2016

Covariate Misclassification under Covariate-Adaptive Randomization: Understanding the Impact and Method for Bias Correction

Liqiong Fan
Medical University of South Carolina

Covariate Misclassification under Covariate-Adaptive Randomization: Understanding the Impact and Method for Bias Correction

by

Liqiong Fan

A dissertation submitted to the faculty of the Medical University of South Carolina
in partial fulfillment of the requirements for the degree of Doctor of Philosophy
in the College of Medicine.

Department of Public Health Sciences
College of Medicine
Medical University of South Carolina

March 2016

APPROVED BY:

COMMITTEE CHAIR Sharon D. Yeatts, Ph.D.

SIGNED: Sharon D. Yeatts

MEMBER: Yuko Y. Palesch, Ph.D.

SIGNED: Yuko Y Palesch

MEMBER: Bethany J. Wolf, Ph.D.

SIGNED: Bethany J. Wolf

MEMBER: Leslie A. McClure, Ph.D.

SIGNED: L. A. McClure

MEMBER: Magdy Selim, M.D, Ph.D.

SIGNED: Magdy M. Selim

Contents

List of Figures

List of Tables

List of Abbreviation

AIC	Akaike Information Criteria
CI	Confidence Interval
CTHMM	Continuous-time Hidden Markov Model
DSMB	Data and Safety Monitoring Board
EVT	Endovascular Therapy
FDA	Food and Drug Administration
GCS	Glasgow Coma Scale
GOS-E	Extended Glasgow Outcome Scale
IMS III	Interventional Management of Stroke III Trial
ITT	Intention to Treat
MCSIMEX	Misclassification Simulation-Extrapolation Method
MM	Markov Model
mRS	modified Rankin Scale
NIH	National Institutes of Health
NIHSS	National Institutes of Health Stroke Scale
NINDS	National Institute of Neurological Disorders and Stroke
rt-PA	recombinant tissue plasminogen activator
OR	Odds Ratio
ProTECT III	Progesterone for Traumatic Brain Injury: Experimental Clinical Treatment III Trial
RCT	Randomized Controlled Trial
RR	Risk Ratio
SAP	Statistical Analyses Plan
TBI	Traumatic Brain Injury

Acknowledgments

I would like to express my deepest gratitude and appreciation to my supervisor Dr. Sharon D. Yeatts, who has provided invaluable mentoring throughout my graduate study. It is her who introduced the interesting misclassification issue from the real data to me, which is the main topic of my dissertation. Dr. Yeatts has provided insightful and instrumental advice on this thesis. Her infinite patience, continuous encouragement, and endless support has helped build up my confidence in doing research and to complete the degree. Her enthusiasm, diligence and dedication to the professional work also impresses me and inspires me to go further. It has been a real pleasure working with Dr. Yeatts.

I would also like to thank all my committee members, Dr. Yuko Y. Palesch, Dr. Bethany J. Wolf, Dr. Leslie A. McClure and Dr. Magdy Selim. I want to thank Dr. Palesch for taking me out while I was stressed out. I want to thank Dr. Wolf for valuable discussions with her and those pleasant conversations. I want to thank Dr. McClure for her encouragement along the way. I want to thank Dr. Selim for his practical points of view as a clinician. Thank all for their constructive comments on my dissertation, which make the thesis stronger and more rigorous.

Special thanks to Dr. Viswanathan Ramakrishnan and Dr. Elizabeth G. Hill. Their great encouragement and being always willing to help has been precious support during my study. Special thanks to Mrs. June Watson for her kind assistance in student affairs in all aspects. I am also grateful to my peers and the entire faculty in the Department of Public Health Sciences, for their warm supports over the past 5 years.

Last but not least, I want to thank my family: my wonderful husband, my lovely daughter, my parents who supported me to come to America to study, and also my friends: thanks for always being there and make life full of fun while studying!

Abstract

LIQIONG FAN. Covariate Misclassification under Covariate-Adaptive Randomization: Understanding the Impact and Method for Bias Correction (Under the direction of SHARON D. YEATTS).

Covariate-adaptive randomization has been frequently used in randomized controlled trials (RCTs) because it can well balance prognostic factors between treatment groups. However when a subject is assigned a wrong covariate value or misplaced in a wrong cohort during the randomization procedure, it may not only impact the balancing of the covariate, but also influence the treatment assignment based on the assigned cohort. Furthermore, it is preferred that covariates that are adjusted during the randomization procedure should also be adjusted for in the primary analysis. It is not clear whether a corrected covariate value, if it could be ascertained, or a misclassified covariate value should be used for the analysis, since the covariate value is tied both to the randomization procedure and analytic model. In this research, the impact of such misclassification on the type I error rate, power for hypothesis testing for the treatment effect and estimation bias of the treatment effect is explored under covariate-adaptive randomization in Aim 1. In Aim 2, a latent class model, the Continuous-time Hidden Markov Model (CTHMM) is used to estimate the misclassification issue with respect to both the estimation of misclassification probabilities and model diagnosis. An AIC based approach, which is calculated from a modified full data likelihood, is developed to test the assumption of misclassification. In Aim 3, a two-stage analysis strategy is proposed, which combines the CTHMM and the Misclassification Simulation-Extrapolation method (MCSIMEX), to correct the estimation bias of the perfectly measured variable caused by covariate misclassification. We apply the proposed analysis strategy to data from the Interventional Management of Stroke III trial to demonstrate the two-stage model.

Keywords: RCT, covariate misclassification, covariate-adaptive randomization, latent class model, CTHMM, MCSIMEX

1 Introduction & Background

1.1 Introduction

Randomized controlled trials (RCTs) have been widely used to test the treatment effect within a specific patient population and are considered the gold standard for claiming the efficacy of an intervention in practice (Chalmers et al., 1981; Moher et al., 2010). One of the key elements of the RCTs is the randomization procedure, which protects the validity and generalizability of the result of a trial. Simple randomization is easy to conduct, but balancing of baseline characteristics among groups is not guaranteed, especially when the sample size is small. Covariate-adaptive randomization has been frequently used (Thall and Wathen, 2005; Krag et al., 2010; Atagi et al., 2012; Sherrington et al., 2014; Ellis et al., 2014; Ybarra et al., 2013; Ersek et al., 2012; Weir and Lees, 2003) in RCTs because it can produce well-balanced groups and therefore, presents a more powerful and generalizable result. However, the operating characteristics of the test of the treatment effect are different under simple randomization and covariate-adaptive randomization (Hu and Hu, 2012; Shao and Yu, 2013). In order to obtain a valid test result under covariate-adaptive randomization, a correctly specified analytic model, which includes adjustments for all covariates used during the randomization procedure, is required.

On the other hand, the impact of covariate adjustment in the logistic regression model differs from that in the classic linear regression (Robinson and Jewell, 1991). With classic linear regression, an unbiased estimate for treatment effect can always be obtained, whether or not the model includes adjustment for covariates associated with the outcome, as long as there is no correlation between the treatment and the covariates (Robinson and Jewell, 1991). However, the treatment effect estimate is less precise, and the corresponding test is less powerful, when the model does not adjust for prognostic covariates. Under logistic regression, failure to adjust for a prognostic covariate (i.e., a covariate which is associated with the outcome) will always lead to biased estimate of the treatment effect, although adjusting for the covariate will increase the variance of the treatment effect estimate. That is, covariate adjustment will reduce the precision of the estimate but still result in a power gain

in terms of the asymptotic relative efficiency (Robinson and Jewell, 1991). Since RCTs often employ a primary analysis based on the logistic regression model in order to analyze binary outcomes, it is important to understand the implications of the selected randomization procedure on the analysis plan.

Because of the requirement of adjusting for the covariates included in the randomization algorithm, the incorrect measurement of those covariates could potentially cause problems for the pursuit of both appropriately balanced treatment arms and effective prognostic adjustments under covariate-adaptive randomization. Measurement error has been a long-standing topic in statistical literature. Misclassification is a special case of measurement error when the mis-measured variable is categorical. When a subject is assigned to a wrong covariate value or mis-placed into a wrong cohort during randomization procedure, it is equivalent to misclassification of the covariate. In an observational study, ignoring measurement error via naïve analysis using mis-measured data will cause biased estimation and power loss with respect to the misclassified variable itself (Buzas, 2006; Carroll et al., 2006). Methods have been developed to correct such bias (Begg and Lagakos, 1993). However, in a RCT setting, the emphasis is much more likely to be the treatment variable, which is considered perfectly recorded, after adjusting for the covariates rather than on the mis-measured covariates themselves. While the impact of covariate misclassification on the estimation of the perfectly measured variable during analysis has been demonstrated in previous observational studies (Buonaccorsi et al., 2005), few researchers have been actually focused on the perfectly measured variable in experimental studies especially that under covariate-adaptive randomization. When a trial employs covariate-adaptive randomization, the misclassified covariate is tied both to the treatment assignment and the analysis. Therefore, it is important to understand the impact of covariate misclassification on the treatment effect estimate, as well as the power and type I error of the corresponding hypothesis test under covariate-adaptive randomization. Whether and how to adjust for error-prone (misclassified) covariates in the analysis needs to be established and rationalized. Furthermore, presumably the bias will apply to the treatment effect estimation, indicating that methods that are specifically designed for bias correction for the treatment effect estimation need to be taken into consideration.

Misclassification error can be well-characterized by a misclassification matrix, if both the truth and the misclassified version of the covariate are known. In many real world applications, however, the truth is unobservable; in order to correct the bias in this case, additional information such as internal/external validation data sets or repeated measurements/replicas are necessary so that misclassification probabilities can be estimated.

In this thesis, the focus will be on effect estimation of the perfectly measured variable, i.e. the treatment effect, when categorical covariates are subject to misclassification. First, the impact of varying rates of covariate misclassification on the type I error rate and power associated with the hypothesis testing for the treatment effect, as well as bias of its estimation, under covariate-adaptive randomization is explored, where the misclassified covariate is adjusted during the randomization procedure. Because binary outcomes and/or event frequency are commonly encountered types of outcomes in medical research, the focus is on the generalized linear regression model. Secondly, given the impact of covariate misclassification on the treatment effect estimation, a method that can account for the uncertainty of the covariate measurement will be developed, thus providing a more accurate estimate of the treatment effect and potentially improving the power to detect it. Finally, the method will be applied to the data from a real trial example, and the results using the developed method compared to those of the naïve analysis, adjusting for the misclassified covariate, and the corrected model, adjusting for the corrected covariate.

1.1.1 Specific Aims

(1) **Specific Aim 1**

Explore the operating characteristics (type I error, power, and bias of treatment effect estimation) of different analysis strategies for dealing with covariate misclassification under covariate-adaptive randomization.

(2) **Specific Aim 2**

Assess the ability of the Continuous-Time Hidden Markov Model (CTHMM) to estimate the misclassification probabilities based on a repeatedly measured error-prone variable including

parameter identifiability and accuracy of the estimates.

(3) **Specific Aim 3**

Combine the CTHMM and MisClassification Simulation-Extrapolation method (MCSIMEX) in order to correct the bias of the treatment effect estimation and assess its robustness. Analyze real trial data (IMS III) using the joint modeling strategy in order to both estimate the misclassification probabilities and appropriately correct the estimation of the treatment effect, comparing with naive analysis.

1.2 Background

1.2.1 Motivation & Clinical Relevance

The motivation of this study came from two multicenter RCTs, both of which employed covariate-adaptive randomization. For the Interventional Management of Stroke (IMS) III Trial (Broderick, 2013), the primary objective was to determine the treatment effect of endovascular therapy following intravenous (IV) rt-PA, initiated within 3 hours of symptom onset of ischemic stroke, compared to the standard intravenous rt-PA approach alone. One of the covariates adjusted for during the randomization procedure was stroke severity defined based on the National Institute of Health Stroke Scale (NIHSS) score, ranging from 0-42. For the severity designation, the baseline NIHSS was dichotomized as $\leq$ 19 VS. $\geq$ 20, representing mild to moderate stroke and severe stroke, respectively. The NIHSS was also evaluated repeatedly at 40 minutes post IV tPA initiation, 24 hours and 5 days or discharge. The primary outcome of the IMS III trial was functional recovery based on the modified Rankin Scale (mRS) evaluated at 3 month after stroke onset, which was also dichotomized as good outcome (mRS 0-2) VS. poor (mRS $>$ 2). The NIHSS is an important prognostic covariate in that it has been shown to predict outcome well after stroke (Muir et al., 1996; Adams et al., 1999; Frankel et al., 2000; De Haan et al., 1993; Harrison et al., 2013; Glymour et al., 2007). However, the NIHSS was categorized for both randomization and analysis, rather than used in its original ordinal scale, due to ease of interpretation and flexibility of model assumptions. During the trial, data were reviewed by independent clinical monitors. About 2% of the subjects with severe stroke (NIHSS $\geq$ 20) were identified as having been misclassified as moderate stroke (NIHSS $\leq$ 19) during randomization, and a similar error rate was found among subjects with moderate stroke for their baseline severity evaluation.

The Progesterone for Traumatic Brain Injury: Experimental Clinical Treatment (ProTECT^{TM}) III Trial (Wright, 2014) was designed to determine the efficacy of intravenous progesterone administration, started within 4 hours of traumatic brain injury (TBI), compared to placebo. In the ProTECT Trial, one of the covariates adjusted for during the randomization procedure was the

severity of TBI based on the Glasgow Coma Scale (GCS) or its subdomain of motor score (Motor) if the subject was intubated. The original scale for GCS ranges from 3 to 15, with 3 representing a deep coma and 15 fully awake. The scale was trichotomized into moderate TBI (GCS 9-12), moderate-to-severe TBI (GCS 6-8 or Motor 4-5) and most severe TBI (GCS 4-5 or Motor 2-3). The GCS was evaluated repeatedly while subjects were hospitalized. The primary outcome was functional recovery defined by the Glasgow Outcome Scale-Extended (GOS-E) score at 6 months after the injury. A favorable outcome was defined as GOS-E 5-8. However, after data were reviewed by the trial monitors, the severity stratum were identified being recorded incorrectly among some subjects. In this trial, 8% of the subjects with most severe traumatic brain injury (TBI) were misclassified as moderate-to-severe TBI at baseline; 7% and 3% of moderate-to-severe TBI were misclassified as most severe and moderate TBI respectively; 1% and 14% of moderate TBI were misclassified as most severe and moderate-to-severe TBI. (Supplement Table 1) However, the use of GCS for evaluation is somewhat controversial due to its poor inter-rater reliability (Green, 2011). Therefore, the possibility of being misclassified based on GCS may be even higher in the ProTECT trial.

It is common in practice to use some scores to quantify disease severity, like the NIHSS for stroke and the GCS for TBI. Usually such scores have a relatively wide range so that different manifestations of the disease can be evaluated and incorporated to better describe the disease severity. However, the original scale may not be very helpful, from a clinical or a statistical perspective. Clinically, and most importantly, the association between disease severity and the outcome may not necessarily change in a linear fashion; for example subjects who scored 10 or 12 on the NIHSS may have the same probability of achieving a favorable outcome defined as mRS 0-2 at 90 days after stroke. Thus, there may be biological rationale for categorizing scores to better reflect the ordinal association between disease severity and the outcome. Statistically, while information may be lost when a continuous or ordinal measure (with wide range) is categorized, such categorization is sometimes necessary for model validity. First of all, the normality assumption is often violated when treating the original score as a continuous predictive variable in the model. Secondly, if the generalized linear regression model is used for analysis, the linearity in the logit assumption may also be questionable, which

may result in unreliable estimates. Therefore, a categorized or ordinal version with fewer levels of the covariate scores may be preferred instead.

However, although generic cutoff points that are commonly used in practice may exist for some scores, the choice of cutoff points is often data driven. For example, Schlegel et al. (2003) use 5 and 13 as cutoff points for the NIHSS to predict care level after acute hospitalization of stroke patients. In another study, the same authors use 5, 10 and 15 as cutoff points for the NIHSS (Schlegel et al., 2004) to predict care level among stroke patients treated with rt-PA. Briggs et al. (2001) define mild stroke as an NIHSS < 8 in order to determine the necessity of intensive care for the study with respect to favorable outcome rate, compared to admission to ward unit. Fischer et al. (2005) also use cutoff points of 10 or 12 to associate NIHSS to arteriographic findings in acute ischemic stroke. Given the uncertainty of these cutoff points, together with the potential evaluation error, categorizing such covariates may introduce more or less misclassification of the data, resulting in a biased estimate for the treatment effect and in power loss. Comparing to the model with the adjustment for the data without misclassification, we will get a biased estimate for the effect of treatment and reduced power. Therefore, statistical consideration that accounts for the existence of misclassification should be taken in order to improve the estimate while analyzing such data.

When a covariate-adaptive randomization scheme is employed and a potentially misclassified covariate is included to achieve the balance, it is not clear whether it is appropriate to use the corrected covariate, if it can be ascertained for the analysis, since the treatment is assigned based on the misclassified variable. Intuitively, corrected data are preferred when available; however, under Intention-to-treat (ITT) analysis, misclassified variables should be analyzed to be true to the randomization scheme. More importantly, it is not clear if using the corrected data will impact the validity and reliability of the trial result. It is important to understand the impact of covariate misclassification in the context of randomized controlled trials, especially those conducted under covariate-adaptive randomization, where the misclassified covariate is closely related to the treatment assignment. In trials conducted under simple randomization or in observational studies, naïve analysis which ignores the misclassification (either via a model with misclassified covariates

or omitting the covariate due to quality issues) may cause biased estimates of the treatment effect and power loss for corresponding hypothesis testing (Luo et al., 2012; Greenland, 1980; Flegal et al., 1986; Reade-Christopher and Kupper, 1991; Spiegelman et al., 2000; John, 1993). In trials conducted under covariate-adaptive randomization, the randomization process may have some "add-on" effect, especially on the power and/or type I error rate, since the misclassified covariate is involved in both the randomization procedure and the analysis.

1.2.2 Misclassification

Misclassification can be well characterized using a misclassification matrix, where each element of the matrix represents the probability of being correctly classified or misclassified given the truth. Suppose $X_i, i = 1, 2, 3, \ldots, n$ is the true class to which individual i belongs, where X_i has k categories with probability $\pi_j = Prob(X_i = j), j = 1, 2, 3, \ldots, k$ and $\boldsymbol{\pi} = (\pi_1, \pi_2, \ldots, \pi_{k-1})^T$. Instead of observing X_i, G_i is observed with certain probabilities given X_i, i.e. $O_{lj} = Prob(G_i = l|X_i = j), l = 1, 2, 3, \ldots, k$. Then the $k \times k$ misclassification matrix can be defined as:

$$\begin{pmatrix} q_{1,1} & q_{1,2} & \cdots & q_{1,k} \\ q_{2,1} & q_{2,2} & \cdots & q_{2,k} \\ \vdots & \vdots & \ddots & \vdots \\ q_{k,1} & q_{k,2} & \cdots & q_{k,k} \end{pmatrix}$$

where each element represents the classification probability of G , the observed value, given the true value. When there are only two categories, q_{11} and q_{22} are usually called sensitivity and specificity, i.e. the probability of correctly identifying true positive and true negative, respectively.

In general, there are two types of misclassification: differential and non-differential misclassification, depending on whether or not the error probabilities change according to the level of the outcome or the covariates. In the RCT setting, covariate misclassification is usually non-differential with respect to the outcome, since the value of the covariate is ascertained before the outcome is observed. However, misclassification may be differential with respect to other covariates or even the treatment

assignment when the randomization procedure does not work well. In particular when the number of the levels of a variable is more than 2, as in the ProTECT trial, the probabilities of being misclassified at each level of the covariate may be different. The misclassification probabilities may also depend on other covariates in the analysis.

1.2.3 Current methods for adjusting covariate misclassification

Many methods have been proposed to correct the bias and improve the accuracy of the estimate for the misclassified variable itself. However, to the best of our knowledge, few methods have been specifically developed for correcting the bias of the perfectly measured variable caused by covariate misclassification. In addition, there is limited research focused on the performance of correcting bias for the perfectly measured variable using methods developed for misclassified variables. The literature review below describes the methods of bias correction for the misclassified variable.

