.1	5.90	0%	0.02	0.20				
				15%	0.01	0.19	0.01	0.15	0.05	0.16
				30%	0.01	0.19	0.01	0.10	NA	NA
30 (L-Design)	30	0.05	2.45	0%	0.02	0.33				
				15%	0.02	0.32	0.02	0.12	0.02	0.15
				30%	0.01	0.14	0.00	0.06	NA	NA
		0.1	3.90	0%	0.01	0.23				
				15%	0.01	0.23	0.01	0.18	NA	NA
				30%	0.02	0.23	0.02	0.13	NA	NA
		0.2	6.80	0%	0.01	0.16				
				15%	0.00	0.15	0.00	0.14	NA	NA
				30%	0.01	0.15	0.00	0.16	NA	NA

Note:
1. m: Number of clusters per trial arm;
2. n: Number of subjects per cluster;
3. ρ: Intracluster correlation coefficient;
4. VIF: Variance inflation factor, i.e. 1+(m-1) x ρ;
5. Standard MI: Standard multiple imputation using logistic regression method;
6. Within-cluster MI: Within-cluster multiple imputation using logistic regression method, which is not applicable (NA) for some L-design of cluster randomized trials;
7. GEE: Generalized estimating equations;
8. RELR: Random-effects logistic regression;
9. For CRTs with 5 clusters per arm, modified standard errors are provided

Table 3. Comparison of root mean squared error

Design of CRTs			VIF [4]	% of missing data	Complete case analysis		Standard MI[5]		Within-cluster MI[6]	
m [1]	n [2]	ρ [3]			GEE[7]	RELR[8]	GEE	RELR	GEE	RELR
5 [9] (S-Design)	500	0.001	1.499	0%	0.07	0.10				
				15%	0.08	0.10	0.08	0.06	0.08	0.06
				30%	0.08	0.11	0.08	0.07	0.09	0.08
		0.01	5.99	0%	0.14	0.17				
				15%	0.14	0.17	0.15	0.15	0.15	0.15
				30%	0.15	0.17	0.15	0.15	0.15	0.15
		0.05	25.95	0%	0.31	0.34				
				15%	0.31	0.34	0.31	0.32	0.31	0.33
				30%	0.31	0.34	0.31	0.32	0.31	0.33
20 (L-Design)	50	0.01	1.49	0%	0.11	0.13				
				15%	0.11	0.13	0.12	0.12	0.12	0.12
				30%	0.12	0.14	0.14	0.12	0.13	0.13
		0.05	3.45	0%	0.18	0.20				
				15%	0.18	0.21	0.18	0.19	0.18	0.19
				30%	0.19	0.20	0.19	0.20	0.19	0.20
		0.1	5.90	0%	0.24	0.26				
				15%	0.24	0.27	0.24	0.26	0.24	0.27
				30%	0.25	0.27	0.25	0.26	NA	NA
30 (L-Design)	30	0.05	2.45	0%	0.15	0.17				
				15%	0.16	0.18	0.16	0.16	0.15	0.17
				30%	0.16	0.17	0.16	0.17	NA	NA
		0.1	3.90	0%	0.20	0.21				
				15%	0.20	0.22	0.20	0.22	NA	NA
				30%	0.20	0.23	0.21	0.22	NA	NA
		0.2	6.80	0%	0.27	0.30				
				15%	0.27	0.30	0.28	0.33	NA	NA
				30%	0.28	0.30	0.28	0.31	NA	NA

Note:
1. m: Number of clusters per trial arm;
2. n: Number of subjects per cluster;
3. ρ: Intracluster correlation coefficient;
4. VIF: Variance inflation factor, i.e. 1+(m-1) x ρ;
5. Standard MI: Standard multiple imputation using logistic regression method;
6. Within-cluster MI: Within-cluster multiple imputation using logistic regression method, which is not applicable (NA) for some L-design of cluster randomized trials;
7. GEE: Generalized estimating equations;
8. RELR: Random-effects logistic regression;
9. For CRTs with 5 clusters per arm, modified standard errors are provided

Table 4. Comparison of coverage probability

Design of CRTs			VIF [4]	% of missing data	Complete case analysis		Standard MI[5]		Within-cluster MI[6]	
m [1]	n [2]	ρ [3]			GEE[7]	RELR[8]	GEE	RELR	GEE	RELR
5 [9] (S-Design)	500	0.001	1.499	0%	0.91	0.96				
				15%	0.92	0.97	0.93	0.97	1.00	0.99
				30%	0.93	0.97	0.95	0.98	1.00	0.99
		0.01	5.99	0%	0.92	0.79				
				15%	0.92	0.81	0.90	0.87	0.95	0.91
				30%	0.94	0.84	0.88	0.84	0.98	0.93
		0.05	25.95	0%	0.91	0.49				
				15%	0.91	0.52	0.89	0.83	0.93	0.89
				30%	0.93	0.52	0.83	0.77	0.96	0.90
20 (L-Design)	50	0.01	1.49	0%	0.94	0.98				
				15%	0.94	0.98	0.93	0.95	0.96	0.97
				30%	0.94	0.98	0.92	0.96	0.98	0.98
		0.05	3.45	0%	0.93	0.91				
				15%	0.93	0.92	0.90	0.89	0.94	0.94
				30%	0.93	0.93	0.87	0.88	0.95	0.96
		0.1	5.90	0%	0.93	0.78				
				15%	0.93	0.82	0.89	0.88	0.93	0.93
				30%	0.92	0.83	0.85	0.85	NA	NA
30 (L-Design)	30	0.05	2.45	0%	0.95	0.95				
				15%	0.96	0.96	0.93	0.93	0.97	0.96
				30%	0.95	0.96	0.91	0.92	NA	NA
		0.1	3.90	0%	0.95	0.91				
				15%	0.95	0.93	0.92	0.92	NA	NA
				30%	0.95	0.94	0.89	0.90	NA	NA
		0.2	6.80	0%	0.94	0.79				
				15%	0.94	0.81	0.90	0.89	NA	NA
				30%	0.94	0.85	0.85	0.85	NA	NA