In order to correct a misclassified covariate, additional information is needed with respect to the mis-measured variable. This information includes: internal/external validation data, repeated measurements or replicas and instrumental variables. Pan et al.(Pan et al., 2006, 2009) describe a transition model to analyze data, where one of the covariates is measured with error. A transition model assumes that, conditioning on the history of the outcome and the covariates, the distribution of the current outcome satisfies the Markov property, i.e. the conditional distribution of the current outcome only depends on the q prior outcomes (q^{th} order Markov Chain). They prove that conditioning on the distribution of the history of the outcome and true covariate, the joint distribution of the current outcome and the history of the observed error-prone covariate belong to the exponential family and therefore, sufficient statistics for the true covariate can be obtained. The estimating equation is then constructed by summing over either the conditional score using the likelihood of the full data or the pseudo conditional score function based on conditional density of the outcome given the past history at each time point, which can be solved using the Newton-Raphson approach. However, their mis-measured variable is continuous and has an additive error which is independent of the truth and independently and identically distributed from a normal distribution with a known

variance. Also, both the outcome and mis-measured covariate are measured at the same time and repeatedly over time.

Other approaches described are the two stage models. In the first stage, extra information is used to estimate the parameters for the misclassification mechanism. The estimated parameters then can be incorporated into the analytic model as weights in the second stage. When internal/external validation data are available, for example, Lyles et al. (2010) propose to estimate the Positive Predictive Value (PPV) and Negative Predictive Value (NPV) via likelihood method by combining the observed data, outcome and an educated guess of sensitivity and specificity. Then, the original dataset is expanded using the estimated PPV and NPV as weights, which the analysis model is built upon to estimate the coefficient of interest. Shardell et al. (2014) treat misclassification in the missing data framework, where error-prone values are treated as missing data. After calculating a propensity score, an inverse-probability weighting method is applied for the analytic model for standardization using the complete-data. Chen et al. (2014) propose a two-stage estimation approach for longitudinal ordinal data with misclassification in both response and covariates based on estimating equations. Other so-called semi-parametric methods (Schafer, 2001; Wang CY, 1997; Roeder K, 1996) estimate parameters of the error distribution parametrically and apply nonparametric methods, such as pseudo conditional likelihood, kernel function or mixture model approach to the analytic model.

Fujisawa et al. (2000) discuss some identifiability conditions for misclassified repeated binary responses when repeated measurements or replicas of the error-prone variable are available. They show that at least three repeated measurements are needed in order to simultaneously estimate the misclassification parameters including false positive and false negative rates. White et al. (2001) demonstrate the possibility of using regression calibration for error-prone categorical variables when more than two replicas are available. Li et al. (2004) propose a two stage semiparametric Asymptotic Corrected Likelihood (ACL) estimator. The distribution of the truth is estimated by empirical characteristic functions and truncated inverse Fourier transform (methods proposed by Li et al. (2002), depending upon which the ACL is maximized. Wang et al. (2000) propose an

Expected Estimating Equation (EEE) method to account for the measurement error in longitudinal data. With misclassification, method of moments is recommended to solve for the EEE. It is also extended to the analysis of survival data with covariate misclassification (Wang and Song, 2013) or measurement error (Sinha and Ma, 2014). Wang et al. (1996) use quasi-likelihood models with misclassified covariate replaced by an estimate of the calibrated expected value of the covariate given other perfect measured variable and misclassified variable. The estimation is carried out both using approximations and through Monte Carlo simulation.

A limitation of all these methods is that they require either assumption of a known distribution of the underlying truth or estimation of its distribution given the observed data. None of the methods described above for handling a misclassified covariate takes into consideration the continuous fashion of disease progression and the ordinal nature of its assoication with the outcome at the same time , which is a reasonable and plausible assumption to make.

1.2.4 Markov chain and Hidden Markov Model

Markov chain is a stochastic process that describes the transitions from one state to another on a state space. It satisfies the Markov property that for a q^{th} order Markov chain, the future state only depends on q states that are prior to that state, but not the entire history of the previous states. Hidden Markov models (HMMs) are built upon an unobservable process, which follow a Markov chain. What is observed is a second stochastic process governed by the underlying unobserved states with certain emission probabilities for each state. Given the latent state, the probability of observing a certain state is called the emission probability. Conditioning on the underlying invisible states, the observed states at each time points are independent, i.e. the latent process contains all information that is needed to explain the observed behavior. Therefore, a HMM is also called a doubly stochastic process (Rabiner, 1990). The model structure can be graphically illustrated as following in Figure 1 (Boyer):

In the context of this research, the latent states $s_1, s_2, \ldots, s_t$ are the unobservable true disease

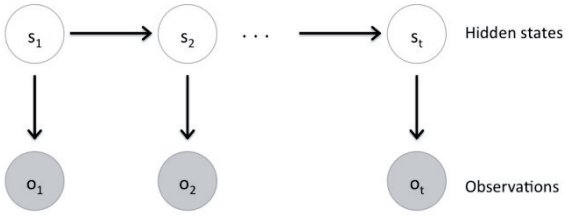

Figure 1: Markov Chain

severity to which a subject belongs at time t. The observed states $o_1, o_2, \ldots, o_t$ are the disease severity to which the subject is assigned based on some categorized score, like NIHSS or GCS at each time point. Therefore, the emission probabilities can be considered as the classification probabilities: the probability that a subject is assigned to a certain severity level, given that the subject truly belongs to a specified severity level, i.e. $Prob(o_t = l|s_t = j), l, j = 1, 2, \ldots, k$, where k is the total number of levels for which the score is categorized.

There are two general forms of the HMM: discrete and continuous time hidden Markov model (DTHMM and CTHMM). DTHMM is considered as the traditional approach to HMM. For DTHMM, the Markov chain is characterized by transition probabilities between hidden states, where the observation times are equally spaced steps and may not be closely relevant . For example, in speech recognition(Rabiner and Juang, 1993; Jelinek, 1997) or biological sequence analysis (Yoon, 2009), DTHMM is a good fit since switching between pronunciations or different functional region on a DNA sequence is observed exactly, thus the DHMM model is widely applied to analyze such data. CTHMM assumes the Markov chain is in a continuous-time fashion and is characterized by transition intensity between hidden states, i.e. the instantaneous transition probabilities. CTHMM is more suitable in medical research, in that disease progression itself is continuous, while the evaluation or observation of disease status is usually irregularly spaced. The exact transition time between states may not be observable, except for absorbing states like death. The application of CTHMM in describing disease progression can be found in the literature (Sweeting et al., 2006; Buter et al., 2008; Terrault et al., 2008; Mayet et al., 2012; Jackson and Sharples, 2002).

Another important advantage of using HMMs is that they automatically account for the potential

misclassification of the observed data, since they simultaneously model the transition between the underlying states and the emission probabilities given the underlying truth. In the presence of misclassification, the emission probabilities can be interpreted as misclassification probabilities(Jackson et al., 2003). Therefore, HMMs provide a convenient framework to estimate the misclassification probabilities by modeling the repeatedly measured error prone variable, while taking into consideration the time varying effect of the covariate. Bureau et al. (2003) apply CTHMM to study misclassified disease outcomes with two examples: the oral lesion hairy leukoplakia and cervical human papillomavirus (HPV) infection. The Markov property and the independence assumption allow simultaneous estimation of all possible transitions between disease states together with the probabilities of being misclassified. The structure of the resulting model is relatively simple. In addition, they also proposed a modification that can be used to include covariates in the model, thus providing a more flexible way to address different misclassification assumptions. However, the model goodness of fit and Markov property need to be assessed using diagnostic techniques. Jackson et al. (2003) also present a multistate Markov model for disease progression with the states of the disease subject to misclassification. The model is used to simultaneously estimate the sensitivity and specificity of ultrasonography screening for aortic aneurysm as well as the progression rate between latent stages. They also point out that pre-specifying the initial conditions of a Markov process was very important in order to get accurate estimation, which would be the starting value of the probability that underlying states occupied in the misclassification case . Van Den Hout et al. (2009) fit a three-state CTHMM to estimate life expectancy in health and ill health, with allowed misclassification in the observed improvement of cognitive ability. Gangnon et al. (2014) also apply CTHMM to investigate the impact of misclassification of age-related macular degeneration on the baseline intensity. After accounting for misclassification, the regression of the disease is rare, which is different from previous results reported in the literature.

1.2.5 Misclassification Simulation-Extrapolation Method (MC-SIMEX)

Misclassification Simulation-Extrapolation Model was an extension of the Simulation-Extrapolation (SIMEX) method for measurement of error model(Kuchenhoff et al., 2006). The SIMEX model was first introduced by Cook and Stefanski (1994). It was a simulation based method of inference which was simple to implement especially for fitting generalized linear models. The basic idea for this method is that by adding an extra known increment of error to the data, the pattern or trend of change in the estimation can be demonstrated. Then, the inference in the case of no measurement error can be obtained by extrapolating this trend back. One requirement for this method is that the trend of change is in general monotone convex or concave in order to get conservative corrections using best linear-tangential approximations. In 2006, Kuchenhoff, Mwalili and Lesaffre (Kuchenhoff et al., 2006) introduced the MCSIMEX model by extending the SIMEX method, which is a general regression method that deals with misclassification cases. It can be applied to misclassification of the outcome, predictors or both.

For the MCSIMEX model, a pre-specified misclassification matrix is needed in order to characterize the misclassification error and simulate the pseudo contaminant data sets as in the SIMEX method. The misclassification matrix is first factorized based on the diagonal matrix of eigenvalues and corresponding matrix of eigenvectors. Then the extra increment of the error is added with a fixed grid of values (integers such as $1, 2, \ldots, n$), raising to the power to the diagonal matrix of the eigenvalues. In the simulation step, for each increment of the error, B new pseudo data sets are simulated using the misclassification matrix with the additional error. The estimate of the coefficients in the model is then an average over those obtained based on the B pseudo data sets. In the extrapolation step, a parametric model is fit on the estimated coefficients obtained in the simulation step and the fixed grid of values by least squares. The forms of the extrapolation function (parametric model) can be linear, quadratic, exponential, or log-linear functions, with the quadratic extrapolation function performing the best in most cases.

According to their simulation results, the MCSIMEX method performs better than simple matrix

method with respect to the coverage probabilities for the parameter of error-prone variable, although they show similar bias correction abilities. Compared to the maximum likelihood method, the MCSIMEX model does not perform as well. However, due to its ease of implementation and flexibility of dealing with more complicated modeling, as well as its comparable ability for bias reduction and good coverage probability, it is still suggested to use the MCSIMEX model when applicable.

2 Original Manuscript 1

Title: The Impact of Covariate Misclassification using Generalized Linear Regression under Covariate-Adaptive Randomization

Authors: Liqiong Fan(a) ,Sharon D. Yeatts (a), Bethany J. Wolf (a), Leslie A. McClure (b) , Magdy Selim (c), Yuko Y. Palesch (a)

Affiliation:

(a) Department of Public Health Sciences, Medical University of South Carolina, Charleston, SC 29425, U.S.A.

(b) Department of Epidemiology & Biostatistics, Drexel University, Philadelphia, PA 19104, U.S.A.

(c) Beth Israel Deaconess Medical Center, Department of Neurology, Division of Cerebrovascular Diseases, Boston MA 02215, USA.

Submission Status: published in the *Statistical Methods in Medical Research* Journal

Abstract

Under covariate adaptive randomization, the covariate is tied to both randomization and analysis. Misclassification of such covariate will impact the intended treatment assignment; further, it is unclear what the appropriate analysis strategy should be. We explore the impact of such misclassification on the trial's statistical operating characteristics. Simulation scenarios were created based on the misclassification rate and the covariate effect on the outcome. Models including unadjusted, adjusted for the misclassified, or adjusted for the corrected covariate, were compared using logistic regression for a binary outcome and Poisson regression for a count outcome. For the binary outcome using logistic regression, type I error can be maintained in the adjusted model but the test is conservative using an unadjusted model. Power decreased with both increasing covariate effect on the outcome as well as the misclassification rate. Treatment effect estimates were biased towards the null for both the misclassified and unadjusted models. For the count outcome using a Poisson model, covariate misclassification led to inflated type

I error probabilities and reduced power in the misclassified and the unadjusted model. The impact of covariate misclassification under covariate-adaptive randomization differs depending on the underlying distribution of the outcome.

Keywords: Covariate-adaptive randomization; Generalized linear regression; Covariate misclassification.

2.1 Introduction

The randomized controlled trial (RCT) is considered the gold standard for valid inference regarding the efficacy of an intervention. One of the key elements of an RCT is its randomization procedure which assures the validity and generalizability of the trial's results. Simple randomization is often used due to its ease of implementation; however, large imbalance in baseline covariates across treatment arms may result, especially when sample size is not sufficiently large. Covariate-adaptive randomization has been frequently used in order to control the imbalance(Thall and Wathen, 2005; Krag et al., 2010; Atagi et al., 2012; Sherrington et al., 2014; Ellis et al., 2014; Ybarra et al., 2013; Ersek et al., 2012; Weir and Lees, 2003).A trial well-balanced on prognostic covariates is more powerful in the comparison of the treatment effect and yield a more convincing result (Kundt, 2009). It will also increase the power for subgroup analyses (Toorawa et al., 2009), since it balances the treatment arms among subjects with a given factor. However, under covariate adaptive randomization, the intended type I error rate can be maintained only when the model is correctly specified, which means that all covariates included in the randomization procedure are also included in the analytic model (Hu and Hu, 2012; Shao and Yu, 2013).

Unfortunately, covariates are sometimes measured with error, and in the case of categorical covariates, we consider these subjects to be misclassified. The rate of misclassification may vary from trial to trial and from covariate to covariate. For example, according to the review of Data and Safety Monitoring Board, in the Interventional Management of Stroke (IMS) III trial (Broderick, 2013), about 2% of subjects with severe stroke were misclassified as having a moderate stroke, and similarly about 2% of subjects with moderate stroke were misclassified as having a severe stroke at the time of randomization. In the Progesterone for Traumatic Brain Injury: Experimental Clinical Treatment (ProTECT III) trial, 13.8% of the subjects with most severe traumatic brain injury (TBI) were misclassified as having a moderate to severe TBI; 7% and 3% of moderate severe TBI were misclassified as having a most severe and moderate TBI, respectively; 1% and 14% of moderate TBI were misclassified as most severe and moderate severe TBI.

Under the covariate-adaptive randomization scheme, the treatment assignment is tied to the misclassified covariate; as a result, it is not clear whether the analysis should be based on the misclassified covariate information or the corrected data. Intuitively, corrected values are preferred when available; however, under intention-to-treat (ITT) analysis, misclassified variables should be analyzed in order to remain true to the randomization scheme. Naïve analysis which ignores the misclassification (either via a model with misclassified covariates or without adjustment) may cause biased estimates of the treatment effect, which may directly impact the power of the trial, as well as its validity and reliability (Luo et al., 2012; Greenland, 1980; Flegal et al., 1986; Reade-Christopher and Kupper, 1991; Spiegelman et al., 2000; John, 1993). Moreover, unlike simple randomization, covariate-adaptive randomization may have some "add-on" effect, especially on the power and/or type I error. Therefore, it is important to understand the impact of misclassification errors in the context of RCTs, where the power, type I error and bias in the treatment effect are important components to the success of a trial.

With classic linear regression, if the covariate is only correlated with the outcome but not the treatment, an unbiased estimate for treatment effect can always be obtained whether or not that covariate is included in the model (Robinson and Jewell, 1991). Therefore, covariate misclassification will not impact the point estimate of the treatment effect but will reduce the power of hypothesis testing for the treatment effect. Under the generalized linear regression frame work, where the outcomes have an underlying binomial or Poisson distribution, the impact of covariate adjustment is different (Robinson and Jewell, 1991). For a binary outcome with logistic regression, if the true outcome distribution is conditioned on the covariate (i.e. there is covariate effect on the outcome), failure to adjust for the covariate results in a biased estimate for the treatment effect (Robinson and Jewell, 1991; Gail et al., 1984). On the other hand, inclusion of the covariate increases the variance of the estimate for the treatment effect. That is, adjusting for a prognostic covariate will reduce the precision of the treatment effect estimate, but still results in power gain for the test of treatment in terms of the asymptotic relative efficiency (Robinson and Jewell, 1991).

Given the above impact of covariate adjustment in both randomization and analysis, it is necessary

to explore the influence of covariate misclassification on type I error, power and estimation bias for the treatment effect under covariate-adaptive randomization. In this paper, we investigate this impact, with a focus on analysis via generalized linear regression models in the scenario where the true covariate value is available at the time of the analysis. We provide the results of a simulation study conducted under covariate-adaptive randomization. Section 2 provides a theoretical explanation for the simulation result. The simulation methods are described in Section 3 and the results shown and explained in Section 4. We provide the conclusion and discussion about the impact in practice and some recommendations in the last section.

2.2 The validity of the tests when covariate is misclassified

2.2.1 Outcome with underlying binomial distribution

Shao et al. (2013) has shown that under covariate-adaptive randomization, the nominal level of the test will be maintained only when the covariates are adjusted for in the analysis; the test without covariates is conservative unless there is no covariate effect on the outcome. When the covariate adjusted for is misclassified, the nominal level of the test will still be maintained. To show this, suppose a pooled table and sub-tables are set up as in Table 1; where Y, Z, X and G represent outcome, treatment assignment, true covariate and misclassified covariate respectively. The prevalence of the covariate in the data is m, i.e.$Prob(X = 1) = m$.

Table 1: Expected cell frequencies for pooled and sub-tables

Pooled Table				Sub-table $X = 1$				Sub-table $X = 0$			
	$Y = 1$	$Y = 0$	Total		$Y = 1$	$Y = 0$	Total		$Y = 1$	$Y = 0$	Total
$Z = 1$	a	b	N_1	$Z = 1$	a_1	b_1	$N_1 m$	$Z = 1$	a_0	b_0	$N_1(1-m)$
$Z = 0$	c	d	N_0	$Z = 0$	c_1	d_1	$N_0 m$	$Z = 0$	c_0	d_0	$N_0(1-m)$

Assume the misclassification probabilities $k_i, i = 0, 1$ are non-differential with respect to the outcome Y, i.e.$Prob(G = 1|X = 0) = k_0, Prob(G = 0|X = 1) = k_1$. Then the resulting expected cell frequencies for the sub-tables stratified by the misclassified covariate G are in Table 2.

Table 2: Expected cell frequencies for sub-tables by misclassified covariate

	Sub-table $G=1$				Sub-table $G=0$		
	$Y=1$	$Y=0$	Total		$Y=1$	$Y=0$	Total
$Z=1$	$a_1{}^*(1-k_1)$ $+a_0{}^*k_0$	$b_1{}^*(1-k_1)$ $+b_0{}^*k_0$	$N_1(1-m)^*k_0$ $+\ N_1m*(1-k_1)$	$Z=1$	$a_1{}^*k_1+$ $a_0{}^*(1-k_0)$	$b_1{}^*k_1+$ $b_0{}^*(1-k_0)$	$N_1m*(1-k_0)$ $+\ N_1(1-m)*k_1$
$Z=0$	$c_1*(1-k_1)$ $+c_0*k_0$	$d_1*(1-k_1)$ $+d_0*k_0$	$N_0(1-m)^*k_0$ $+\ N_0m^*(1-k_1)$	$Z=0$	$c_1{}^*k_1+$ $c_0{}^*(1-k_0)$	$d_1{}^*k_1+$ $d_0{}^*(1-k_0)$	$N_0m\ ^*(1-k_0)$ $+\ N_0(1-m)^*k_1$

Then,

$$Prob(Y=1|X=1,Z=1)=\frac{a_1}{N_1m} \tag{1}$$

$$Prob(Y=1|X=1,Z=0)=\frac{c_1}{N_0m} \tag{2}$$

Under the null distribution where there is no treatment effect, given $X=1$, the probabilities in (1) and (2) are equal, i.e. $\frac{a_1}{N_1m}=\frac{c_1}{N_0m}$. This can also be written as $\frac{a_1}{c_1}=\frac{N_1m}{N_0m}=t$, where t is the constant ratio between a_1 and c_1,as well as N_1 and N_0.Then a_1 and N_1 can be expressed as $a_1=tc_1$ and $N_1=tN_0$. The ratio between a_0 and c_0 equals to t as well because given $X=0$, under the null distribution, $\frac{a_0}{N_1(1-m)}=\frac{c_0}{N_0(1-m)}$ and $\frac{a_0}{c_0}=\frac{N_1}{N_0}=t$. When there is misclassification as in Table 2, the probability of $Y=1$ can be expressed as:

$$Prob(Y=1|G=1,Z=1)=\frac{a_1(1-k_1)+a_0k_0}{N_1m(1-k_1)+N_1(1-m)k_0} \tag{3}$$

$$Prob(Y=1|G=1,Z=0)=\frac{c_1(1-k_1)+c_0k_0}{N_0m(1-k_1)+N_0(1-m)k_0} \tag{4}$$

Using the ratio notation defined above, then (3) becomes

$$\begin{aligned} &Prob(Y=1|G=1,Z=1)\\ &\quad=\frac{tc_1(1-k_1)+tc_0k_0}{tN_0m(1-k_1)+tN_0(1-m)k_0}\\ &\quad=\frac{c_1(1-k_1)+c_0k_0}{N_0m(1-k_1)+N_0(1-m)k_0}\\ &\quad=Prob(Y=1|G=1,Z=0) \end{aligned}$$

That is, although the proportion of $Y = 1$ is altered by the misclassification process from (1) to (3), the magnitude of the change is the same across treatment arms under the null distribution. Thus, if there is no treatment effect given the true covariate value $X = 1$, there is no treatment effect given the misclassified covariate $G = 1$. The same approach can be used to demonstrate the impact given $X = 0$ and $G = 0$. Therefore, the type I error probability is not affected by the misclassification, regardless of whether adjustment is made for the true covariate or the misclassified covariate.

2.2.2 Outcome with underlying Poisson distribution

The asymptotic property of the Poisson log-linear regression models was also developed by Shao et al. (2013). With the quasi-Poisson regression models, the estimated variability of the outcome based on the unadjusted model then becomes $\varphi E(Y_{ij})$, where φ is the estimated over-dispersion parameter and is greater than 1 if the data are over-dispersed. Therefore, an extension to the equation in Shao et al. becomes:

$$lim_{N\to\infty} pr_y(|T_s| > C\alpha) = 2\Phi\left(\frac{-C_\alpha \varphi E(Y_{ij})}{E\{E(Y_{ij}|X_i)\}}\right)$$

which reduces to $2\Phi(-C_\alpha\varphi)$, and the test is conservative in the unadjusted model of quasi-Poisson regression under covariate-adaptive randomization. However, the estimated over-dispersion parameter in a naive model of quasi-Poisson regression will depend on the amount of misclassification in G, the misclassified covariate. And the overall variability could be estimated correctly. Thus the type I error could be maintained. The same rationale works for the negative binomial regression model.

2.3 Simulation scenarios and hypothesis testing

Suppose a randomized controlled trial is going to be carried out with a goal of detecting an absolute 10% difference in a favorable outcome, with a 40% success rate assumed in the control group. This

yields an unadjusted odds ratio of 1.5 for the treatment effect. A sample size of 1,000 (500 subjects per group) is estimated in order to obtain 90% power to detect this unadjusted effect at a 5% significance level. This is similar to the assumptions specified in the IMS III trial design (Broderick, 2013). For each scenario described below, 10,000 trials are simulated using Monte Carlo simulation method. The operating characteristics – type I error, power and bias with respect to the treatment effect – are compared between simple randomization and covariate-adaptive randomization.

In each trial, there are three variables of primary interest: Z is a dichotomous treatment assignment, X is a prognostic covariate which is dichotomous and subject to misclassification, and Y is the outcome of interest. We assume Z is perfectly measured (i.e., recorded without error). We also assume that the misclassification error of X is non-differential with respect to the treatment and the covariate itself, i.e. the misclassification probabilities are the same for both treatment arms as well as at both levels of the covariate. No interaction between the treatment effect and the covariate is considered. Dichotomized variables X and Z are generated from a Bernoulli distribution with $p = 0.5$, i.e. $Prob(X = 1) = Prob(X = 0) = Prob(Z = 1) = Prob(Z = 0) = 0.5$. A misclassified version of X, denoted as G, is also generated, with the same misclassification rates for each level of X (i.e. $Prob(G = 1|X = 0) = Prob(G = 0|X = 1)$), varying from 0 to 40%. Under simple randomization, X (as well as G) and Z are generated independently, while under covariate-adaptive randomization, Z is generated within each level of G, the misclassified version of the covariate. Both permuted block randomization and biased-coin randomization are incorporated for covariate-adaptive randomization. Based on published recommendations, the block size for permuted block randomization is set at 4 (Efron, 1971; Matts and Lachin, 1988) and the probability assigned to the biased-coin is 0.85 (Smith, 1984; Zelen, 1974).

We investigate two different types of outcome, binary and count data, both of which are analyzed using generalized linear regression models. For the dichotomous outcome, the response variable Y is generated from a Bernoulli distribution; when the outcome represents count data, the response variable Y is generated from a Poisson distribution. In both cases, the linear combination of X(the correct version of the covariate), Z(treatment assignment), and their corresponding prespecified

coefficients are used to define the distributional parameters: the probability for the Bernoulli distribution and the rate for the Poisson distribution. For both outcomes, the beta coefficient for Z is fixed at 0.405 under all scenarios. This is equivalent to either an unadjusted OR of 1.5 or rate ratio of 1.5 between treatment groups. For logistic regression, the coefficient for X also varies from -3 to +3, which results in a wide range of X effects on Y on the scale of odds ratio (ranged from 0.064 to 23.2). For Poisson log-linear regression, the coefficient for X is fixed at 1.6 (rate ratio = 4.95) with varying misclassification rate.

Hypothesis testing for the effect of treatment Z is based on three analysis approaches (described below) applied to the simulated data. All models include treatment Z. Model (5) is the true model, adjusting for the corrected version of the covariate X. Model (6) is the misclassified model, adjusting for the misclassified covariate G. Model (7) is the unadjusted model including only treatment Z:

$$g_1(p) = \beta_{x0} + \beta_{xz}Z + \beta_x X \tag{5}$$

$$g_2(p) = \beta_{g0} + \beta_{gz}Z + \beta_g G \tag{6}$$

$$g_3(p) = \beta_0 + \beta_z Z \tag{7}$$

In each model, $g(p)$ is the link function – logit when Y is binary or log when Y is count; p is the parameter for the distribution, either $Prob(Y = 1|\cdot)$ or $Pois(\lambda|\cdot)$. For count data, in addition to the Poisson log-linear regression, quasi-Poisson and negative binomial regression are also considered given the potential in practice for over-dispersion.

2.4 Simulation study

Figures 2 – 8 show the results under simple randomization (Panel A) and that under covariate-adaptive randomization (Panel B). We only present the results under stratified permuted block randomization for covariate-adaptive randomization, because the results are similar under stratified permuted block randomization and stratified biased coin randomization. Supplemental Tables 3 –

5 provide detailed numerical examples with pre-specified coefficients describing and comparing the change pattern for type I error, power and bias. Throughout the simulation, the bias is defined as the difference between the estimated beta coefficient and 0.405, the true parameter value.

2.4.1 Dichotomized outcome: logistic regression with logit link

The operating characteristics for each model are presented as follows: the true model (Model (5)) is represented in red, the misclassified or naive model (Model (6)) is represented in blue, and the unadjusted model (Model (7)) is represented in black.

2.4.1.1 Under simple randomization Under simple randomization, where covariate information is not adjusted for during the randomization process, misclassification impacts the hypothesis testing and parameter estimation only through covariate adjustment in the analysis. The type I error is maintained in all three models (Figure 2, Panel A). The true model has the smallest power loss, while the unadjusted model has maximum power loss (Figure 3, Panel A). The power loss of the misclassified model depends on the misclassification rate, as well as the magnitude of the effect of covariate X on the outcome Y. With increasing rate of misclassification or increasing magnitude of the effect of covariate X, the amount of power loss increases. As expected based on the literature (John, 1993; Gail et al., 1984), the estimated beta coefficient of treatment Z is biased towards the null in the unadjusted model, while the estimate in the true model is unbiased (Figure 4, Panel A). The direction of the bias caused by misclassification in the misclassified model is the same as that caused by failing to adjust, and is towards the null. The magnitude of this bias in the adjusted misclassified model increases with increasing misclassification rates as well as with increasing absolute magnitude of β_x, the covariate effect.

2.4.1.2 Under covariate-adaptive randomization Under covariate-adaptive randomization, information from covariate G is adjusted for during the randomization. Either X or G is adjusted for in the final analytic model. Hypothesis testing for the treatment effect is conservative using

the unadjusted model, while both the misclassified model and the true model preserve the nominal type I error (Figure 2, Panel B). With increasing misclassification rate, the type I error of the unadjusted model approaches the nominal level.

The impact of covariate misclassification on power under covariate-adaptive randomization is the same as that under simple randomization (Figure 3) with respect to both the pattern and magnitude of the impact. The true models have the smallest power loss under all scenarios. The magnitude of power loss is not trivial using the unadjusted model when the covariate effect is large relative to the treatment effect. The power loss caused by adjusting for the misclassified covariate depends on both the magnitude of the covariate effect as well as the misclassification rate, similar to what is observed for simple randomization. For example, when $\beta_x = -2.0$ and $\beta_z = 0.405$, the power of the true model is about 80%, while the power of the unadjusted model is only 70%, although the study is designed to have 90% power. The power for the misclassified model is in between that of the true model and the unadjusted model. A direct comparison between simple randomization and covariate-adaptive randomization can be seen in the supplementary document (Figure **??**). The resulting power for the unadjusted/misclassified/true model under covariate-adaptive randomization is very close to each of those under simple randomization.

Also, the pattern in the changes in the bias under covariate-adaptive randomization was similar to that under simple randomization (Figure 4). The maximum bias was observed in the unadjusted model; with approximately 10% misclassification, the bias observed in the misclassified model was about a half of what was observed in the unadjusted model. At higher rates of misclassification (30% or higher), adjusting for a misclassified covariate was similar to not adjusting for the covariate in terms of bias in the estimated treatment effect.

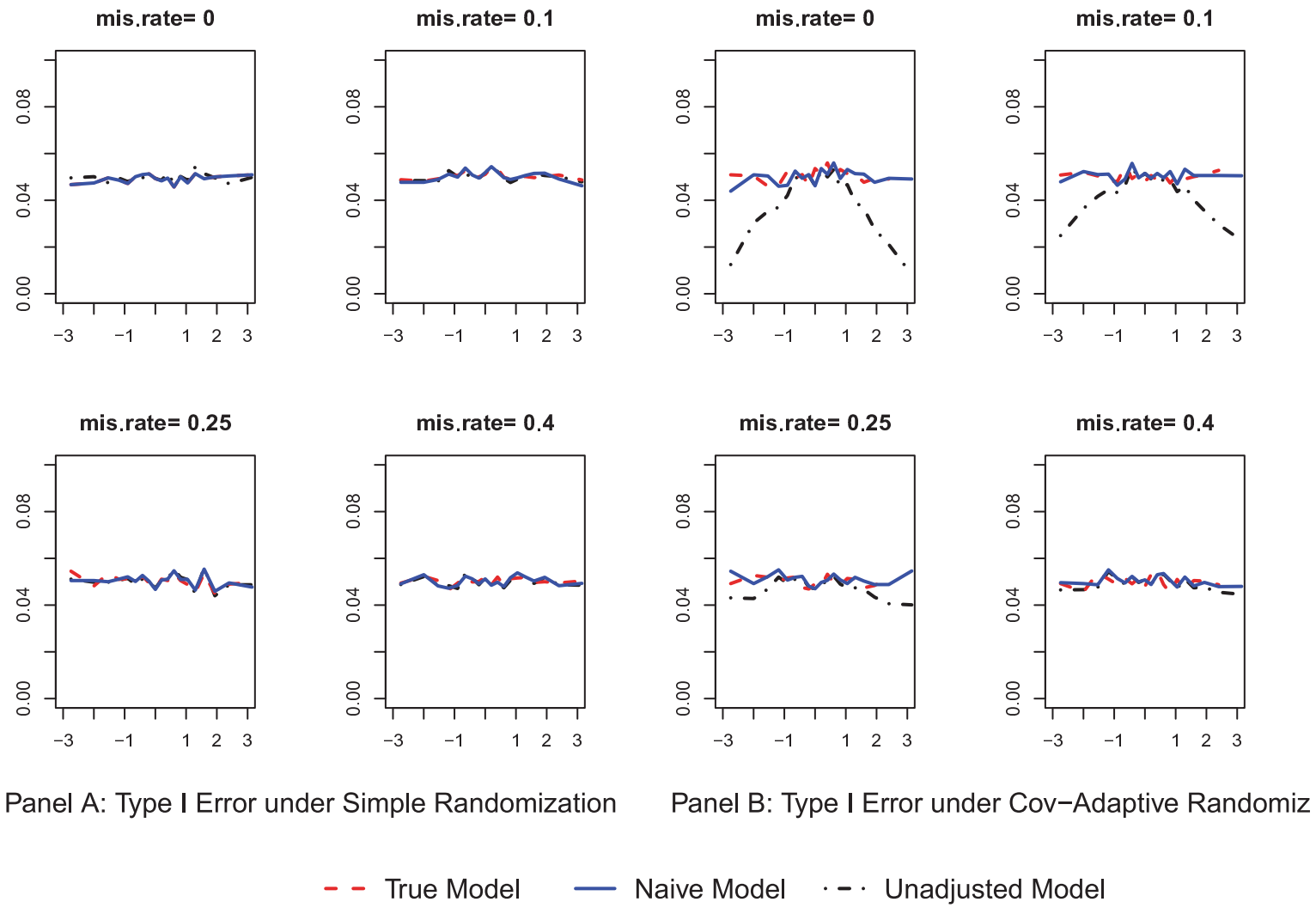

Figure 2: Change pattern of Type I error under simple randomization (Panel A) and under covariate-adaptive randomization (Panel B).

Each plot represents a different misclassification rate of 0, 0.10, 0.25, and 0.4. The X-axis represents the beta coefficients of the covariate, which ranged from -3 to +3.

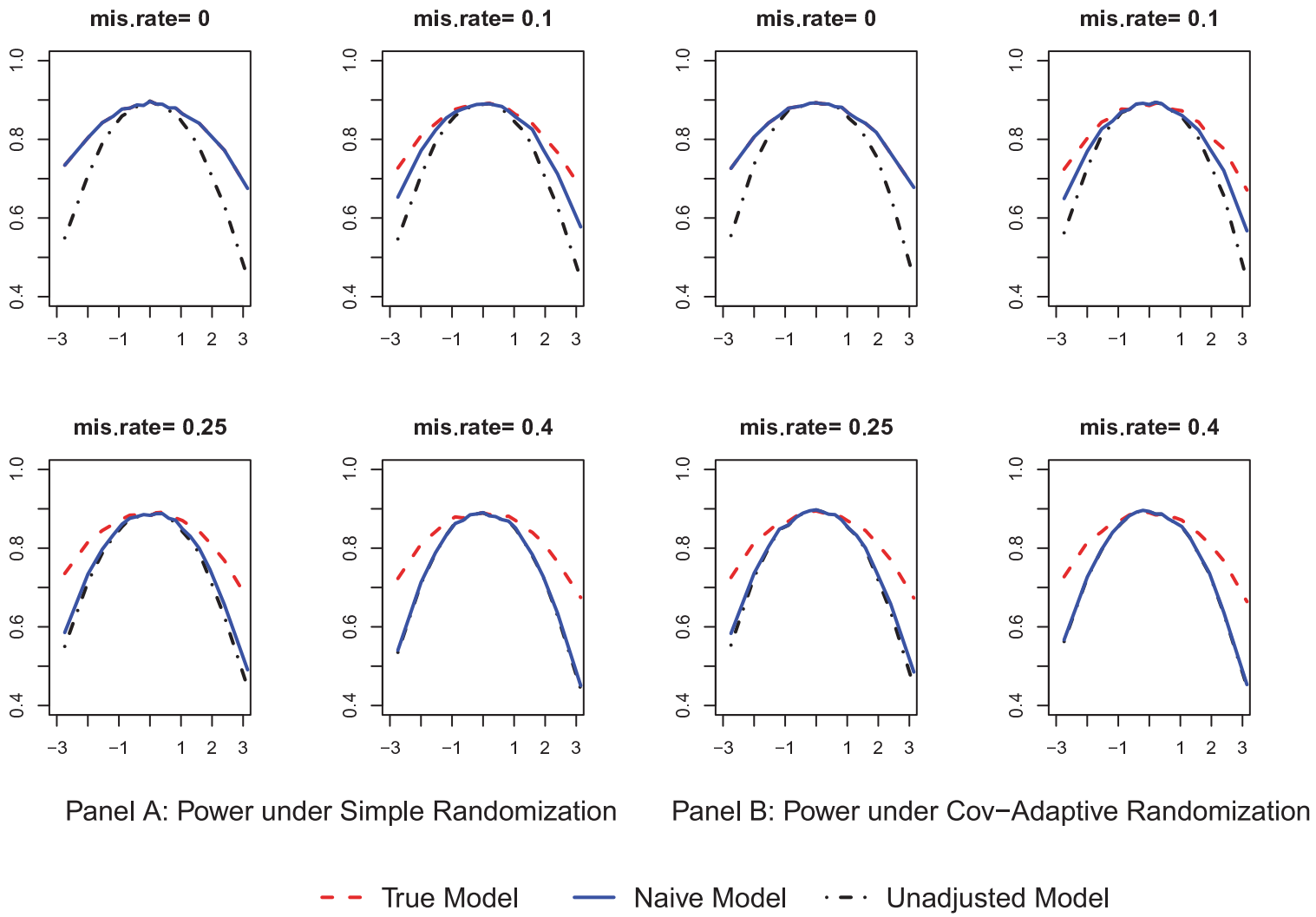

Figure 3: Change pattern of power under simple randomization (Panel A) and under covariate-adaptive randomization (Panel B).

Each plot represents a different misclassification rate of 0, 0.10, 0.25, and 0.4. The X-axis represents the beta coefficients of the covariate, which ranged from -3 to +3.