Note:
1. m: Number of clusters per trial arm;
2. n: Number of subjects per cluster;
3. ρ: Intracluster correlation coefficient;
4. VIF: Variance inflation factor, i.e. 1+(m-1) x ρ;
5. Standard MI: Standard multiple imputation using logistic regression method;
6. Within-cluster MI: Within-cluster multiple imputation using logistic regression method, which is not applicable (NA) for some L-design of cluster randomized trials;
7. GEE: Generalized estimating equations;
8. RELR: Random-effects logistic regression;
9. For CRTs with 5 clusters per arm, modified standard errors are provided

Table 5. Summary of results

Design effect of CRTs[1]	Percentage of missing data	Missing data strategies	Validity of statistical analysis	
			GEE[5]	RELR[6]
VIF[2]<3	<15%	Complete case analysis	√	√
		Standard MI[3]	√	X
		Within-cluster MI[4]	X	X
	≥15%	Complete case analysis	X	X
		Standard MI[3]	√	X
		Within-cluster MI[4]	X	X
VIF[2]≥3	<15%	Complete case analysis	√	√
		Standard MI[3]	X	X
		Within-cluster MI[4]	√	X
	≥15%	Complete case analysis	X	X
		Standard MI[3]	X	X
		Within-cluster MI[4]	√	X

Note:
1. CRTs: Cluster randomized trials
2. VIF: Variance inflation factor
3. Standard MI: Standard multiple imputation
4. Within-cluster MI: Within-cluster multiple imputation, which is not applicable for CRTs with small cluster size
5. GEE: Generalized Estimating Equations
6. RELR: Random-effects logistic regression