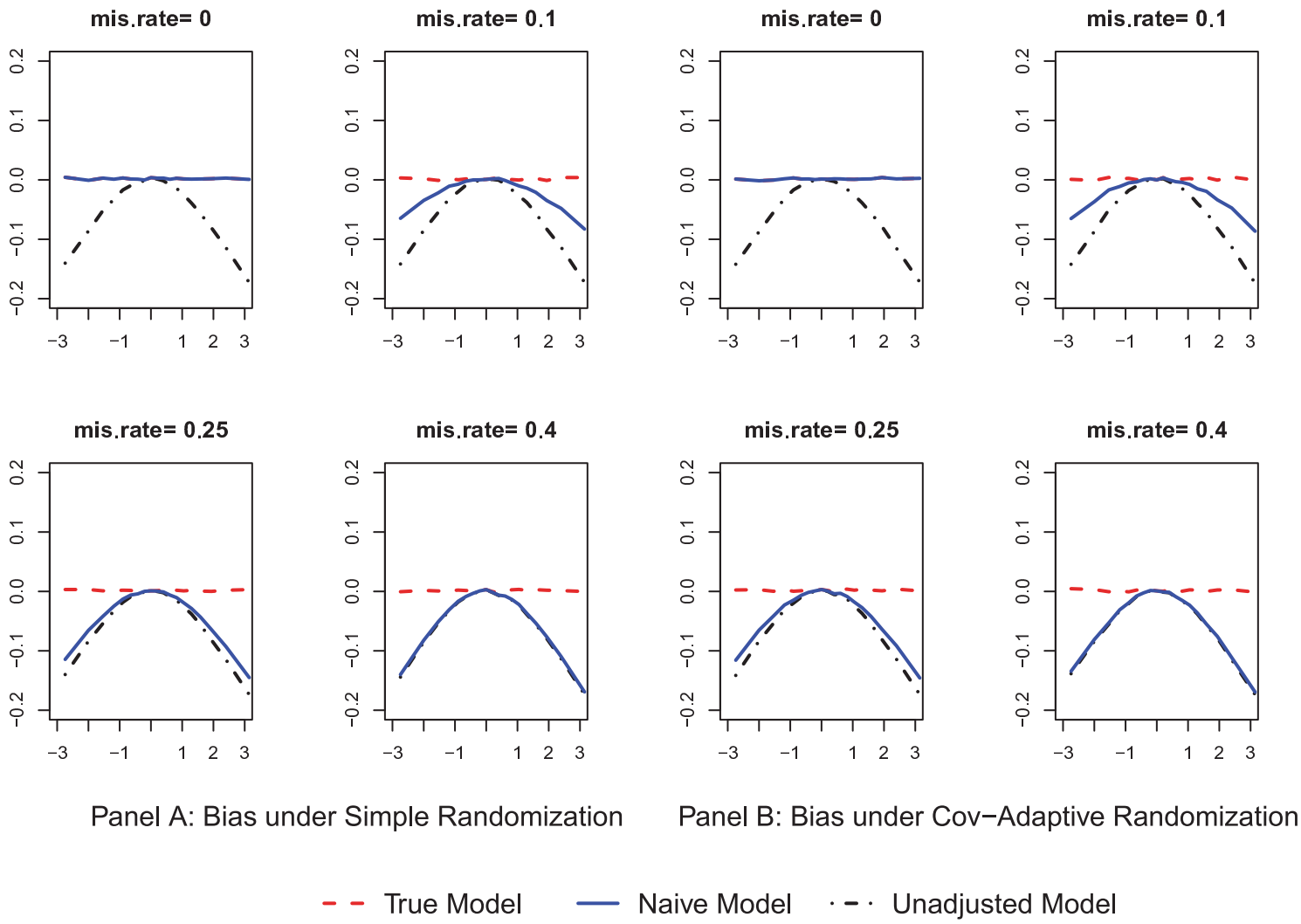

Figure 4: Change pattern of bias under simple randomization (Panel A) and under covariate-adaptive randomization (Panel B).

Each plot represents a different misclassification rate of 0, 0.10, 0.25, and 0.4. The X-axis represents the beta coefficients of the covariate, which ranged from -3 to +3.

2.4.2 Count of events as outcome: Poisson, quasi-Poisson and negative binomial regression with log link.

For the simulated count data, we fit 3 sets of log-linear regression models: Poisson regression, quasi-Poisson regression and negative binomial regression. Each set of regression models includes an unadjusted model, a model adjusted for the misclassified covariate and a model adjusted for the corrected covariate.

With Poisson regression (Figure 6 – 8, solid lines), under simple randomization the type I error (Figure 6, Panel A) is maintained only in the true model (red line). Both the unadjusted (black line) and misclassified (blue line) models have inflated type I error probabilities, with the magnitude of inflation increasing with increasing misclassification rate in the misclassified model, and maximized in the unadjusted model. Pre-specified power (Figure 7, Panel A) is obtained using the true model only. The unadjusted model has the maximum power loss, and the power loss observed in the misclassified model increases with increasing misclassification rate. No bias (Figure 8, Panel A) is observed for the estimate of the treatment effect in any of three models. Under covariate-adaptive randomization, the true model (red line) maintained the type I error (Figure 6, Panel B) at the nominal level. In data with no misclassification, the type I error is also maintained in the unadjusted model (black line). However, with increasing misclassification rate, type I error probabilities are inflated in both the unadjusted (black line) and misclassified (blue line) models with a similar magnitude. The loss in power under covariate-adaptive randomization (Figure 7, Panel B) is similar to that under simple randomization. Bias is neglegible for all three models (Figure 8, Panel B).

Quasi-Poisson regression and negative binomial regression behave similarly to each other. To better demonstrate graphically, only results of quasi-Poisson regression are shown in Figure 6 – 8 (dashed lines). While use of the quasi-Poisson or negative binomial regression captures the over-dispersion described earlier, the target power is maintained only by the true model. In fact, failure to adjust is associated with an even larger power reduction for these over-dispersed models than for the

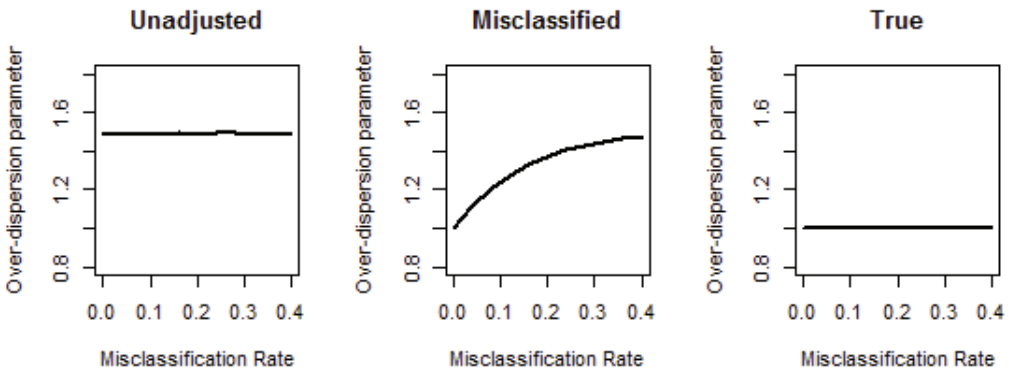

Figure 5: Estimated overdispersion parameter for quasi-Poisson model

typical Poisson regression model. As previously demonstrated, the magnitude of power loss in the misclassified models depends on the misclassification rate. Under simple randomization (Figure 6, Panel A), all models maintain the nominal type I error. However, using the unadjusted quasi-Poisson (green line) or negative binomial regression model (result not shown), more significant power loss than that of the unadjusted Poisson regression model (black line) is identified. The magnitude of power loss in the misclassified models (purple line) depends on the misclassification rate. Under covariate-adaptive randomization (Figure 6, Panel B), both of the adjusted models with either the true covariate (orange line) or the misclassified covariate (purple line) maintains type I error, while the unadjusted models are conservative. The misclassified models again demonstrate substantial power loss (almost doubled) compared to that of the misclassified Poisson model (blue solid line). The estimated dispersion parameter increases in the misclassified model (Figure 5) as well, where the estimated dispersion parameter is maximized in the unadjusted model and fixed, regardless of the misclassification rate. Again, no bias is identified in any of the models under either randomization procedure.

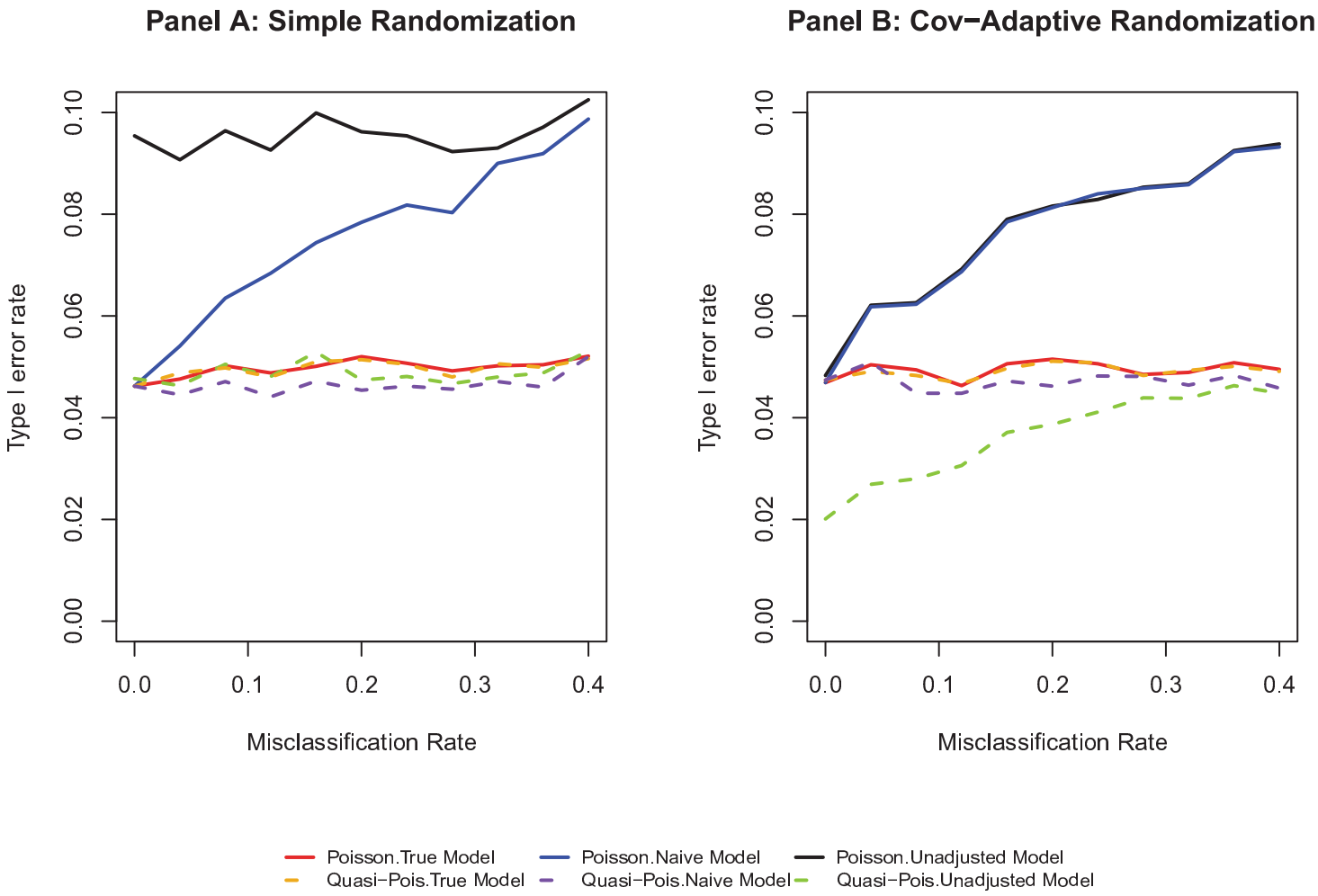

Figure 6: Change pattern of Type I error for count data under simple randomization (Panel A) and covariate-adaptive randomization (Panel B)$\beta_x = 1.6$; $\beta_z = 0.4$; $\lambda_{treat}/\lambda_{control} = 1.5$.

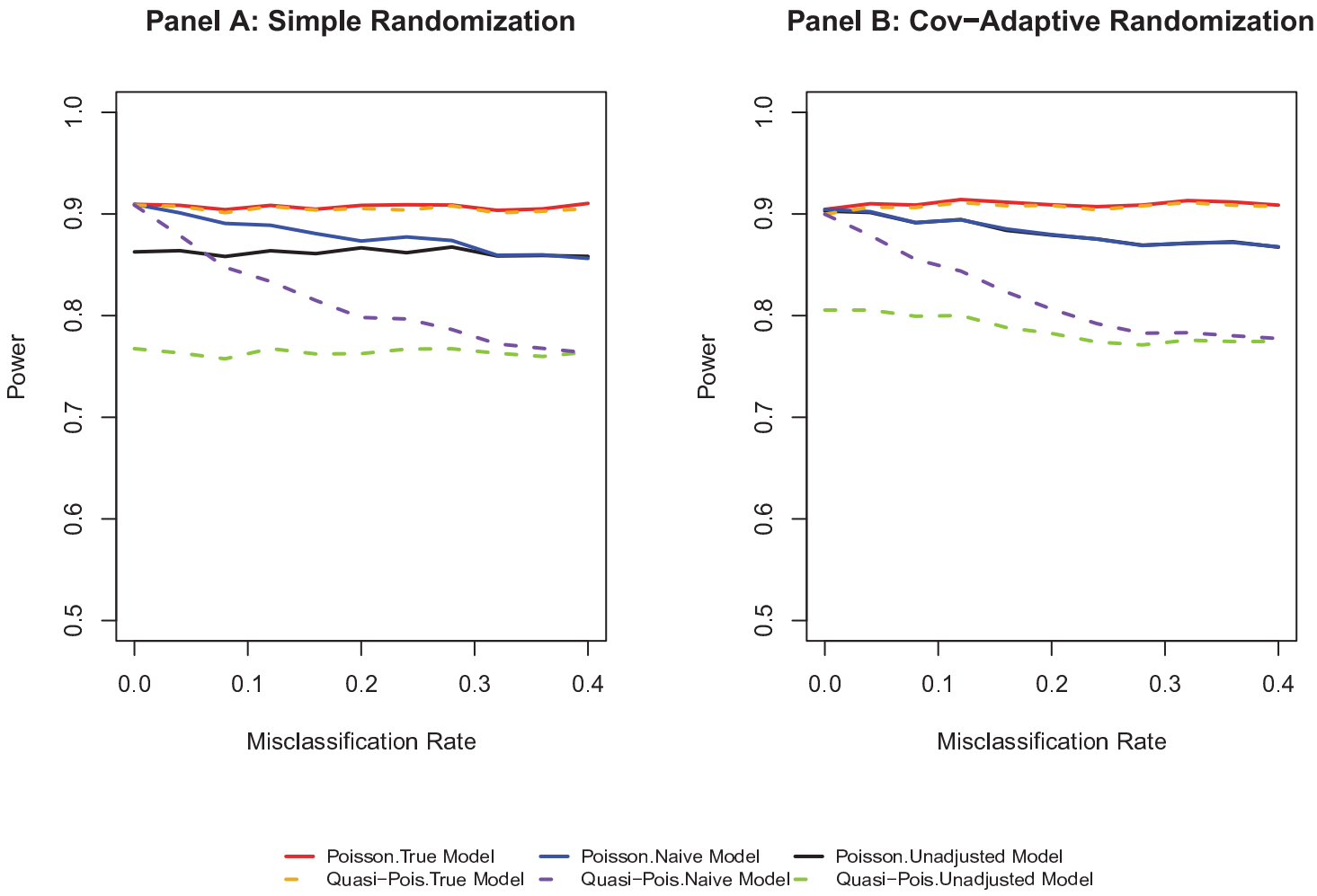

Figure 7: Change pattern of power for count data under simple randomization (Panel A) and covariate-adaptive randomization (Panel B) $\beta_x = 1.6$; $\beta_z = 0.4$; $\lambda_{treat}/\lambda_{control} = 1.5$.

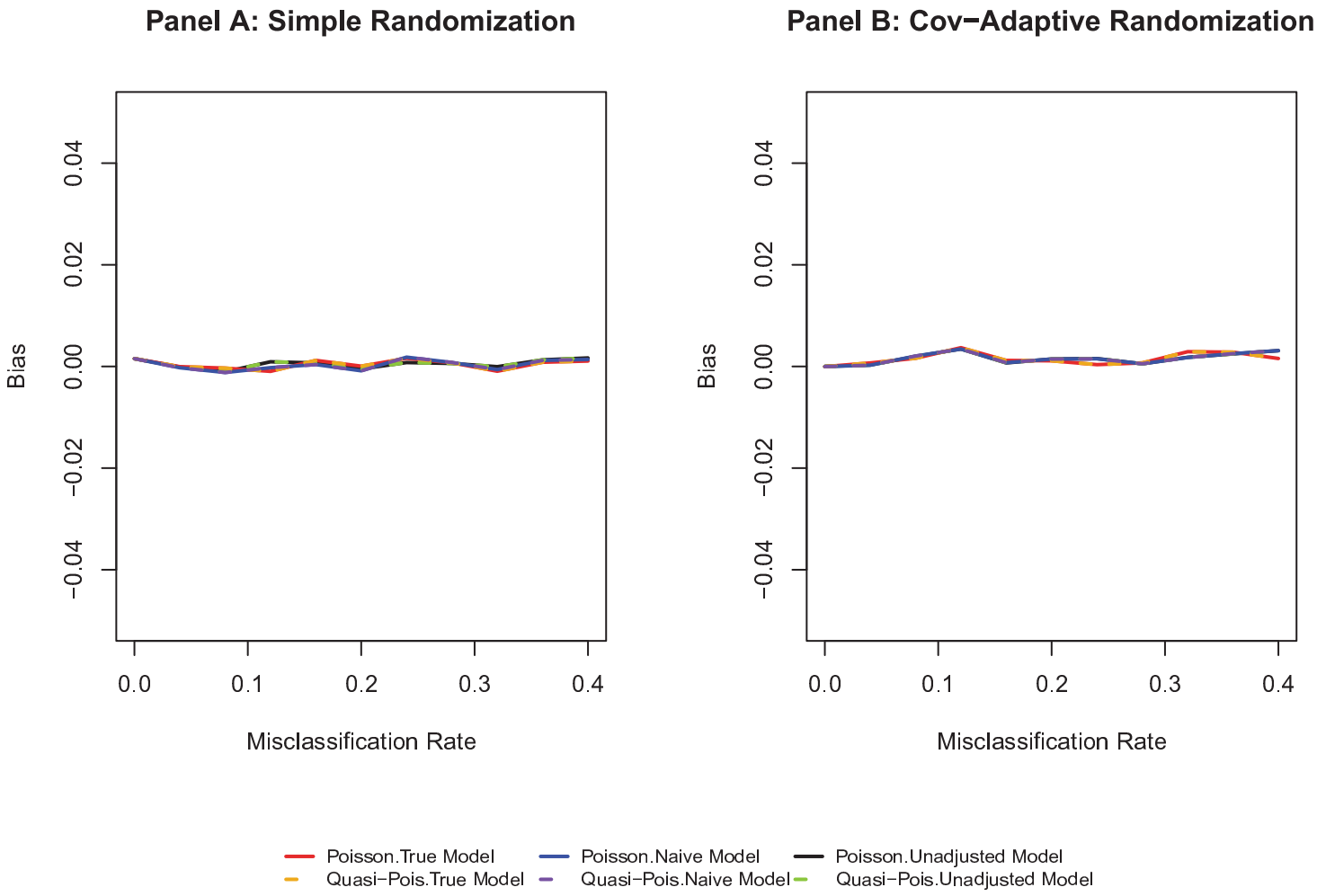

Figure 8: Change pattern of bias for count data under simple randomization (Panel A) and covariate-adaptive randomization (Panel B) $\beta_x = 1.6$; $\beta_z = 0.4$; $\lambda_{treat}/\lambda_{control} = 1.5$.

2.5 Conclusions and discussion

When the misclassification rate is high, covariate-adaptive randomization performs more like simple randomization. With increasing misclassification rate, especially when the misclassification rate is relatively large (e.g. 40%), the covariates are not informative during the covariate-adaptive randomization process. Under simple randomization, where the covariate is not involved during randomization procedure, covariate misclassification will only cause power loss and biased estimation through analytic models, and this result is consistent with the literature (Greenland, 1980; John, 1993; Robinson and Jewell, 1991). However, the results under covariate-adaptive randomization indicate that the impact of the randomization scheme depends on the underlying distribution of the outcome.