Figure 1: Schematic overview of the simulation study

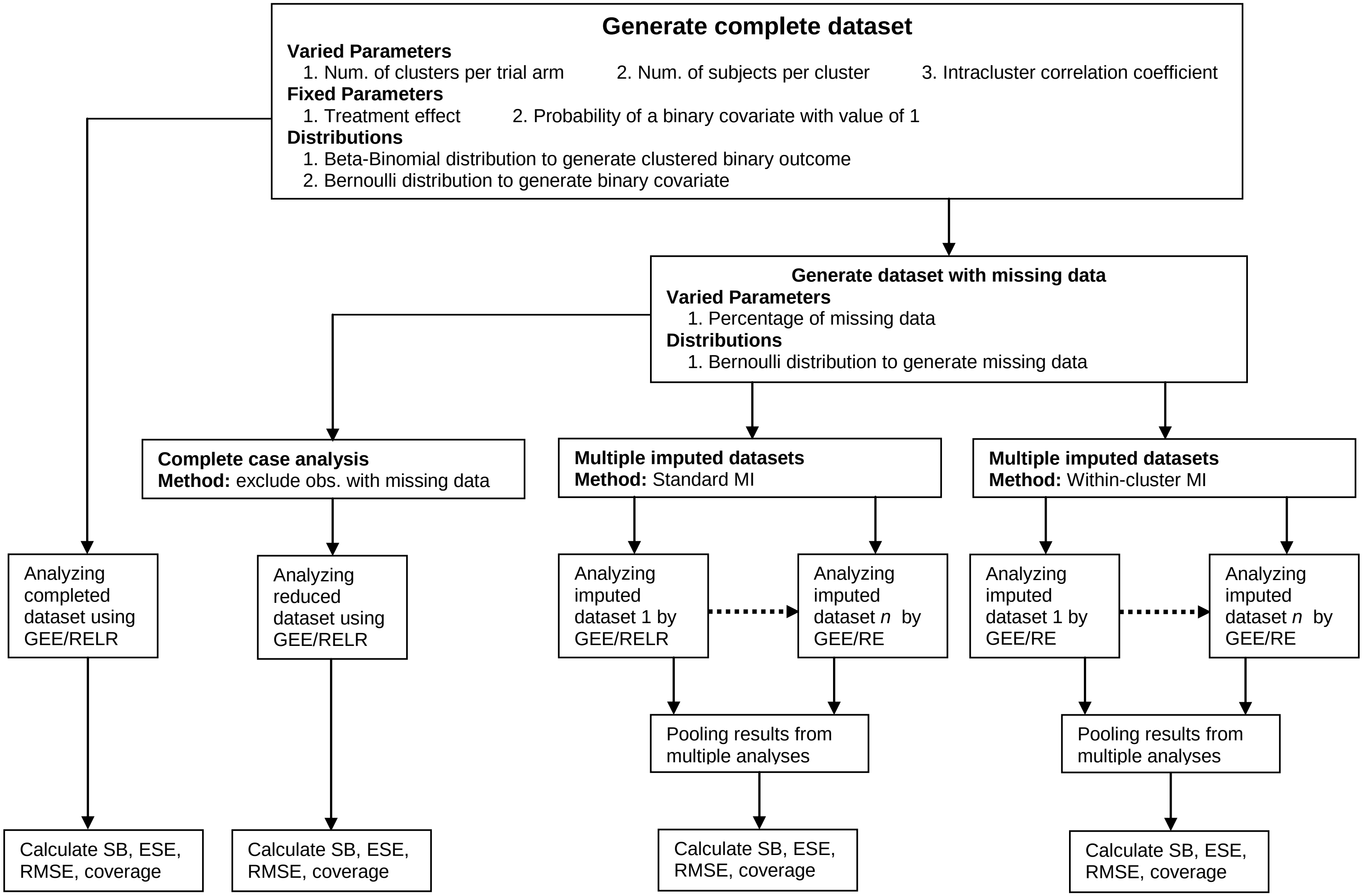

Generate complete dataset
Varied Parameters
1. Num. of clusters per trial arm
2. Num. of subjects per cluster
3. Intracluster correlation coefficient
Fixed Parameters
1. Treatment effect
2. Probability of a binary covariate with value of 1
Distributions
1. Beta-Binomial distribution to generate clustered binary outcome
2. Bernoulli distribution to generate binary covariate
Generate dataset with missing data
Varied Parameters
1. Percentage of missing data
Distributions
1. Bernoulli distribution to generate missing data
Complete case analysis
Method: exclude obs. with missing data
Multiple imputed datasets
Method: Standard MI
Multiple imputed datasets
Method: Within-cluster MI
Analyzing completed dataset using GEE/RELR
Analyzing reduced dataset using GEE/RELR
Analyzing imputed dataset 1 by GEE/RELR
Analyzing imputed dataset n by GEE/RE
Analyzing imputed dataset 1 by GEE/RE
Analyzing imputed dataset n by GEE/RE
Pooling results from multiple analyses
Pooling results from multiple analyses
Calculate SB, ESE, RMSE, coverage
Calculate SB, ESE, RMSE, coverage
Calculate SB, ESE, RMSE, coverage
Calculate SB, ESE, RMSE, coverage

Figure 2 Comparison of empirical standard error

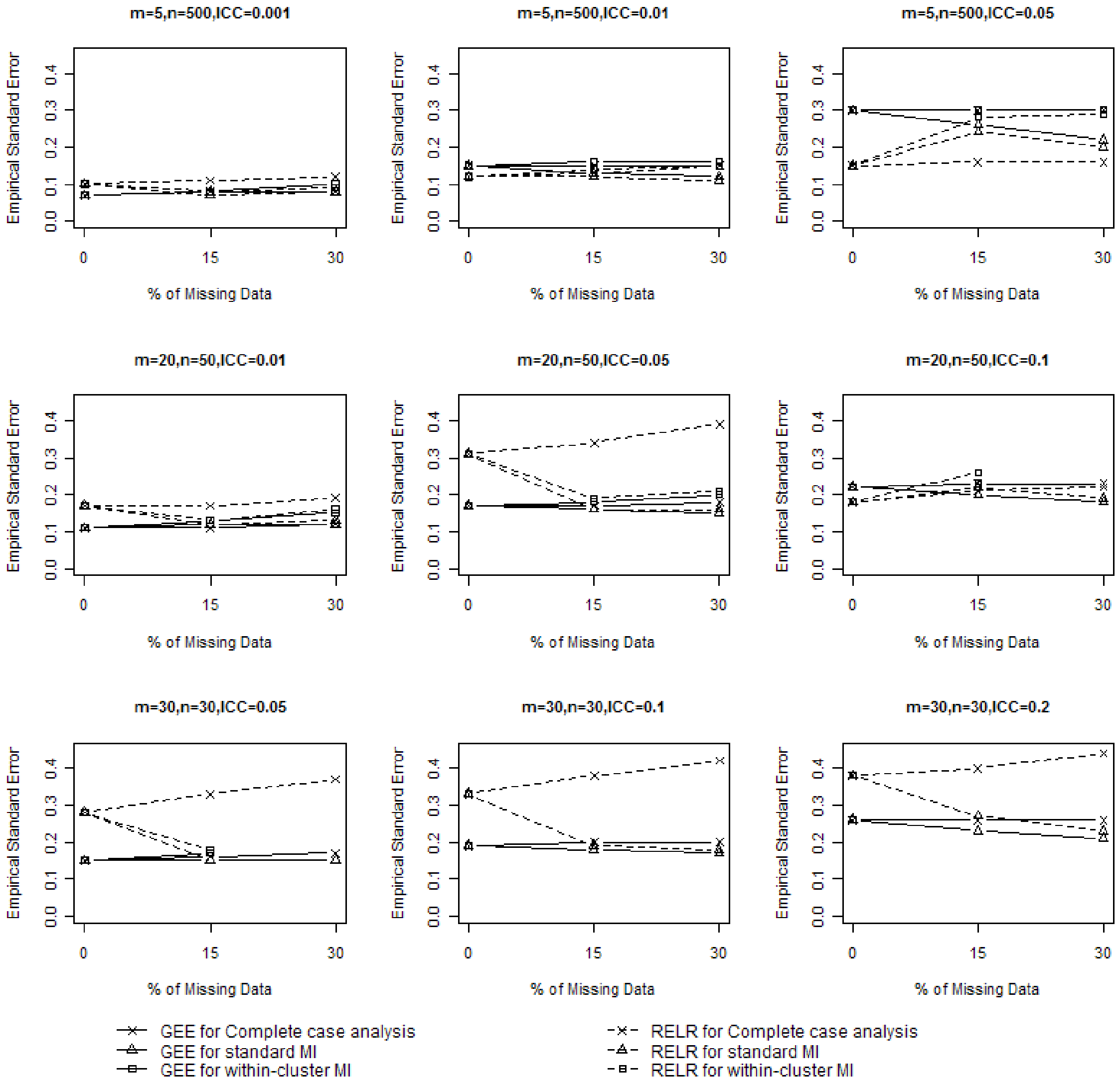


Figure 3 Comparison of standardized bias