For binary outcome with logistic regression, covariate-adaptive randomization does not have an additive effect on bias and power loss caused by covariate misclassification, even when the misclassified covariate is adjusted for in the randomization procedure and in the model. While the type I error probability can be maintained through adjustment for the misclassified covariate, the resulting power is less than targeted, whereas adjustment for the corrected covariate will minimize the power loss and maintain nominal type I error probability. The bias of the estimate for the treatment effect has the same direction in the misclassified model as that in the unadjusted model, with a smaller magnitude in the misclassified model. The magnitude of the bias depends on the misclassification rate as well as the effect of the covariate on the outcome. The amount of bias and power loss is not trivial, especially when the covariate effect on the outcome is relatively large compared to the treatment effect. The operating characteristics of misclassified models are more similar to that of the unadjusted models as the misclassification rate increases. The results presented pertain to the scenario where the covariate prevalence is 50% (i.e. $Prob(X = 0) = Prob(X = 1)$). Additional simulations (results presented in the supplementary document, Figure 10) demonstrate that, for a given covariate effect and misclassification rate, the effect of the misclassification on the power and the bias is slightly lessened when the covariate prevalence is away from 50%, but the pattern remains the same. This is likely because the variability in the covariate is maximized when

the prevalence is 50%. When the prevalence is much larger than 50%, the information gained by inclusion of the covariate is reduced, as is the noise introduced by the misclassification.

For count of events with Poisson regression, on the other hand, the randomization scheme does have additional influence on power and type I error. No bias is observed using either an adjusted or unadjusted model under either randomization scheme. This is consistent with the finding by John et al. (1993) and Gail et al. (1984). Under simple randomization, the magnitude of power loss and type I error inflation is maximized in the unadjusted model regardless of covariate misclassification, but increases with increasing misclassification rates in the misclassified model. However, under covariate-adaptive randomization, the operating characteristics of the misclassified model and the unadjusted model are similar, where power loss and type I error rate inflation in both models depends on the rate of misclassification.

Due to misclassification, only part of the variability of the covariate is accounted for through the randomization procedure. With either the unadjusted model which excludes the covariate or the misclassified model, the "residual" variability due to the covariate results in over-dispersion. This over-dispersion will not be estimated correctly using Poisson regression since the estimated variance is assumed to be equal to the expected count. Thus under covariate-adaptive randomization, with Poisson regression, the variability for the outcome in an unadjusted model or misclassified model will be underestimated with the amount depending on the misclassification rate. As a result, power is reduced and type I error is inflated. Quasi-Poisson and negative binomial models are very often used for data with anticipated over-dispersion. Given the flexibility to estimate the variance differently from the expected count in these two models, the variability in the covariate, which is not sufficiently accounted for because of covariate misclassification during randomization, can be estimated through the adjusted model. However, with the unadjusted model, where no information about the covariate is included in the model, the amount of "residual" variability can't be accurately estimated. Instead, the unadjusted quasi-Poisson and negative binomial models will be estimating over-dispersion, which does not truly exist (or at least partly) since the variability (or part of the variability) of the covariate is eliminated under covariate-adaptive randomization.

The work described herein assumes that the true value of the prognostic covariate is observable, which is not always the case. The main purpose is to show how misclassification in the covariates affects the estimation of the treatment effect. The results remain relevant when the truth is not observable, although the manner in which such misclassification would be identified and corrected is left as a topic for future work. One limitation of this work is that we only considered non-differential misclassification of the covariate with respect to treatment assignment; differential misclassification may have more complicated impact on the treatment effect estimation if the treatment effect itself is expected to vary according to the covariate. In addition, we only investigated two types of covariate-adaptive randomization, other randomization algorithms may also be considered in the future.

Overall, adjustment for the covariate is always recommended in the final analysis even when the quality of the covariate is questionable. Effort should be made to identify potential misclassification and adjust for the corrected covariate in the analysis in order to minimize power loss and to more accurately estimate the treatment effect. On the other hand, given the fact that the effect of prognostic covariates on the outcome reported in the literature varies, statistical methods, and perhaps sample size reassessment, should be considered to correct the bias in estimating treatment effect if there is anticipated or observed covariate misclassification.

Acknowledgements

This work is supported by a National Institute of Neurological Disorders and Stroke (NINDS) grant, U01 NS054630, P.I.: Yuko Y. Palesch, Ph.D. and Neurological Emergencies Treatment Trials (NETT) Network grant, U01 NS059041, P.I.: Yuko Y. Palesch, Ph.D.

2.6 Supplement

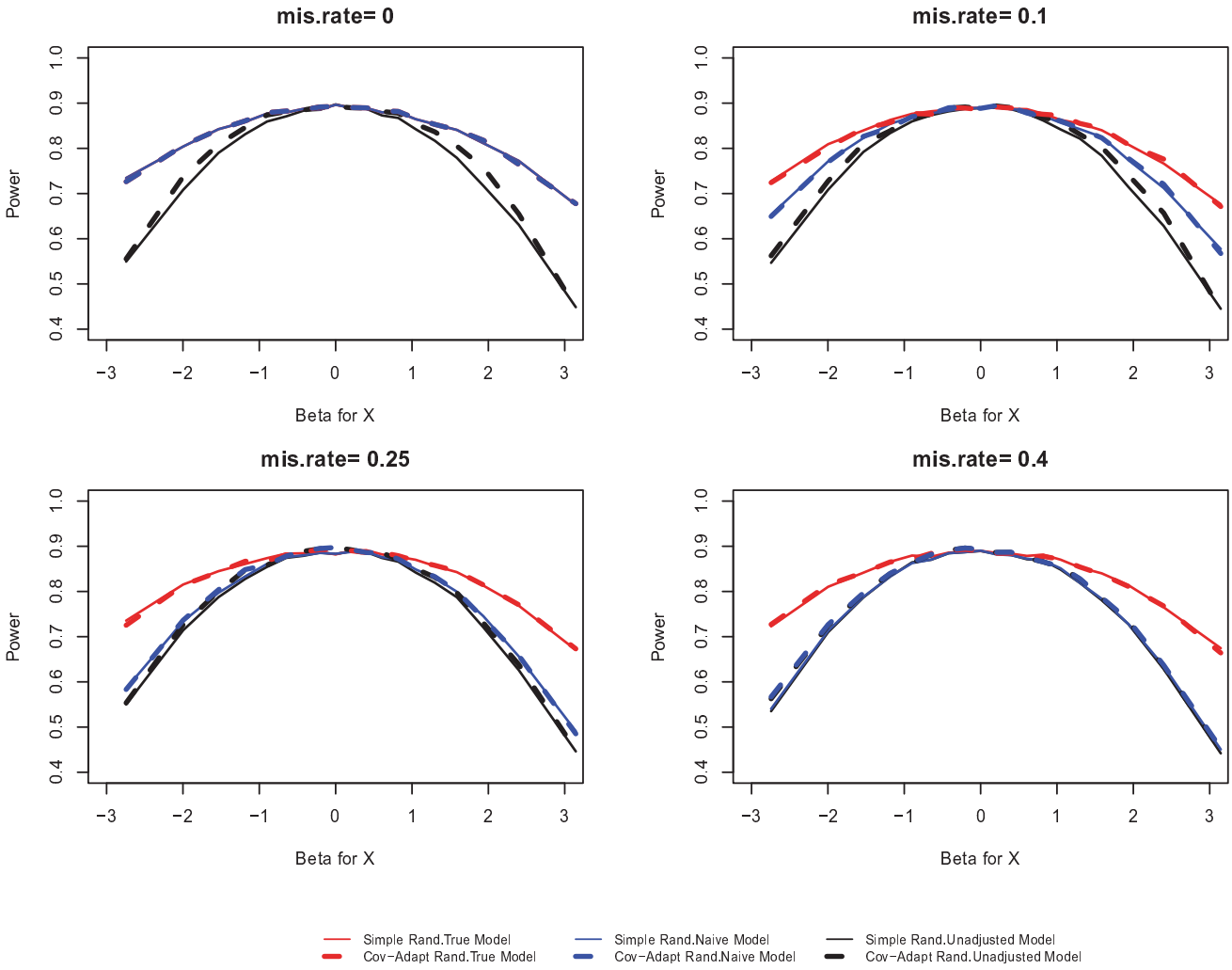

Figure 9: Direct comparison of power between simple randomization and covariate-adaptive randomization

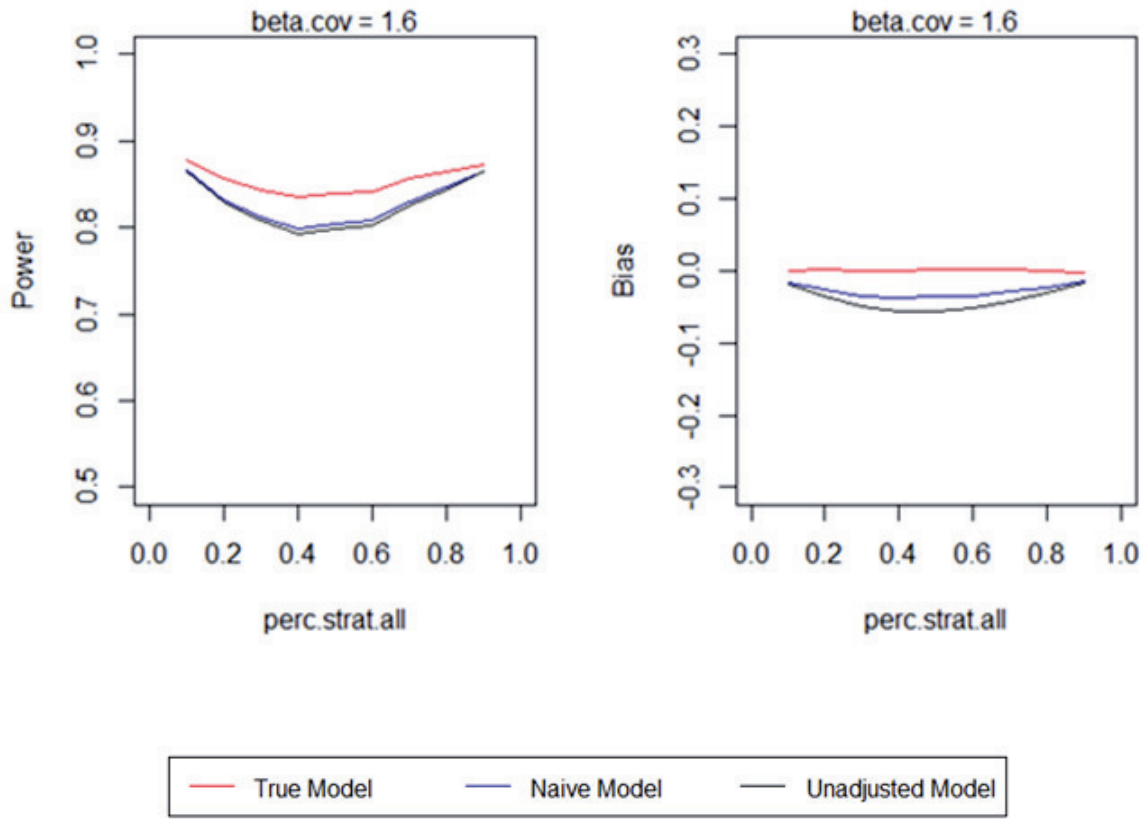

Figure 10: Impact of covariate misclassification on power and bias with varying covariate prevalence.

The misclassification probabilities are fixed at 20% for both levels of the covariate. The crude covariate effect on the outcome in terms of Odds Ratio is 4.95. The X-axis represents varying prevalence of the true covariate, i.e. $Prob(X = 1)$, from 0.1 to 0.9.

Table 3: Logistic regression model, $\beta_x = -2.0, OR_{x|z} = 0.14; \beta_z = 0.4, OR_{z|x} = 1.5$

Misclassification Rate (%)		BIAS			POWER			TYPE I ERROR		
		True	Misclassified	Unadjusted	True	Misclassified	Unadjusted	True	Misclassified	Unadjusted
0	Simple Randomization	-0.00077	-0.00077	-0.08595	0.804	0.804	0.7082	0.0474	0.0474	0.0501
	Stratified Perm. Block	-0.00153	-0.00153	-0.08711	0.8063	0.8063	0.7373	0.0509	0.0509	0.0303
	Stratified Biased Coin	0.002781	0.002781	-0.08411	0.8127	0.8127	0.7324	0.0461	0.0461	0.0275
10	Simple Randomization	0.001811	-0.03453	-0.08486	0.8093	0.7713	0.7086	0.0482	0.0477	0.0484
	Stratified Perm. Block	-0.00091	-0.03669	-0.08673	0.8028	0.7697	0.7278	0.0508	0.0523	0.0365
	Stratified Biased Coin	0.001959	-0.03303	-0.08352	0.8132	0.776	0.7294	0.0512	0.0502	0.0359
20	Simple Randomization	0.002804	-0.05681	-0.08352	0.8091	0.7494	0.7157	0.047	0.0475	0.0492
	Stratified Perm. Block	0.000973	-0.05793	-0.08459	0.8079	0.747	0.7259	0.0512	0.0515	0.0415
	Stratified Biased Coin	0.002197	-0.0572	-0.084	0.8058	0.7474	0.7197	0.0524	0.053	0.0466
30	Simple Randomization	0.002623	-0.0716	-0.08313	0.8067	0.7291	0.7128	0.0514	0.0491	0.0504
	Stratified Perm. Block	0.00025	-0.0751	-0.08658	0.8067	0.721	0.7172	0.0505	0.0492	0.0443
	Stratified Biased Coin	0.000451	-0.07347	-0.08512	0.8169	0.7342	0.7191	0.0485	0.0504	0.0469
40	Simple Randomization	0.001673	-0.08213	-0.08508	0.8104	0.7132	0.7101	0.0523	0.053	0.0522
	Stratified Perm. Block	0.003136	-0.08087	-0.08392	0.8149	0.7268	0.7249	0.0492	0.0492	0.0466
	Stratified Biased Coin	0.001022	-0.0824	-0.08546	0.8101	0.718	0.7134	0.0498	0.05	0.0484

Table 4: Logistic regression model, $\beta_x = 1.59, OR_{x|z} = 4.90; \beta_z = 0.4, OR_{z|x} = 1.5$

Misclassification Rate (%)		BIAS			POWER			TYPE I ERROR		
		True	Misclassified	Unadjusted	True	Misclassified	Unadjusted	True	Misclassified	Unadjusted
0	Simple Randomization	0.001484	0.001484	-0.05687	0.8408	0.8408	0.7791	0.0492	0.0492	0.0514
	Stratified Perm. Block	0.001612	0.001612	-0.05628	0.8407	0.8407	0.8052	0.0512	0.0512	0.0365
	Stratified Biased Coin	0.001753	0.001753	-0.05623	0.8417	0.8417	0.796	0.0484	0.0484	0.0354
10	Simple Randomization	0.002423	-0.02099	-0.05593	0.84	0.8252	0.7836	0.0497	0.0515	0.0513
	Stratified Perm. Block	0.003954	-0.01931	-0.0546	0.8441	0.8232	0.8024	0.0491	0.0506	0.0402
	Stratified Biased Coin	0.002313	-0.02044	-0.05524	0.8414	0.8198	0.7933	0.0491	0.0506	0.0411
20	Simple Randomization	0.001497	-0.03735	-0.05675	0.8421	0.8024	0.7834	0.0526	0.0513	0.0511
	Stratified Perm. Block	0.000985	-0.03695	-0.05602	0.839	0.8064	0.7963	0.0489	0.0483	0.0438
	Stratified Biased Coin	0.002478	-0.03584	-0.05495	0.8476	0.8075	0.7907	0.0493	0.0487	0.0448
30	Simple Randomization	0.000064	-0.04965	-0.05805	0.838	0.7868	0.78	0.0502	0.051	0.0519
	Stratified Perm. Block	0.003387	-0.04633	-0.05475	0.8447	0.7963	0.7952	0.0476	0.0479	0.045
	Stratified Biased Coin	0.002588	-0.04748	-0.0559	0.8434	0.7954	0.7896	0.0478	0.0488	0.0465
40	Simple Randomization	0.002907	-0.05335	-0.05566	0.8401	0.7837	0.7806	0.0497	0.0503	0.0493
	Stratified Perm. Block	.000019	-0.055	-0.05733	0.8379	0.7865	0.7848	0.0465	0.0482	0.0474
	Stratified Biased Coin	0.002946	-0.05268	-0.05499	0.8416	0.785	0.783	0.053	0.0504	0.0501

Table 5: Poisson regression model,$\beta_x = 1.6, \beta_z = 0.4$, $\lambda_{treat}/\lambda_{control} = 1.5$

Misclassification Rate (%)		BIAS			POWER			TYPE I ERROR		
		True	Misclassified	Unadjusted	True	Misclassified	Unadjusted	True	Misclassified	Unadjusted
0	Simple Randomization	0.00113	0.00113	0.001165	0.9107	0.9107	0.8632	0.0456	0.0456	0.0953
	Stratified Perm. Block	0.003001	0.003001	0.003074	0.9157	0.9157	0.9159	0.0514	0.0514	0.0517
10	Simple Randomization	-0.00013	-2.65E-05	-0.00054	0.9062	0.888	0.8593	0.0499	0.066	0.096
	Stratified Perm. Block	0.00245	0.002532	0.002503	0.9116	0.8957	0.8946	0.0502	0.0685	0.0687
20	Simple Randomization	0.003529	0.003277	0.003436	0.9104	0.8816	0.8675	0.0482	0.0795	0.0998
	Stratified Perm. Block	0.001719	0.001499	0.001596	0.9124	0.8798	0.8802	0.0538	0.0804	0.0802
30	Simple Randomization	0.001738	0.001444	0.000962	0.901	0.8644	0.8594	0.0519	0.0907	0.0971
	Stratified Perm. Block	0.002117	0.00231	0.002309	0.9143	0.8694	0.8692	0.0512	0.092	0.0923
40	Simple Randomization	0.000659	-0.00049	-0.00044	0.9133	0.8657	0.8649	0.0478	0.0932	0.0962
	Stratified Perm. Block	0.000744	0.000798	0.000794	0.9072	0.8639	0.8639	0.0496	0.0962	0.0961

3 Original Manuscript 2

Title: Using Continuous-Time Hidden Markov Model to Estimate the Misclassification Probabilities: the Performance and Model Diagnostic Issue.

Authors: Liqiong Fan(a) ,Sharon D. Yeatts (a), Bethany J. Wolf (a), Leslie A. McClure (b) , Magdy Selim (c), Yuko Y. Palesch (a)

Affiliation:

(a) Department of Public Health Sciences, Medical University of South Carolina, Charleston, SC 29425, U.S.A.

(b) Department of Epidemiology & Biostatistics, Drexel University, Philadelphia, PA 19104, U.S.A.

(c) Beth Israel Deaconess Medical Center, Department of Neurology, Division of Cerebrovascular Diseases, Boston MA 02215, USA.

Submission Status: drafted

Abstract

In medical research, many diseases are characterized by discrete states although the transition between the states is continuous. The evaluation of the disease states are sometimes subject to misclassification due to either patient's characteristics which interfere with the accuracy of the assessment or the sensitivity of diagnostic tests. In addition, patient's evaluation schedule is usually pre-specified resulting in panel-observed data. Applications of the continuous time Hidden Markov Model (CTHMM) for data with misclassification have been proposed in the literature. However, model performance, in terms of estimation and model diagnosis, is not well described. We provide a simulation study to demonstrate the performance of the model with a focus on misclassification. We also propose an alternative AIC based approach derived according to modified full data likelihood to directly compare CTHMM and basic Markov Model in order to identify the misclassification error in the data. Simulation scenarios were generated with varying parameters including transition intensities, misclassification probabilities, number

of observations per subject, and sample size. Results based on both basic Markov Model (MM) and CTHMM were compared. The performance of our proposed AIC based approach was also compared to the Pearson type goodness-of-fit test in terms of identifying a correct model for the data. For a two-state CTHMM, the performance of the model in terms of point estimate and model goodness-of-fit highly depends on the size of transition intensity relative to the observation schedule and the amount of misclassification in the data. Increasing total sample size, increasing number of observations per subject, and increasing variability in the observation schedule improves the performance. The CTHMM can be used to estimate the misclassification probabilities in the data. The proposed AIC based approach can identify the correct model for data with or without misclassification.