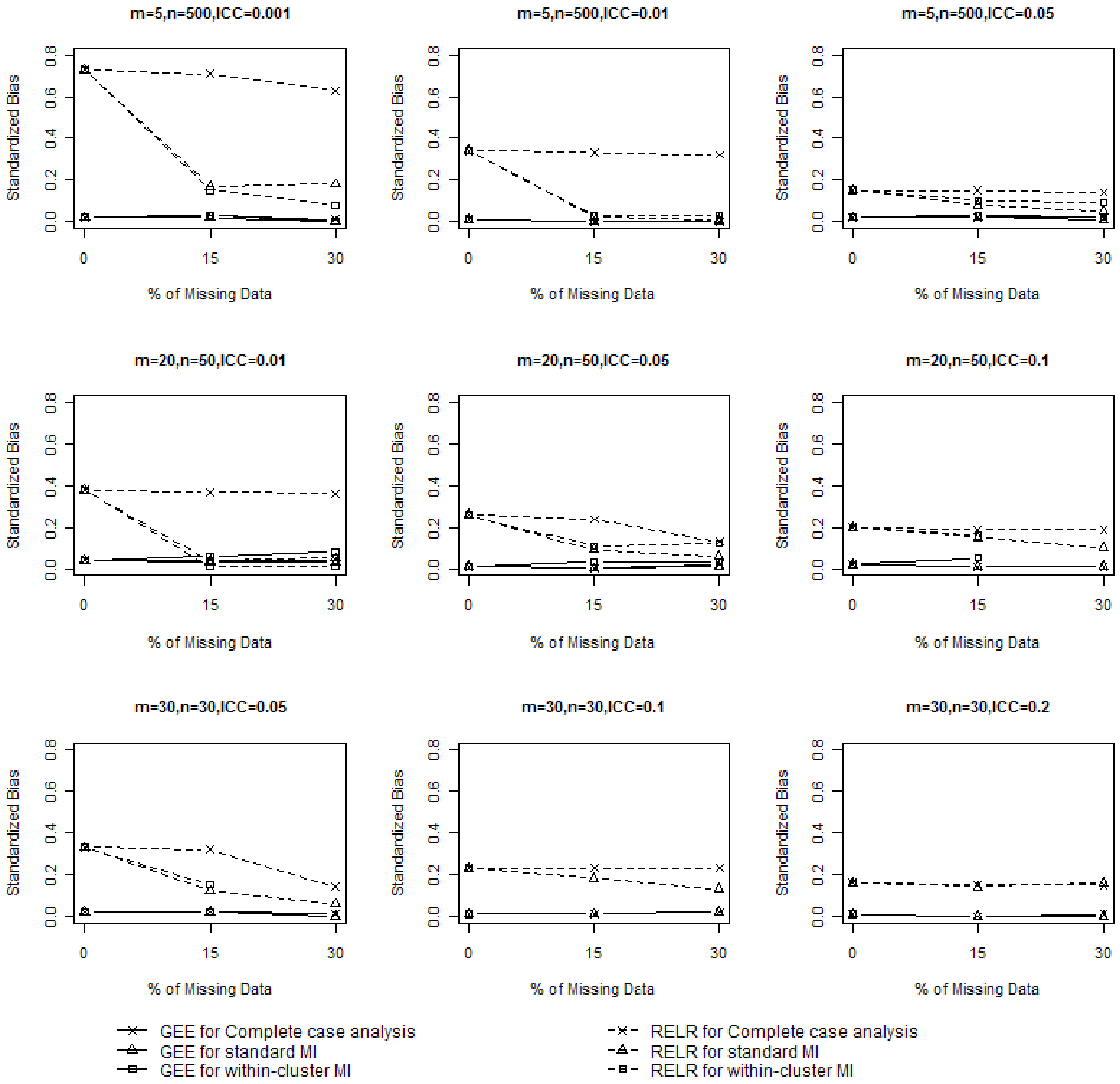


Figure 4 Comparison of root mean squared error

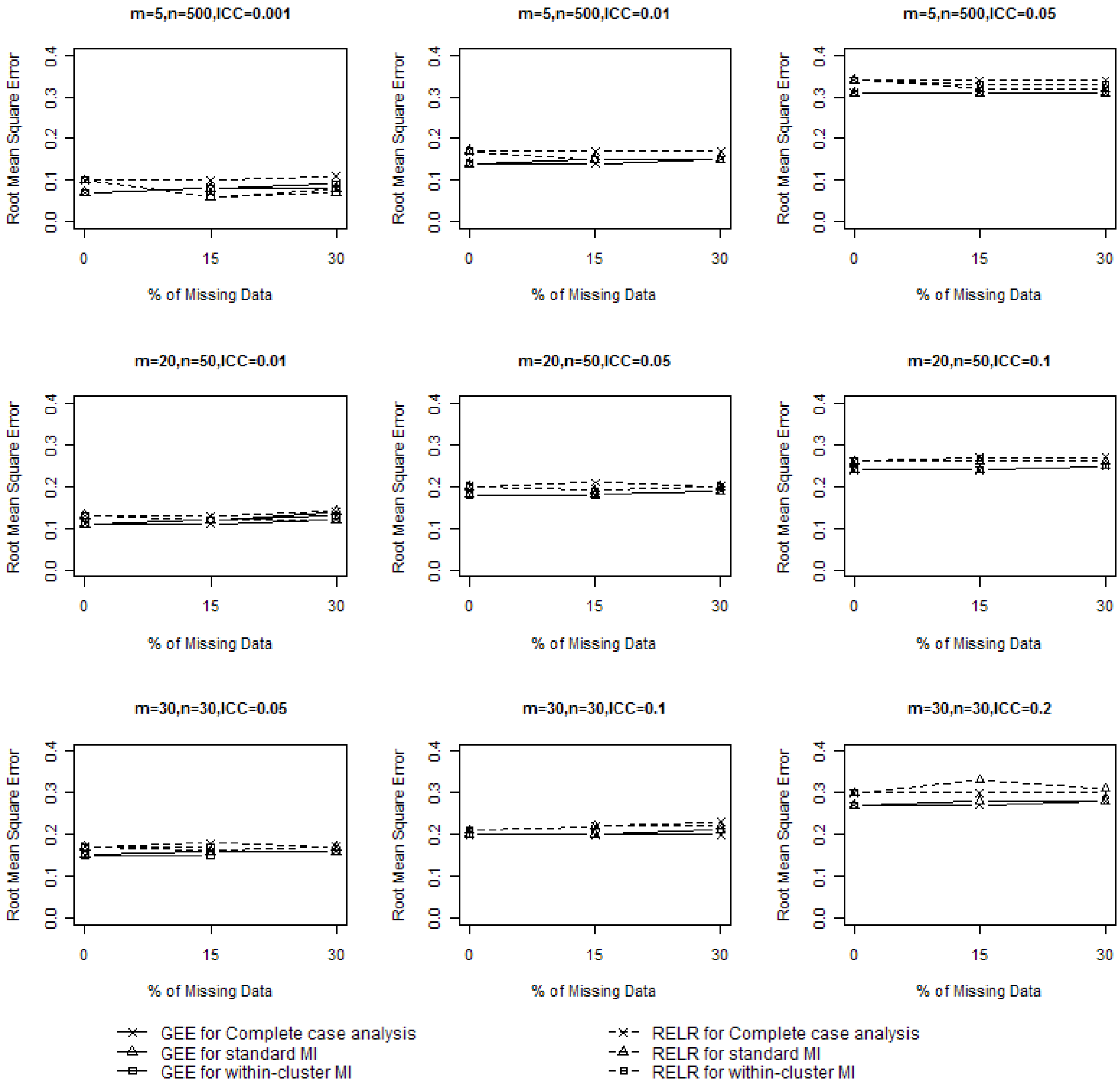
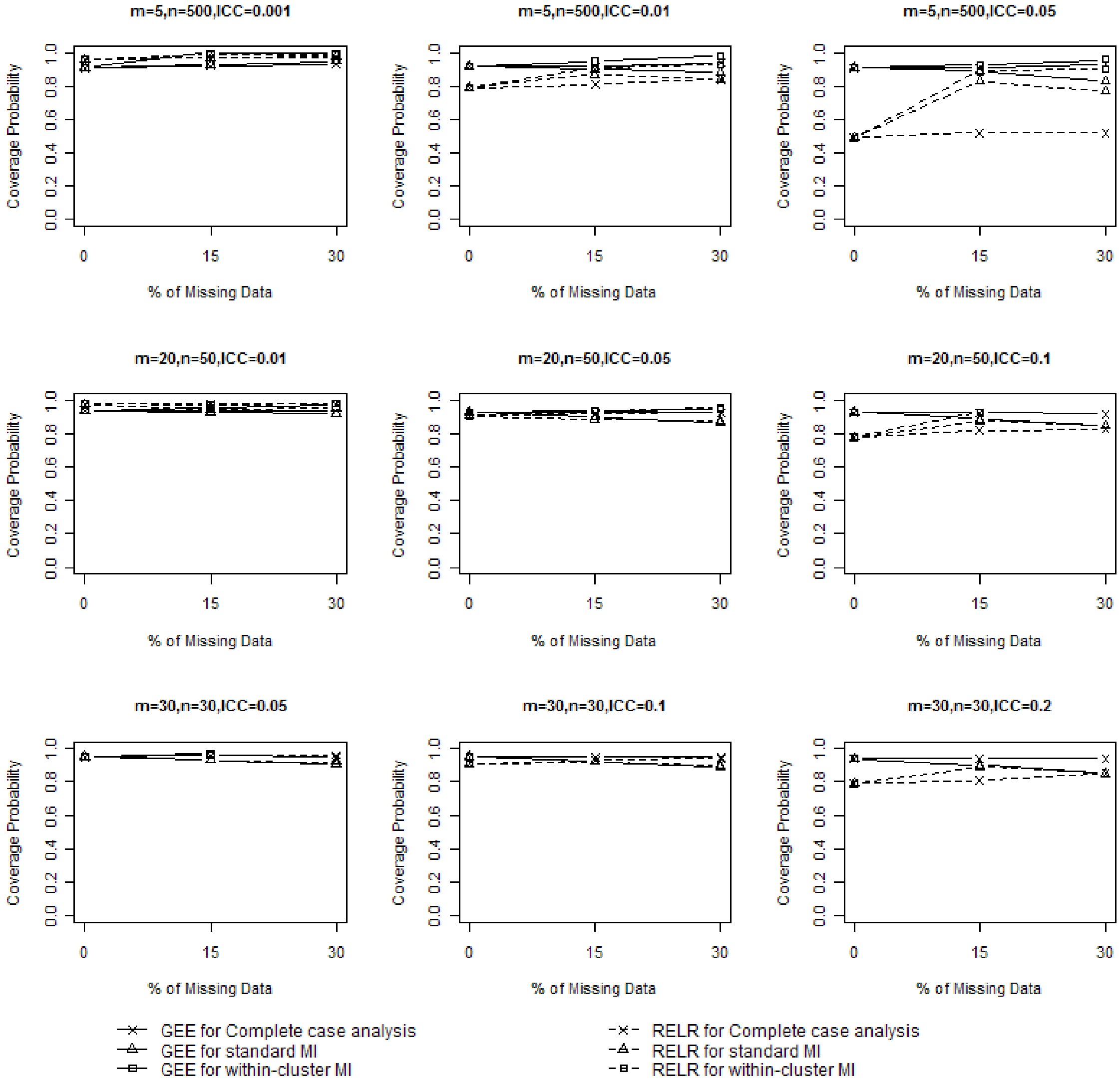


Figure 5 Comparison of coverage



CHAPTER 5

CONCLUSIONS

There are many methodological issues in the design and analysis of studies or trials with correlated data. A subset of these issues include: 1) rigorous sample size and statistical power calculation for longitudinal studies with interests in investigating the gene-environment interaction in disease susceptibility and progression; 2) determination of the most appropriate strategy to handle missing binary outcomes in cluster randomized trials (CRTs) according to their design settings; 3) determination of most appropriate statistical analysis method to analyze data from CRTs with missing binary responses. We conducted comprehensive simulation studies to address these important topics in a ”sandwich” thesis, with each chapter dedicated to investigating each of the issues. In this chapter, the key findings were summarized, and limitations and implications were discussed.

In Chapter 2, we investigated the power profile for the environmental and genotype risk exposures and their interaction on the transition from healthy to diseased state based on the design of Canadian Longitudinal Study on Aging (CLSA) [1], considering the transition from healthy to dead state as a competing risk. The measurement error on risk exposures and unmeasured etiological determinants were taken into account in the simulation to achieve a more accurate and realistic power profile.

The key findings in this chapter include:

(1) Given statistical power of 80%, significance level of 0.05 for environmental risk exposure and 0.0001 for genotype risk exposure and gene-environment

interaction, the design of CLSA, which involves 30,000 participants measured every three years for at least twenty years, enables moderate (1.5<HR≤2.0) or large (2.0<HR≤3.0) hazard ratio (HR) to be detected for environmental risk exposures.

(2) For genotype risk exposure, the CLSA is capable of detecting moderate HR when the incidence of disease is high, or the prevalence of genotype risk factor is high (≥0.1).

(3) For gene-environment interaction, even large HR can not be detected when the prevalence of genotype and environmental risk factors is low (<0.1).

(4) Misclassification on risk factors substantially reduces the statistical power.

(5) The HRs for designs involving data collection every three years are slightly larger than those obtained assuming exact event time is observed.