Keywords:CTHMM, misclassification, AIC, model diagnosis

3.1 Introduction

In medical research, a patient's condition is often categorized into one of several finite disease states (eg, mild, moderate, or severe) based on lab results, clinical assessments, or physical examinations. However, such evaluation is sometimes easily confounded and may result in misclassification due to either patient or examiner characteristics that interfere with the accuracy of the assessment or the sensitivity of the diagnostic test used for evaluation. Naive analysis of the data assuming no misclassification error will lead to biased estimates, resulting in misleading conclusions (Fan et al., 2015). Multistate models (MSM) have been widely used for longitudinal categorical data and are useful tools to model disease dynamics when the disease is characterized into states. However, the transition of the disease from one state to another may not be observed exactly. In clinical trials, the timing of follow-up visits is often determined by the protocol rather than by the occurrence of an event; this results in panel-observed data, where the transition occurs between the current assessment and the previous assessment. A continuous-time hidden Markov model (CTHMM) relates the true disease condition to the transition of the latent status, which follows a continuous-time Markov chain (Bureau et al., 2003; Jackson et al., 2003). Conditional on the underlying true disease status, the observed data can be characterized via the emission probabilities. The structure of the CTHMM provides a convenient framework to simultaneously estimate the disease dynamics and the potential probability of data misclassification. Such applications have been proposed in the literature (Bureau et al., 2003; Jackson et al., 2003; Jackson and Sharples, 2002) and applied in different disease areas including cognitive decline and dementia (Marioni et al., 2012; Norton et al., 2013; Buter et al., 2008), HIV/HPV (Human Papilloma Virus) infection (Blitz et al., 2013), as well as liver cirrhosis and hepatitis C (Bartolomeo et al., 2011; Terrault et al., 2008; Sweeting et al., 2006).

Despite the use of the CTHMM in clinical applications, the evaluation of the CTHMM's performance in the case of misclassification is limited. Rosychuk et al. (Rosychuk and Thompson, 2003; Rosychuk and Islam, 2009; Rosychuk RJ, 2004) discussed the impact of misclassification on the accuracy of transition probability estimates for an alternating binary Markov disease process. How-

ever, they only investigated equally spaced observations with small misclassification. Other authors (Leroux, 1992; Petrie, 1969; Baum and Petrie, 1966) have studied the identifiability issue for hidden Markov model without addressing the impact of misclassification on the performance. In addition, as recommended in the literature (Bureau et al., 2003), comparing the estimated transition intensity with and without the latent structure is a useful initial step to evaluate the goodness-of-fit of the hidden Markov model. Aguirre-Hernandez and Farewell (Aguirre-Hernandez and Farewell, 2002) originally proposed using the Pearson-type goodness-of-fit test to evaluate the model fit, and Titman and Sharples (Titman and Sharples, 2008) extended it to the case of misclassification. However, the performance of the test has not been well studied and may potentially suffer from power due to grouping the time intervals and from other model assumptions (Titman and Sharples, 2008).

The purpose of this paper is to investigate the performance of the CTHMM with a focus on misclassification with respect to both the point estimates and model goodness-of-fit. We expand Rosychunk et al.'s simulation to more sophisticated scenarios in order to fully examine the performance of CTHMM with a focus on misclassification. We also propose an AIC based approach based on the modified full data likelihood to directly compare Markov models with and without the latent structures (i.e. the hidden Markov vs. basic Markov model) and compare the performance to that of the modified Pearson-type goodness-of-fit test. The rest of the paper will be organized as follows. In section 2, the CTHMM will be introduced in detail, including assumptions, notation, the algorithm for estimation, and model diagnostics. The simulation used to evaluate the performance of the model will be described and the results will be presented in section 3. Section 4 will introduce the proposed modified full data likelihood and AIC, and demonstrate its performance through Monte Carlo simulation. Finally, we will close the paper with discussion and recommendations based on the simulation results.

3.2 Continuous-time hidden Markov model (CTHMM)

3.2.1 Model setup and likelihood for CTHMM

Consider a two-state CTHMM for alternating binary longitudinal outcomes, as described by (Bureau et al., 2003; Jackson et al., 2003; Rosychuk and Thompson, 2003). Let $O = \{O^{(t_1)}, O^{(t_2)}, \ldots, O^{(t_m)}\}$ represent m observed states with state space $\{0, 1\}$ for individual subject. The observed states: $O^{(t_1)}, O^{(t_{t2})}, \ldots, O^{(t_m)}$ are independent given their underlying true states, denoted as $S = \{S^{(t_1)}, S^{(t_2)}, \ldots, S^{(t_m)}\}$ where the underlying latent states $\boldsymbol{S}$ are assumed to follow a first-order continuous-time Markov process with the same state space $\{0, 1\}$. That is, given the current state at time $t_m(S^{(t_m)})$, the future state $(S^{(t_{m+1})})$ is independent of the past state $(S^{(T<t_m)})$. Let $\Delta t_m = t_{m+1} - t_m$ denotes the time interval between the current and the future observation. The continuous-time indicates that the probability of transition out of the current state depends on the time (Δt_m) spent in that state. Then $P_{ij}\{\Delta t_m\}$, the probability of transitioning from state i at time t_m to state j at time t_{m+1} with interval time Δt_m, can be expressed as:

$$P_{ij}\{\Delta t_m\} = P\{S^{(t_{m+1})} = j | S^{(t_m)} = i, S^{(T<t_m)} = s\} = P\{S^{(t_{m+1})} = j | S^{(t_m)} = i\}$$

The latent process is described by the instantaneous transition rate from one state to another, referred to as transition intensities. For a homogeneous continuous-time Markov process, these intensities are the same over time. The transition intensities can be expressed in a matrix form Λ with rows representing current states and columns representing future states. That is:

$$\Lambda = \begin{pmatrix} \lambda_{0,0} & \lambda_{0,1} \\ \lambda_{1,0} & \lambda_{1,1} \end{pmatrix}$$

where $\lambda_{ij} = \lim_{\Delta t \to 0} \frac{prob\{S^{(t_{m+1})}=j | S^{(t_m)}=l\}}{\Delta t}$, and the mean time spent on a state before moving out is referred to as the mean sojourn time, which can then be calculated as $1/\lambda_{01}$ and $1/\lambda_{10}$ for state 0 and 1 respectively. The transition probability between two time points with interval time Δt_m,

given the state status, is the corresponding entry in the exponentiation of the transition intensity matrix

$$P(\Delta t_m) = e^{\Delta t_m \Lambda}$$

where the $(i,j)^{th}$ entry represents the probability of transitioning from state i to state j after interval time Δt_m. For the two-state case, the transition probability can be calculated analytically as,

$$P_{01}(\Delta t_m) = P(S^{(t_{m+1})} = 1 | S^{(t_m)} = 0) = \frac{\lambda_{01}}{\lambda_{01}+\lambda_{10}}\{1 - e^{-\Delta t_m(\lambda_{01}+\lambda_{10})}\}$$

$$P_{10}(\Delta t_m) = P(S^{(t_{m+1})} = 0 | S^{(t_m)} = 1) = \frac{\lambda_{10}}{\lambda_{01}+\lambda_{10}}\{1 - e^{-\Delta t_m(\lambda_{01}+\lambda_{10})}\}$$

The emission probabilities are the conditional probabilities of the observed states given the latent states, i.e. $\pi_{o|s} = p(O = i|S = j), i, j = 0, 1$ with the initial distribution probabilities for the latent states $\pi(s) = p(S^{(t_1)} = s), s = 0, 1$. Based on all the above parameters, the likelihood of the observed data can be expressed as:

$$\begin{aligned} L(\mathbf{O}) &= L(O^{(t_1)}, O^{(t_2)}, \ldots, O^{(t_m)}) \\ &= \sum_{s^{(t_1)}} \sum_{s^{(t_2)}} \cdots \sum_{s^{(t_m)}} \pi(s^{(t_1)}) \prod_{m=1}^{n-1} P_{s^{(t_m)} s^{(t_{m+1})}}\{t_{m+1}, t_m | S = s^{(t_m)}\} \prod_{m=1}^{n} \pi_{o|s} \end{aligned}$$

The latent states $S^{(t_1)}, S^{(t_2)}, \ldots, S^{(t_m)}$ are the unobservable true disease severity that a subject belongs to at time $(t_1, t_2, \ldots, t_m)$. The observed states $O^{(t_1)}, O^{(t_2)}, \ldots, O^{(t_m)}$ represent the disease severity to which the subject is assigned and is subject to misclassification error. Then the emission probabilities $\pi_{o|s}$ can be considered as the classification probabilities, which are assumed to be the same over time in this study.

To solve the likelihood, the expectation-maximization (EM) algorithm can be applied with the weights in the E step calculated via the forward-backward algorithm developed by Baum et al.

(1970). One typical issue for solving the likelihood using the EM algorithm is that the EM algorithm may converge to a local maximum or a stationary point instead of the global maximum. Therefore, it is important to use different starting values to solve the likelihood. An (2013) proposed a starting value based on three consecutive observations to improve the convergence. As an alternative, Lange and Minin (2013) developed a robust and efficient expectation-maximization algorithm based on a complete data likelihood for parameter estimation.

3.2.2 Model Goodness-of-fit diagnostic tools

As described in the previous section, there are three key assumptions of the CTHMM: (i) conditional on the current state, the future and the past are independent, i.e. the Markov property of the underlying states; (ii) for a time-homogeneous Markov model, the transition intensities are constant over time; (iii) the misclassification (emission) probabilities do not vary by time. It is also essential to evaluate if this additional latent structure is necessary for the data compared to a basic Markov model. Most of the diagnostic approaches in the literature for the misclassification CTHMM are graphical. In general, the modified Pearson type goodness-of-fit test (Titman and Sharples, 2008) could be used for either the basic Markov model or the CTHMM to evaluate the overall model fit. Here we briefly review the diagnostic tools.

State-change plots compare the empirical Kaplan-Meier estimates of the transition between a specific pair of states to those predicted by the fitted model (Bureau et al., 2003; Titman and Sharples, 2010). Although the construction of the plots requires simulation to get the empirical estimates, they provide informal but important information to visually assess the model's overall goodness-of-fit. Other methods of informal assessment include generalization of prevalence counts and contingency tables for prediction of future observations (Titman and Sharples, 2008). The generalized contingency table is more appropriate for unequally-spaced observations, in that the counts are averaged over time periods instead of at a specific time point. Bureau et al. (2003) also recommends comparing the estimated transition intensities from a fitted CTHMM to those from the basic Markov model without the latent structure in order to determine whether or not the additional

latent structure provides a better fit to the data. Meanwhile, the estimated misclassification probabilities can be compared to those in the existing literature and practical experience, if available.

A modified Pearson-type goodness-of-fit test by Titman and Sharples (2008) could be applied to formally evaluate the model fit in the case of the misclassification CTHMM. The observations are grouped by assessment number, time intervals, covariate categories, and quantiles of the estimated transition intensities. The test can also accommodate the situation where there is an absorbing state with a known exact entry time such as death in the data. Parametric bootstrapping is needed for the test to estimate the asymptotic null distribution.

Finally, the invariance of misclassification probabilities over time can be examined to some extent by both state change plots and the Pearson-type goodness-of-fit test in terms of overall model goodness-of-fit. The likelihood ratio test between models with and without a time-varying covariate of the misclassification probabilities can also be constructed for this purpose (Titman and Sharples, 2010).

3.3 Monte Carlo Simulation

3.3.1 Simulation setup

The simulation is set up as follows: for a given subject, a hidden continuous-time Markov process which represents the true disease dynamics is simulated first with pre-specified transition intensity matrix (Λ) assuming a homogeneous Markov chain and fixed, unequally-spaced observation schedule. Conditioning on the true disease status, the observed states can then be generated with given misclassification rate (MCR) $\pi_{o|s}$. Varying parameters include size of transition intensities, the magnitude of misclassification in the data, number of observations per subject, and total number of subjects in the study. We consider three different sizes of transition (Λ): small, medium and large, resulting in long, medium-long and short mean sojourn times. Six different misclassification rates are simulated ranging from no misclassification to 40% misclassification. Without loss of generality,

we assume the misclassification probabilities are the same at both levels of the disease status in the simulation, i.e.$\pi_{o|s} = P(O = 1|S = 0) = P(O = 0|S = 1)$. For each combination of Λ and $\pi_{o|s}$, 5 and 10 unequally spaced observations per subject are generated with constraints that the same first 5 observations are also included in the 10-observation scenarios. Thus an unequally-spaced panel-observed timeframe is set up with comparable scenarios regarding the number of observations. The observation schedule is fixed at day 1, 3, 5, 8, and 12 for the 5-observation-time-points scenario and additionally at day 13, 16, 20, 26, and 28 for the 10-observation-time-points scenario. Compared to the three sizes of the transition, these observation schedules correspond to less than 1/3, around 0.5, and about the same length of the mean sojourn times. Moderate to large size trials, with sample size equal to 500, are considered in the simulation for all combinations of the above parameters. A smaller sample size equal to 100 is also examined in some scenarios. Detailed parameter values are given in Table 6. The initial distribution is the distribution of the true states at 1^{st} time point: i.e.$P(S^{(t_1)} = 1) = 0.4$.

Table 6: Transition intensities and misclassification probabilities used in the simulation

Scenario	Initial Dist'n	λ_{01}	λ_{10}	$\pi_{o\|s}$	Obs. Time Points
Small	0.4	0.04	0.1	0.05,0.2,0.25,0.3,0.4	5,10
Medium	0.4	0.2	0.15	0.05,0.2,0.25,0.3,0.4	5,10
Large	0.4	0.4	0.3	0.05,0.2,0.25,0.3,0.4	5,10

After data are generated, two types of models are fit to the simulated data with both true disease and observed disease statuses: 1) the Basic Markov model (MM), estimating only the transition intensities of the data, assuming no misclassification; 2) the CTHMM, estimating the transition intensities and misclassification probabilities of the data simultaneously. For each scenario described above, 500 trials are simulated using Monte Carlo simulation method. All simulations are carried out using software R version 3.0.2 with package *msm*.

For all models, a Pearson-type goodness-of-fit test statistic is calculated, and corresponding p-values are estimated. The significance level of the test is set at 0.05, meaning that a p-value less than 0.05 indicates poor fit of the model.

3.3.2 Simulation results

In all results presented below, the colors of the lines represent different transition intensities: blue for small intensities ([0.04, 0.1]), black for medium intensities ([0.2, 0.15]) and red for large ([0.4, 0.3]) intensities. Each panel represents various misclassification probabilities from 0.05, 0.2, 0.25, 0.3 to 0.4. Within each panel, the number of observation varies as either 5 or 10. The dashed horizontal lines represent the true parameter values in the simulation.

3.3.2.1 Estimated transition intensities Figures 11 and 12 show the results of the estimated transition intensities $\hat{\lambda}_{01}$ and $\hat{\lambda}_{10}$, respectively. The error bars are the 95% empirical confidence interval around the point estimates. Similar pattern can be seen for both $\hat{\lambda}_{01}$ and $\hat{\lambda}_{10}$. The accuracy of the estimates is highly dependent on the size of the transition intensity and the misclassification probabilities. The estimated transition intensities are almost unbiased for all sizes of transitions with very tight 95% confidence interval when there is no misclassification error in the data. With small (in blue)and medium (in black) true transition intensities, the estimated transition intensities are nearly unbiased for almost all magnitude of MCR, except when the MCR is extremely large (e.g. MCR = 40%). With large misclassification in the data, none of the models estimate the transition intensities well, even with the small transition intensity case. The estimates are greatly biased when the true transition intensity is large. However, the estimation is improved with more observed time points except for the scenario with large transition intensity and large MCR.

3.3.2.2 Estimated misclassification probabilities Figures 13 and 14 show the results of the estimated misclassification probabilities $\hat{\pi}_{o=1|s=0}$ and $\hat{\pi}_{o=0|s=1}$, respectively for all scenarios. In panels labeled "No MC", data modeled are correctly classified while still allowing the CTHMM to estimate the misclassification probabilities. The dashed horizontal lines represent the true misclassification probabilities in the simulation. For scenarios with small (in blue) and medium (in black) transition intensities, the estimated misclassification probabilities are almost unbiased even for relatively large misclassification rate (MCR = 0.3). In the case of large transition intensities,

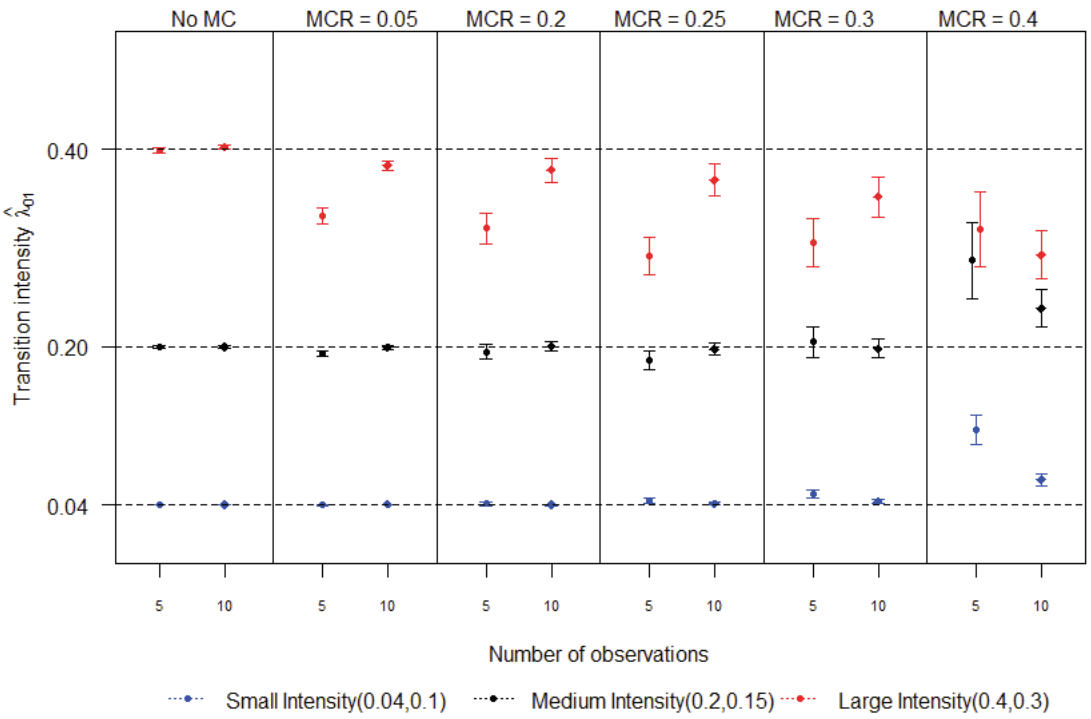

Figure 11: Estimated transition intensity $\hat{\lambda}_{01}$

the estimated misclassification rates are biased. No clear change pattern can be found for the estimates with increasing MCR. However estimates are improved with more observed time points. Since in reality, it is unknown whether data are misclassified, we fit CTHMM on the correct data, allowing the model to estimate the misclassification probabilities as well. The estimated misclassification probabilities in the case, where actually there is no misclassification, are $\leq 1\%$ for the small/medium transition intensities scenarios. For large intensities with 5 observed time points, the estimated misclassification probabilities are around 5%, which may be misleading in practice. However, these probabilities decrease if more observed time points are available.

3.3.2.3 Model diagnosis: Pearson type goodness-of-fit test A p-value based on the Pearson-type goodness-of-fit test less than 0.05 indicates that the model is a poor fit. If the model assumption is correct - that is, a basic Markov model fit the true data and a CTHMM fit the misclassified data - a p-value of the test less than 0.05 should be observed 5% of the times among all simulation

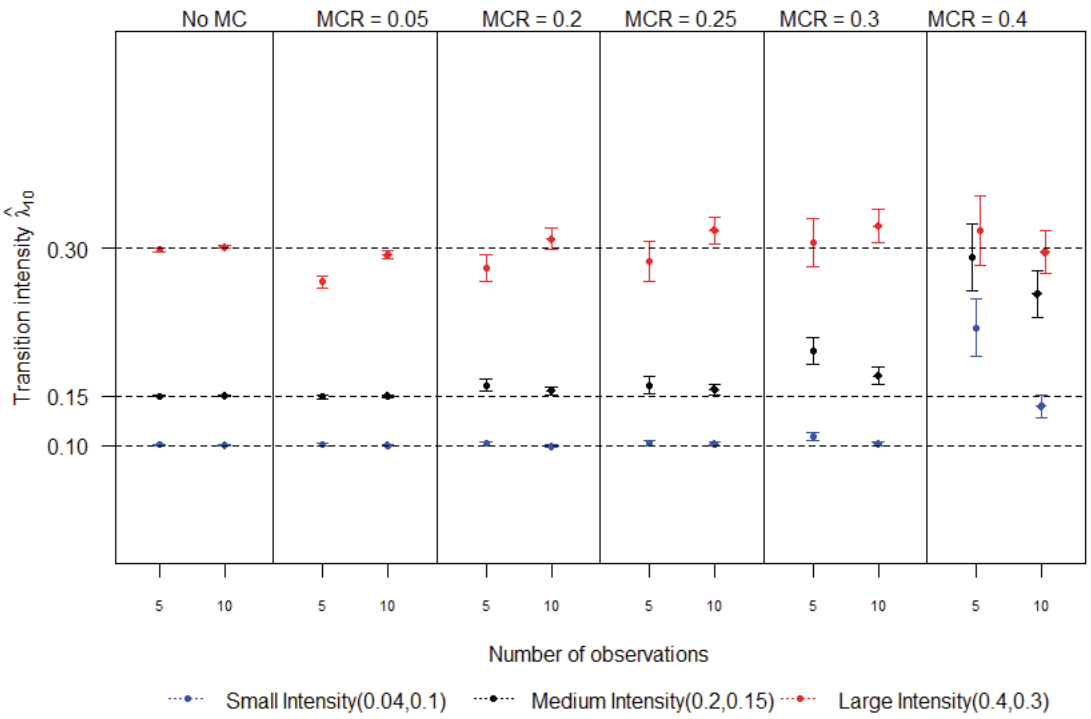

Figure 12: Estimated transition intensity $\hat{\lambda}_{10}$

iterations. This represents the false likelihood of lack of fit (type I error) since the assumption about the misclassification error is correctly specified or accounted for in the models. In contrast, a higher percentage of p-value less than 0.05 should be expected if the model assumption is wrong. In other words, a basic Markov model fitted on misclassified data and a CTHMM fitted on true data should lead to a p-value less than 0.05 with greater frequency among all simulation iterations depending how powerful the test is in the different scenarios.

Figure 15 shows the results of the Pearson-type goodness-of-fit test for 5 observed time points. For CTHMM, the asymptotic null distribution of the test statistics is estimated via simulation due to the latent structure (Titman and Sharples, 2008), and an upper and lower bound of the p-values are obtained and presented for the test. When modeling data without misclassification, 5% of the tests indicate lack of fit with the basic Markov model, while almost none of the tests advocate the CTHMM for the same data despite the relatively fewer data points (N = 500, 5 observed time points). However, when data are subject to misclassification, irrespective of the

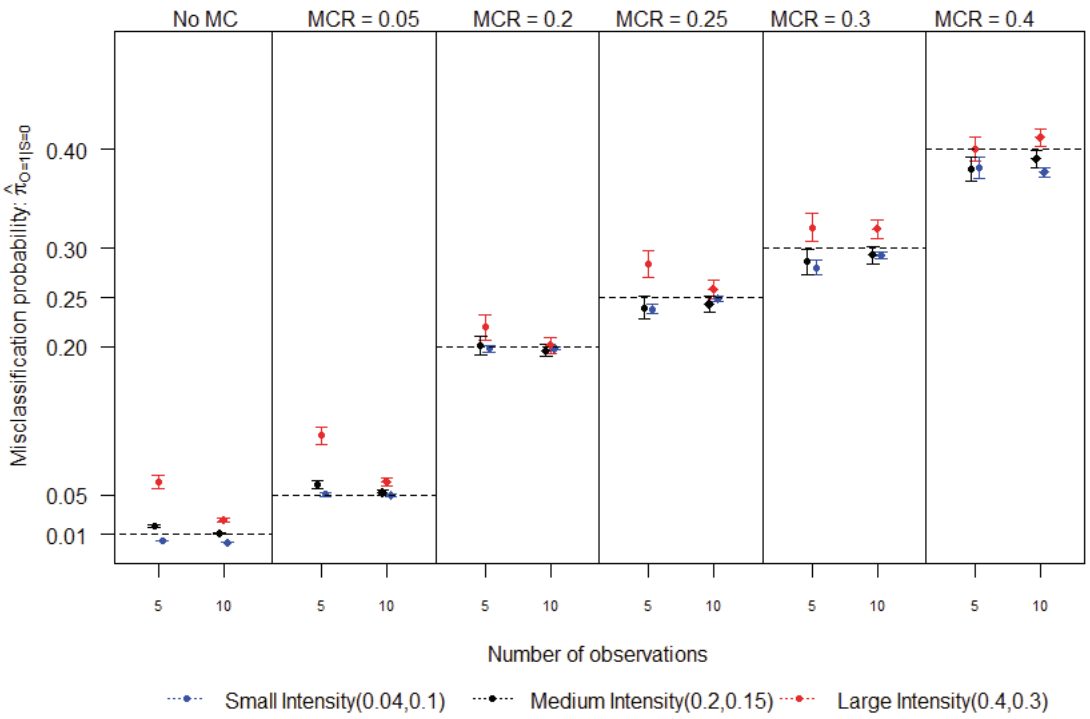

Figure 13: Estimated misclassification probabilities for the observed data: $\hat{\pi}_{o=1|s=0}$

magnitude of misclassification in the data, the Pearson-type Goodness-of-fit test fails to identify the correct model (i.e. the CTHMM) for the misclassified data most of the times. Although with small transition intensities, chances of the tests having a p-value smaller than 0.05 for the wrong models (MM) are greater than those with medium/large transition intensities, the tests ascertain lack of fit for the correct model (CTHMM) too with high frequencies. A trend can be seen that with more misclassification in the data, the tests are less likely to indicate lack of fit for the correct models (CTHMM). However, with increasing misclassification, the same trend can also be found for the wrong model (MM).With 10 observed time points (Figure 16), similar pattern can be seen as in Figure 15 to diagnose the correct model (CTHMM) with improved performance. For the wrong model (MM), with small transition intensities, almost all tests suggest poor model fit irrespective of the misclassification rate except when the MCR is really high (MCR = 0.4). Increasing observed time points does not improve the performance of the tests in the case with medium/large transition intensities for the wrong model (MM).

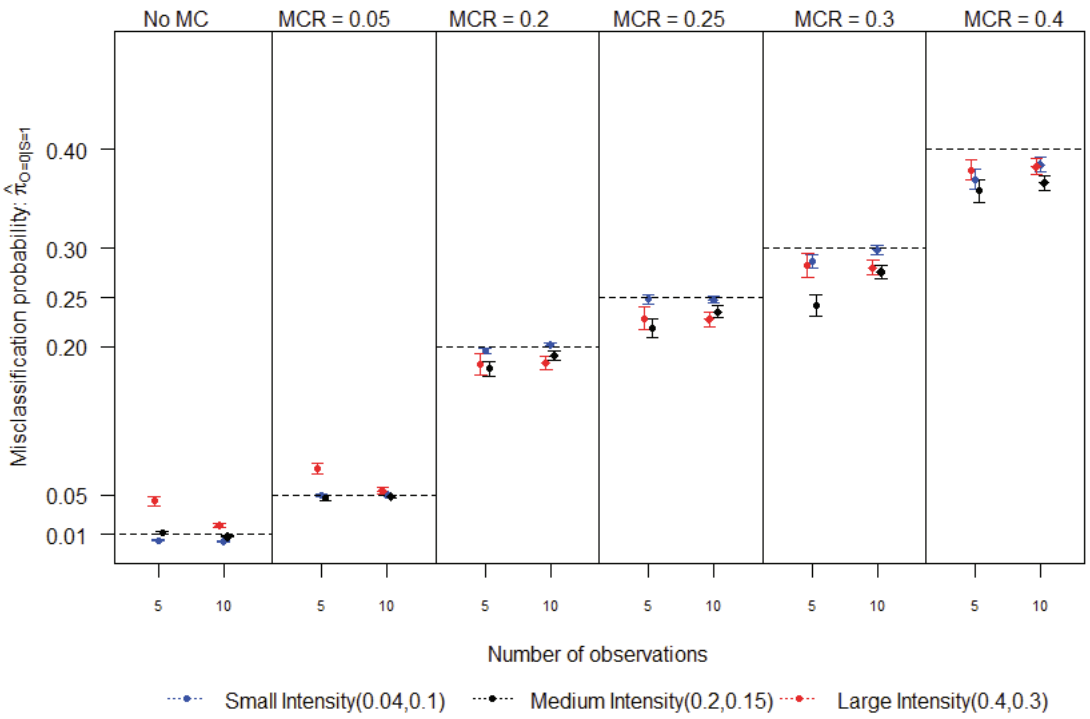

Figure 14: Estimated misclassification probabilities for the observed data: $\hat{\pi}_{o=0|s=1}$

3.4 Modified likelihood and AIC

As seen in the previous section, the Pearson-type goodness-of-fit test does not do a good job distinguishing between CTHMM and MM for data with misclassification. An alternative way to compare the CTHMM and MM for misclassified data is to compare their likelihoods directly. However, due to the structure of the model and the algorithm used to estimate the parameters, the resulting likelihoods for both models are not comparable. For the CTHMM, the likelihood is constructed for all observations as shown in section 2. For the MM, however, the observations at the first time point are not included but are conditioned on while the likelihood is maximized. The

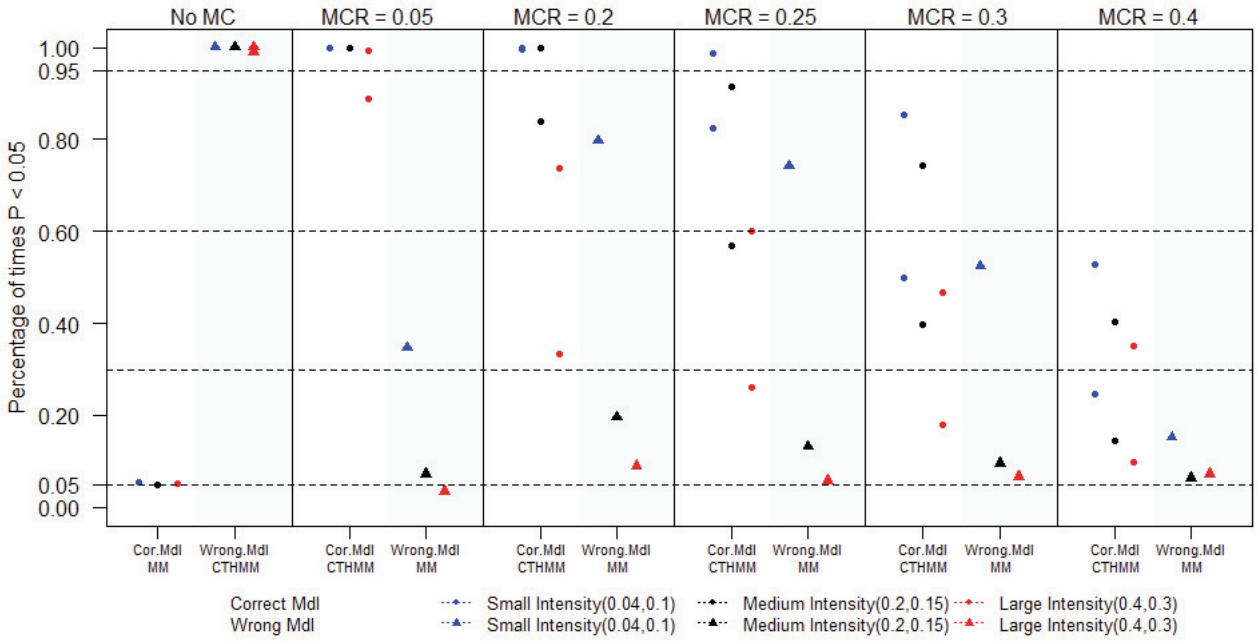

Figure 15: Pearson type goodness-of-fit test for 5 observed time points

likelihood of the MM can be expressed as:

$$L(\Theta|s^{(t_2)}, s^{(t_3)}, \ldots, s^{(t_m)}|s^{(t_1)}) = f(s^{(t_2)}, s^{(t_3)}, \ldots, s^{(t_m)}|s^{(t_1)}, \Theta)$$
$$= \prod_{l=1}^{m} \left\{ \prod_{i,j=0,1} \hat{p}_{i,j}(t_{l-1}, t_l)^{n_{i,j,l}} \right\}$$

where $n_{i,j,l}$ represents the total number of observed transitions from state i to state j at time t_l. The full likelihood, which includes the observations at the first time point, then can be modified as:

$$f(s^{(t_1)}, s^{(t_2)}, \ldots, s^{(t_m)}|\Theta)$$
$$= f(s^{(t_2)}, s^{(t_3)}, \ldots, s^{(t_m)}|s^{(t_1)}, \Theta) \times f(s^{(t_1)}|\Theta)$$
$$= \prod_{l=1}^{m} \left\{ \prod_{i,j=0,1} \hat{p}_{i,j}(t_{l-1}, t_l)^{n_{i,j,l}} \right\} \times \prod_{i=1}^{k} \hat{\pi}_i(t_1)^{n_{i,t_1}}$$

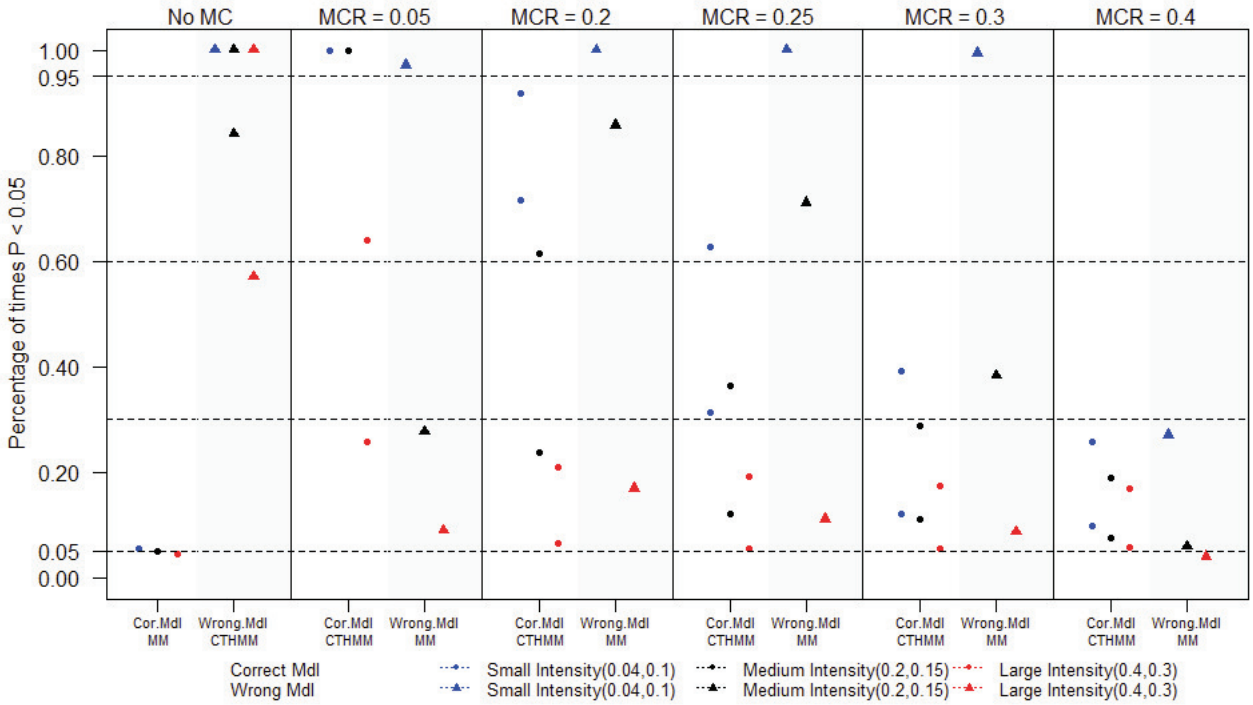

Figure 16: Pearson type goodness-of-fit test for 10 observed time points

where $\hat{\pi}_i(t_1)$ is the estimated probability that the observed state at time t_1 is i and n_{i,t_1} is the observed number of subjects belonging to state i at time t_i. For a two-state MM, $\hat{\pi}_i = p(s^{(t_1)} = i), i = 0, 1$. Define the state distribution at t_2 as $\hat{c}_i$, where $\hat{c}_i = p(s^{(t_2)} = i), i = 0, 1$. Then, given the state distribution $\hat{\pi}_i$ at t_1, the state distribution $\hat{c}_0$ at t_2 is:

$$\hat{c}_0 = p_{00}(t_1, t_2) \times \hat{\pi}_0 + p_{10}(t_1, t_2) \times \hat{\pi}_1$$

where $p_{00}(t_1, t_2)$ is the probability of staying at state 0 from t_1 to t_2 and $p_{10}(t_1, t_2)$ is the probability of transitioning from state 1 to state 0 from t_1 to t_2. Because $\hat{\pi}_1 = 1 - \hat{\pi}_0$ and after a little algebra, the estimated distribution of $\hat{\pi}_0$ is:

$$\hat{\pi}_0 = \frac{\hat{c}_0 - \hat{p}_{10}(t_1,t_2)}{\hat{p}_{00}(t_1,t_2) - \hat{p}_{10}(t_1,t_2)}$$

Therefore, the estimated distribution of the states at t_1 can be calculated based on the estimated

transition probability $\hat{p}_{ij}$ from t_1 to t_2 and the estimated state distribution $\hat{c}_i$ at t_2. For more than two states, the estimated distribution of the states at t_1 can be calculated in a similar way (See appendix for the results of three states).

Then AIC for both CTHMM and MM can be calculated and compared using:

$$\text{AIC} = 2k - 2ln(L_{full})$$

where k is the number of the parameters estimated in the models and L_{full} is the likelihood of the full data. A smaller AIC indicates better model fit.

To investigate the performance of the AIC, we conducted Monte Carlo simulations. The simulation is set up in the same way as in the previous section. A total of 500 data sets are simulated for each scenario. The performance of the AIC is described as the percentage of times that the AIC identifies the correct model. That is, for data without misclassification, MM fit the data better and should have a smaller AIC compared to fitting CTHMM on the same data; for data with misclassification, CTHMM should have a smaller AIC instead.

Figure 17 presents the results of AIC for all scenarios. When there is no misclassification (No MC), based on the AIC, CTHMM are selected as the correct model for the data 5% among the simulation iterations with all three different sizes of transition intensities. This represents the probability that a wrong decision is made based on the AIC, where the MM should be chosen for data without misclassification. For data with misclassification, in the case with small transition intensity (in blue), the CTHMM are correctly selected by AIC most of the time, except when the MCR is high (e.g. MCR = 40%). With medium-size transition intensity (in black), the performance of the AIC declines, especially with small (e.g. 5%) and relatively large (e.g. 30%) MCR. The performance of the AIC is even worse in the scenarios with large transition intensity (in red). In less than 20% of the simulation iterations, the AIC advocates the CTHMM as the correct model for the misclassified data. However, the performance of the AIC is improved in all scenarios with increasing number of the observations from 5 to 10, even with very large MCR. The improvement is dramatic for

scenarios with medium transition intensity (in black). That is, with more observations per subject, more information can be gathered to compare the models.