Though this simulation study is motivated and conducted based on the design of CLSA, its findings are generalizable to the design of similar population based longitudinal studies. First, an appropriate choice of frequency and timing of repeated measurements are closely related to the mean sojourn in healthy and diseased state. If the frequency of repeated measurements is considerably larger than the mean sojourn in diseased state, it is very likely that both the transition from healthy to diseased state and the subsequent transition to dead state will occur within the same data collection interval. Consequently, the statistical power will decrease since the transition from healthy to diseased state will not be observed. Second, the larger the time interval between two adjacent measurements, the higher chance that subjects are lost to follow-up within that time interval. In this case, transitions within that time interval will not be observed. Third, when assessments are expensive, increasing the frequency of repeated measurements will increase the cost. Therefore, the optimal design may only be determined once a cost function is specified. Forth, the above findings also suggest that higher statistical power may be achieved without increasing the cost through increasing the frequency of measurements for subjects with high risk of developing diseases or loss to follow-up, and slightly decreasing the frequency of measurements for other subjects. Since the incidence of aging related chronic diseases is usually higher for older people, to achieve higher statistical power, researcher may recruit higher proportion of seniors for achieving more incidence cases in a relatively short period. However, increasing the number of seniors in the sample may also cause a decrease in power due to two reasons: 1) transition from healthy to diseased state will not be observed during the follow-up for those subjects since they are more likely develop diseases before entering into the study; 2) incidence of new cases may not be observed since seniors are more likely to develop the disease and then die or lost to follow-up before the next measurement in 3 years comparing with mid-age subjects. Fifth, misclassification of the environmental and genotype risk exposures substantially increases the minimum detectable hazard ratio (MDHR). This is consistent with the finding from Garcia-Closas [2] *et al* that misclassification of environmental or genotype risk factors can substantially increase the sample size required to evaluate the gene-environment interaction in case-control studies. The magnitude of the increase in sample size is highly dependent on the misclassification rate on the risk exposures. Therefore, improving the accuracy in measuring both genotype and environmental risk exposures is crucial, especially for a valid assessment of gene-environment interaction. In addition, to the best of our knowledge, the design of this simulation is the first attempt to investigate the statistical power of a longitudinal study with analytic complexities such as delayed entry into the study, loss to follow-up, and increasing intensity of hazard to develop chronic diseases with time etc., being taken into account. It affords future researchers experiences and lessons that merit attention.

This project has some limitations. First, we assume the loss to follow-up rate is constant overtime. In practice, it may change with time and other variables. Second, accrual period is not considered since we assume all subjects enter into the study at the beginning of the CLSA. However, the influence of ignoring accrual period on the estimate of MDHR may be compensated by assuming all the subjects are followed up for 21 years in the simulation study. Third, we only estimate the MDHR for time-independent risk exposures. Fifth, we consider the misclassification of the risk exposure in this simulation study. However, the measurement error for the response variable, i.e. the accuracy of disease diagnosis, is not considered in the present study.

In Chapter 3, the performance of different strategies for handling missing binary outcomes in CRTs under different design settings was assessed.

The key findings include:

(1) The design of CRTs, including factors such as the number of clusters in each intervention group, the number of subjects within each cluster, the intracluster correlation coefficient (ICC) and the variance inflation factor (VIF), are important determinants for selecting an appropriate missing data strategy.

(2) Under the assumption of covariate dependent missingness (CDM) and application of the generalized estimating equations (GEE) approach for statistical analysis, complete case analysis can be used to obtain valid inference when the percentage of missing binary outcomes is small (<20%).

(3) Standard multiple imputation (MI) strategies using logistic regression or Markov chain Monte Carlo (MCMC) method can be used to impute the missing values when the design effect is small (VIF≤3); however, they tend to underestimate the standard error of the treatment effect when the design effect is large, though the underestimation of the standard MI using MCMC method is not substantial.

(4) Within-cluster MI using logistic regression may be an appropriate strategy to impute missing binary outcomes in CRTs, especially for CRTs with large cluster size and design effect. The performance of within-cluster MI using MCMC method is good for CRTs with large cluster size and design effect (VIF>3); however, it may yield biased estimates of treatment effect for CRTs with a small cluster size.

(5) The MI using logistic regression with the cluster as a fixed effect substantially overestimates standard error of the treatment effect for CRTs with small ICC (<0.05) and may result in biased estimates of treatment effect for CRTs with a small cluster size.

Findings from this study indicate that when the design effect of CRTs varies, different strategies may lead to varying results; therefore, the appropriate strategy needs to be chosen carefully to obtain valid inferences and mitigate design issues. However, very limited attention has been paid in the current literature on how to handle the missing binary outcomes in CRTs. Findings from this simulation study provide researchers with quantitative evidence to guide the selection of appropriate strategies based on the design settings of CRTs.

In Chapter 4, we compared the accuracy and efficiency of population-averaged (PA) and cluster-specific (CS) models through a simulation study, in particular, the GEE method and the random-effects logistic regression (RELR) respectively, for analyzing binary outcomes in CRTs with missing data. Results show that under the assumption of CDM, the GEE method performs well as long as an appropriate strategy is applied to handle the missing data based on the percentage of missing data and the design of CRTs. The appropriate strategy in this instance is to use complete case analysis for any CRTs with a small percentage of missing outcomes (<15%); use standard MI to impute missing outcomes for CRTs with a small design effect (VIF<3); or use within-cluster MI for CRTs with a large design effect (VIF≥3) and cluster size (>50). In contrast, the RELR performs poorly when either standard or within-cluster MI strategy is used to impute missing data prior to the analysis.

These findings have significant implications for future practice in health research. First, MI using random-effects logistic regression may not appropriate for imputing binary outcomes in CRTs. This is because that if the underlying data structure assumes a

PA treatment effect, the MI using random-effects logistic regression, which impute missing data based on the CS treatment effect, may distort the original data structure and lead to invalid inference. This conclusion seems to be in contradiction with current literature: for example, Taljaard *et al* [3] showed that the mixed-effects regression imputation strategy takes into account between-cluster variance; therefore, it provides valid inferences for the treatment effect when used to impute missing continuous outcomes in CRTs. In a previous study [4], we proposed MI using random-effects logistic regression to impute missing binary outcomes in CRTs and found that this strategy may be valid for imputing binary outcomes in CRTs. These two studies reached a different conclusion from the present study since the mixed-effects regression imputation strategy by Taljaard *et al* is used to handle the missing continuous outcome, and the MI using random-effects logistic regression by Ma *et al* is based on a real dataset which has relatively large ICC, number of clusters per arm, and number of subjects per cluster, which limited the generalizability of their conclusions to more general settings. Second, MI has been accepted as a solution for missing data problems in many settings. Both the GEE method and RELR are commonly used for analyzing binary data in CRTs [5]. Results from this paper also imply that the choice of statistical analysis method and imputation method should reflect the same data structure as the inherent structure of the original data; otherwise, valid or improved inferences will not be achieved.

For researchers with thorough understandings of the GEE method, the RELR, CRTs, and the MI, results from this present study may not entirely surprising; however, the application of imputation and analysis methods in practice for CRTs does not reflect this finding. Some CRTs used mixed effects models for statistical analysis, but fixed-effects for clusters in imputation [6, 7, 8, 9]. In some other CRTs [10, 11, 12], no details were provided on which imputation procedure was applied. Findings from this simulation study urge caution on the use of RELR in the analysis of data from CRTs when missing binary outcomes are imputed by either standard or within-cluster MI strategy, thus improve the statistical practice in epidemiological research.

The simulation studies in Chapter 3 and 4 have some common limitations. First, performance of missing data strategies and the statistical analysis models were assessed only for CRTs with a completely randomized design. Other designs, such as the matched pairs design and stratified randomized design, were not considered. Second, only CRTs with balanced design were considered; however, settings found more often in empirical situations, such as unequal numbers of subjects per cluster, or unequal number of clusters in each trial arm, were not considered in this study. These design restrictions were made to understand the performance of the methods in simple scenarios. Further research is required to assess whether these conclusions can be extended to more general settings.

References

1. Raina PS, Wolfson C, Kirkland SA, Griffith LE, Oremus M, et al. The Canadian longitudinal study on aging (CLSA). *Can J Aging* 2009; 28(3): 221-229.

2. Garcia-Closas M, Rothman N, Lubin J. Misclassification in case-control studies of gene-environment interactions: Assessment of bias and sample size. *Cancer Epidemiol Biomarkers Prev* 1999; 8(12): 1043-1050.

3. Taljaard M, Donner A, Klar N, Imputation strategies for missing continuous outcomes in cluster randomized trials. *Biom J* 2008; 50(3): 329-345.

4. Ma J, Akhtar-Danesh N, Dolovich L, Thabane L, CHAT investigators, Imputation strategies for missing binary outcomes in cluster randomized trials. *BMC Med Res Methodol* . 201; 11: 18.

5. Donner A, Klar N. Design and Analysis of Cluster Randomization Trials in Health Research. New York: John Wiley & Sons; 2000.

6. Brown EC, Graham JW, Hawkins JD, Arthur MW, Baldwin MM, Oesterle S, Briney JS, Catalano RF, Abbott RD. Design and analysis of the community youth development study longitudinal cohort sample. *Evaluation Review* 2009; 33:311–24.

7. Hawkins JD, Brown EC, Oesterle S, Arthur MW, Abbott RD, Catalano R F. Early effects of communities that care on targeted risks and initiation of delinquent behavior and substance use. *Journal of Adolescent Health* 2008; 43:15–22.

8. Hawkins JD, Oesterle S, Brown EC, Arthur MW, Abbott RD, Fagan AA, Catalano RF. Results of a type 2 translational research trial to prevent adolescent drug use and delinquency. *Archives of Pediatrics and Adolescent Medicine* 2009; 163:789–98.

9. Clark NM, Shah S, Dodge JA, Thomas LJ, Andridge RR, Awad D, Little RJA. An evaluation of asthma interventions for preteen students. *Journal of School Health* 2010; 80:80–87.

10. French SA, Story M, Fulkerson JA, Himes JH, Hannan P, Neumark-Sztainer D, Ensrud K. Increasing weight-bearing physical activity and calcium-rich foods to promote bone mass gains among 9–11 year old girls: outcomes of the cal-girls study. International Journal of Behavioral Nutrition and Physical Activity 2005; 2: 8.

11. Pate RR, Ward DS, Saunders RP, Felton G, Dishman RK, Dowda M. Promotion of physical activity among high-school girls: a randomized controlled trial. *American Journal of Public Health* 2005; 95:1582–87.

12. Ganz PA, Farmer MM, Belman MJ, Garcia CA, Streja L, Dietrich AJ, Winchell C, Bastani R, Kahn KL. Results of a randomized controlled trial to increase colorectal cancer screening in a managed care health plan. *Cancer* 2005; 104:2072–83.