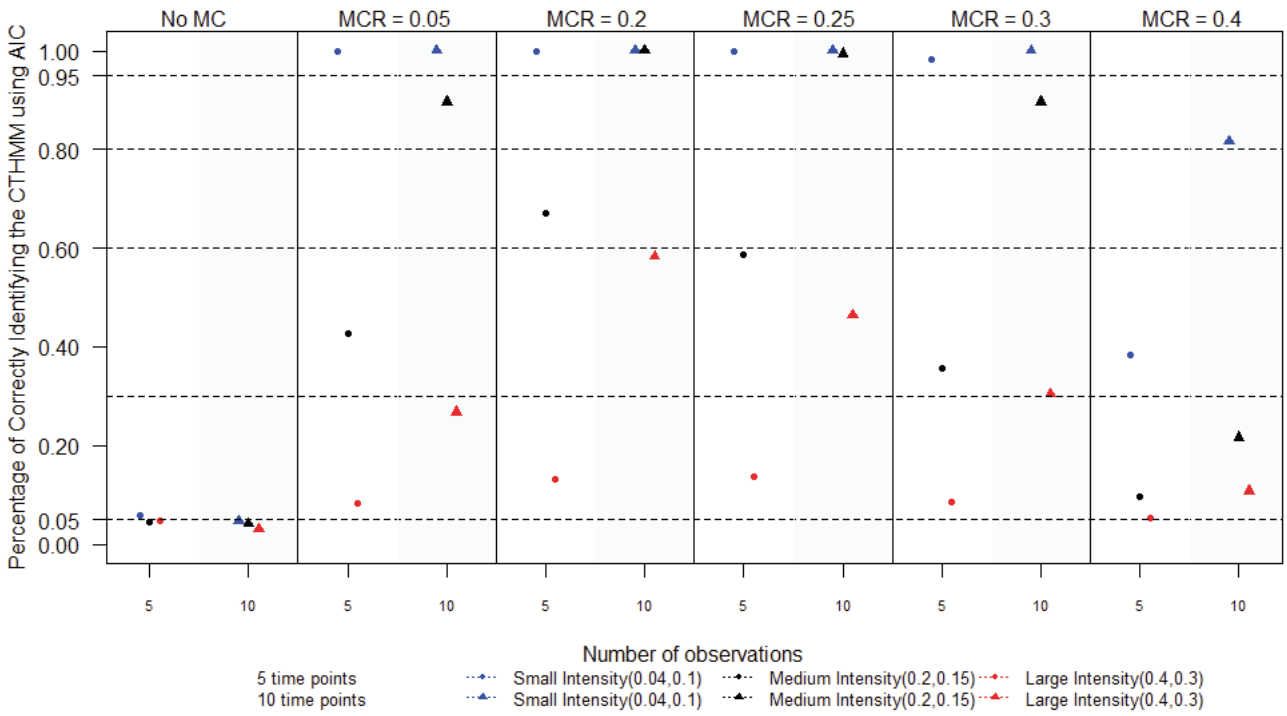

Figure 17: Performance of AIC based on the modified likelihood of MM

3.5 Discussion

The CTHMM has been proposed to model panel-observed data subject to misclassification. While the two-layer structures provide a flexible framework to model the data, the estimability and/or accuracy of all parameters and model diagnosis are challenging due to the uncertainty of the hidden layer and the inherent drawbacks of the EM algorithm. Although studies have been done to improve the estimation and the assessment of the model fit, limited work has shown the performance of CTHMM with application to misclassified data. We evaluate the performance of such models with a focus on the estimation of the misclassification probabilities solved by the most common algorithm implemented in the literature. We also propose an alternative approach for model diagnosis using AIC based on the modified full-data likelihood.

In our simulation, all parameters are estimable; however, the accuracy of the estimation varies. One key parameter that impacts the estimation is the transition intensity. For small to medium transition intensities corresponding to long to mid-length mean sojourn time relative to the observation intervals, parameters including transition and misclassification can be well estimated as long as the classification error is not too high (MCR $< 30\%$). A similar result was discussed by Rosychuk et al. (2004). In their simulation, all the interval times are equal to 7 days with mild misclassification (2% and 10%). They found that in order to estimate the parameters accurately, the mean sojourn time needs to be at least three times the observation interval time. In other words, we need to collect the data often enough to get enough information about the parameter. In our simulation, because our observation schedule is unequally spaced, this may introduce additional variability - more information - to the model. Thus in the case with medium transition intensities, where the mean sojourn time is about twice as long as the simulated interval time, parameters can still be well estimated. For the large intensities case, although the estimated transition intensities are quite biased, the estimated misclassification probabilities are reasonably close to the true parameter values. In such case, we can still glean some insight into the potential misclassification error even though the transition intensities are poorly estimated. For higher misclassification rates ($> 30\%$), the point estimates of the misclassification probabilities are biased, even with small transition intensities. Changing the initial values of the parameters does not improve the accuracy of the estimation. We explored this issue (results not shown) for the two-state cases and find that with a large amount misclassification, the surface of the likelihood is flattened, which makes it difficult for the algorithm to converge to the true parameter value, no matter which starting value is given, a problem discussed by Rosychuk et al. (Rosychuk and Thompson, 2003; Rosychuk RJ, 2004). Furthermore, with more observations per subject (e.g. 10 observations per subject), the point estimates are improved and the uncertainty of the estimates is reduced.

We also found that the modified Pearson-type goodness-of-fit test does not work well to identify a correct model for the misclassified data in our simulation. The poor performance may be partly due to lack of power. As we have seen in the simulation, with more observations, the performance of the Pearson-type goodness-of-fit test is improved. On the other hand, due to the latent structure

of the CTHMM, the estimated counts of individual cell are calculated by summing over all possible latent states, which may potentially impose more difficulties on identifying the fit of the models. Our proposed AIC approach works well in terms of selecting the right model for the data with or without misclassification. However, the performance of our proposed approach is also greatly impacted by the relative size of the transition intensities compared to the observation schedule, although it can be improved with more observed time points. In addition, violation of other assumptions may also influence the performance of the test; applying different procedures including graphical approaches are recommended to assess the model fit thoroughly.

We also explored the impact of small sample size (n = 100) on the performance of the model (results are not shown). In cases with small transition intensities, the performance of the CTHMM with respect to point estimates and our proposed AIC approach are similar to that when the sample size is medium to large (n = 500). In scenarios with medium and large transition intensity, the estimated misclassification probabilities are still close to the true values although biased. However, our proposed AIC approach suffers when the size of the transition intensity is medium when n = 100. Among all simulation iterations, less than 30% of the times the models are identified correctly with only 5 observations per subject. Increasing the total number of observations per subject improves the performance from 30% to around 65%. With large transition intensity, the performance is worse than the small/median transition intensity cases. That is to say, in order to identify the misclassification error, either a large enough sample size or a frequent sampling scheme (i.e. follow-up schedule) is needed to draw the conclusion.

There are some limitations of this study. First, we assume the state space is known beforehand. However, the state space may not be known, especially if we use categorized scores to represent the state space. Various ways are described to categorize a single score, which may potentially impose issues on the resulting estimates. Secondly, the assumption of constant misclassification probabilities may also be questionable in practice. In a longitudinal study setting, physicians may make fewer errors after they see the patients more often and repeatedly. Therefore it may be more realistic to assume that the misclassification probabilities decrease as time goes on. This

assumption needs to be assessed and will be a future topic for study.

In summary, CTHMM can be applied to estimate misclassification probabilities while taking into account changes of disease statuses. Although the assumptions of such models are strong, the results still provide useful information about potential error, especially if misclassification probabilities are not too high. If the Markov assumption is reasonable for given data with repeated measurements, one could start with CTHMM since, as we have shown in the simulation, with correctly classified data, estimates from CTHMM are small and AIC based on the modified likelihood can identify the correct model. The estimated misclassification probabilities then can be used for the purpose of bias correction. In addition, if the estimated transition intensities are small, i.e. the mean sojourn time is long relative to the interval time of the observation schedule, one can be confident about the results. Otherwise, data need to be collected more often in order to get better estimates or the accuracy of the estimates might be questionable.

Acknowledgements

This work is supported by a National Institute of Neurological Disorders and Stroke (NINDS) grant, U01 NS054630, P.I.: Yuko Y. Palesch, Ph.D. and Neurological Emergencies Treatment Trials (NETT) Network grant, U01 NS059041, P.I.: Yuko Y. Palesch, Ph.D.

Appendix 1. The estimated state distribution at t_1 for three states with state space 0, 1, 2

The estimated state distribution at t_2 for three states with state space $\{0, 1, 2\}$:

$$\begin{cases} \hat{c}_0 = \hat{p}_{00}\hat{\pi}_0 + \hat{p}_{10}\hat{\pi}_1 + \hat{p}_{20}\hat{\pi}_2 \\ \hat{c}_1 = \hat{p}_{01}\hat{\pi}_0 + \hat{p}_{11}\hat{\pi}_1 + \hat{p}_{21}\hat{\pi}_2 \\ \hat{c}_2 = \hat{p}_{02}\hat{\pi}_0 + \hat{p}_{12}\hat{\pi}_1 + \hat{p}_{22}\hat{\pi}_2 \end{cases}$$

And $\hat{\pi}_2 = 1-\hat{\pi}_0-\hat{\pi}_1$. After some algebra, the estimated state distribution at t_1 can be calculated as:

$$\begin{cases} \hat{\pi}_0 = \frac{(\hat{c}_0-\hat{p}_{20})(\hat{p}_{11}-\hat{p}_{21})-(\hat{c}_1-\hat{p}_{21})(\hat{p}_{10}-\hat{p}_{20})}{(\hat{p}_{00}-\hat{p}_{20})(\hat{p}_{11}-\hat{p}_{21})-(\hat{p}_{01}-\hat{p}_{21})(\hat{p}_{10}-\hat{p}_{20})} \\ \hat{\pi}_1 = \frac{(\hat{c}_0-\hat{p}_{20})(\hat{p}_{01}-\hat{p}_{21})-(\hat{c}_1-\hat{p}_{21})(\hat{p}_{00}-\hat{p}_{20})}{(\hat{p}_{00}-\hat{p}_{20})(\hat{p}_{01}-\hat{p}_{21})-(\hat{p}_{11}-\hat{p}_{21})(\hat{p}_{00}-\hat{p}_{20})} \end{cases}$$

4 Original Manuscript 3

Title: A Joint Modeling Analysis Strategy for Bias Correction Caused by Covariate Misclassification

Authors: Liqiong Fan(a) ,Sharon D. Yeatts (a), Bethany J. Wolf (a), Leslie A. McClure (b) , Magdy Selim (c), Yuko Y. Palesch (a)

Affiliation:

(a) Department of Public Health Sciences, Medical University of South Carolina, Charleston, SC 29425, U.S.A.

(b) Department of Epidemiology & Biostatistics, Drexel University, Philadelphia, PA 19104, U.S.A.

(c) Beth Israel Deaconess Medical Center, Department of Neurology,Division of Cerebrovascular Diseases, Boston MA 02215, USA.

Submission Status: drafted, ready for submission.

Abstract

Covariate-adjusted regression analysis is very common to analyze data from randomized clinical trials to demonstrate the effect of an intervention. However, adjusting for misclassified covariate will result in biased estimate for the treatment effect, which is assumed to be measured without error. We propose a joint modeling analysis strategy, which combine the continuous-time hidden Markov model (CTHMM) and the misclassification simulation extrapolation (MCSIMEX) method to correct the bias of the estimate for the treatment effect when the misclassified covariate is measured repeatedly. We demonstrate the performance of the CTM-MCSIMEX estimator in terms of type I error, power for the hypothesis testing of the treatment effect and bias correction of the estimates. With mild to moderate misclassification in the covariate, more than 50% of the bias can be corrected by CTM-MCSIMEX estimator given that the misclassification probabilities can be estimated by CTHMM accurately in the first stage. However, the power of the hypothesis testing is not improved and the type I error is slightly

inflated using CTM-MCSIMEX estimator. Estimates based on CTM-MCSIMEX estimator can provide some insight of the treatment effect when no misclassification error can be assumed and can be used as reference for future study design.

Key words: CTHMM, MCSIMEX, covariate misclassification, bias correction

4.1 Introduction

Covariate-adjusted regression analysis is very common to analyze data from randomized clinical trials to demonstrate the effect of an intervention. Incorrect measurement of the required covariates could potentially cause problems for the pursuit of effective prognostic adjustment, resulting in both biased treatment effect estimation and power loss for hypothesis testing. When the mis-measured variable is categorical, it is referred to as misclassification. Measurement error has been a long-standing topic in the statistical literature, primarily in the context of observational studies, where the misclassified variable is of main interest. Methods have been proposed to correct the bias associated with the estimate of the misclassified variable. However, in the clinical trials setting, emphasis is more likely to be on the estimation of the effect associated with the treatment, which is considered to be recorded without error. Although these proposed methods developed for a misclassified variable may be applied for a perfectly measured variable as well, limited research has focused on the performance of bias correction with respect to the perfectly measured variable using such methods.

It is not uncommon for prognostic covariates to be measured with error or misclassified. For example, disease severity is an important prognostic covariate for most diseases and is often adjusted for in an analytic model for the primary analysis. However, the evaluation of the severity may be easily confounded by inter-observer variability or clinical conditions such as sedation and/or intubation or drug or alcohol usage, resulting in inaccurate assessment. The National Institutes of Health Stroke Scale (NIHSS) for stroke and the Glasgow Coma Scale (GCS) for traumatic brain injury (TBI) are two examples. In addition, the adjustment of the disease severity is often based on categorized scales for both biological and statistical consideration. No direct assessment is available to validate the accuracy of such evaluations and categorizations. Therefore, the true severity is unknown in most cases.

Many approaches developed for correcting covariate misclassification are two-stage models. In the first stage, the parameters describing the characteristics of misclassification are estimated using

an internal/external validation data set, repeated measurements, or instrument variables. Those parameters can then be incorporated in the analytic model as weights in the second stage. Since the true severity is unknown in the examples given earlier, we will focus on the methods using repeated measurements. White et al. (2001) demonstrate the possibility of using regression calibration for error-prone categorical variables when more than two measurements of the variable are available. Wang et al. (2000) propose an Expected Estimating Equation (EEE) method to account for the measurement error in longitudinal data. With misclassification, method of moments is recommended to solve for the EEE. Chen et al. (2014) propose a two-stage estimation approach for longitudinal ordinal data with misclassification in both the response and covariates based on estimating equations.

Multistate models (MSM) are very useful tools to model disease dynamics and have been widely used for longitudinal categorical data (Diggle et al., 2002). The continuous-time hidden Markov Models (CTHMM) are one type of MSM, which provide a convenient framework to simultaneously estimate disease development and potential probability of misclassification of the data (Bureau et al., 2003). The disease dynamics are captured in the transition through latent states, which follows a continuous-time Markov chain; the misclassification is characterized via the emission probabilities, which are the probabilities of the observed status given the underlying true disease status. Jackson et al. (2003) proposed such application of CTHMM for data subject to misclassification and developed an R package (msm) to ease model fitting. CTHMMs have also been applied to data in different disease areas including cognitive decline and dementia (Marioni et al., 2012; Norton et al., 2013; Buter et al., 2008), HIV/HPV (Human Papilloma Virus) infection (Blitz et al., 2013), liver cirrhosis and hepatitis C (Bartolomeo et al., 2011; Terrault et al., 2008; Sweeting et al., 2006).

The Misclassification Simulation Extrapolation model (MCSIMEX) is an extension of the Simulation Extrapolation model (SIMEX), which was originally designed for continuous measurement error (Kuchenhoff et al., 2006). The basic idea of MCSIMEX is that, given the known misclassification probabilities in the data, the behavior of the estimation from the model can be "calibrated"

by adding extra known increments of misclassification into the data. Therefore, the inference in the case where there is no misclassification can be obtained by extrapolating the trend/behavior back. It can be applied to covariate misclassification, outcome misclassification or both. Due to its ease of implementation and flexibility of dealing with more complicated modeling, MCSIMEX has received considerable attention with application to genotyping data (Lamina et al., 2010; Heid et al., 2008), periodontal disease (Slate and Bandyopadhyay, 2009), social science (Hopkins and King, 2010) and occupational disease (Costas et al., 2015). However, there has not been an extensive assessment of the performance of MCSIMEX in the literature.

In this paper we demonstrate the feasibility and performance of combining the two models, the CTHMM and the MCSIMEX in the context of estimating an exactly measured variable when a potentially confounding variable is mis-measured. The joint modeling provides a flexible way to adjust for the biased estimate of the treatment effect caused by covariate misclassification in the situation where the misclassification probabilities are unknown, but repeated measurements for the misclassified variables are available. We first assess the performance of MCSIMEX regarding bias correction for the perfectly measured variable. Then the two models are combined to provide the estimation of the treatment effect given the prognostic covariate with correction for misclassification. All evaluations are carried out using Monte Carlo Simulation methods. The paper is organized as follows. We introduce a motivating example in section 2. The analysis strategy is described in section 3. In section 4, the simulation used to evaluate the performance of the MCSIMEX model and the combined approach is described, and the results are presented. The combined approach is then applied to analyze the motivating data, and the results are presented in section 5. We close the paper with discussion and recommendations based on the simulation results.

4.2 Motivating example: The IMS III trial

The Interventional Management of Stroke (IMS) III trial (Broderick, 2013) was a randomized controlled phase III trial, designed to determine the efficacy of endovascular therapy following intravenous rt-PA initiated within 3 hours of symptom onset of ischemic stroke. The primary

outcome was functional recovery at 3 months post-stroke, defined based on the modified Rankin Scale (mRS), which was dichotomized as favorable outcome (mRS 0-2) vs. unfavorable outcome (mRS > 2). Adjustment for baseline stroke severity was prespecified.

The stroke severity was defined based on dichotomized NIHSS (the National Institutes of Health Stroke Scale) with NIHSS ≤ 20 representing a moderate stroke and NIHSS ≥ 21 a severe stroke. Some misclassification of the stroke severity was discovered during the data monitoring process. Approximately 2% misclassification was identified at each level of the stroke severity by the review of the Data and Safety Monitoring Board during the trial. Moreover, some residual error may still exist either due to the uncertainty of the selected cutoff point for dichotomization or other confounding factors while evaluating patients' stroke severity. Therefore, the true misclassification rate of stroke severity was unknown and no validation data were available. In addition to baseline NIHSS, all subjects were evaluated repeatedly using NIHSS at 40 minutes, 24 hours, and 5 days, which provided extra information to estimate the true magnitude of misclassification.

For ease of illustration, we assume stroke severity, which is referred to as the misclassified covariate, is the only covariate that needs to be adjusted for in the primary analysis. We will refer to the treatment assignment as a perfectly measured variable in the model, which is a common assumption in practice for intention to treat analysis.

4.3 General consideration for analysis strategy

We will focus on the logistic regression model with additional information of repeated measurements for the error-prone covariate since this is the scenarios described in our motivating example. In order to correctly estimate the treatment effect accounting for the covariate misclassification, we propose a two-stage modelling strategy, which combines the CTHMM and the MCSIMEX model. In the first stage, the error-prone covariate will be treated as the outcome, and the misclassification probabilities will be estimated via CTHMM. Then in the second stage, the estimated misclassification probabilities from the CTHMM will be incorporated into the MCSIMEX algorithm, and

the inference about the treatment effect will be made according to the pre-specified analytic model after MCSIMEX correction. Detailed setup for the two-stage modelling is described below. For ease of notation, we will omit the subscript i, representing the i^{th} subject in the study throughout.

4.3.1 Misclassification Model

Suppose the true stroke severity at time point t_m, denoted as $S^{(t_m)}, m = 1, 2, \ldots, n$, follows a continuous time Markov chain with state space $S = \{0, 1\}$, where 0 and 1 represent a mild and a severe stroke respectively. Then the transition of the true severity between time points t_m and t_{m+1} is captured via transition intensity, denoted as λ_{lj}, with l, j representing states occupied,$l, j = \{0, 1\}$, and $\lambda_{lj} = \lim_{\Delta t \to 0} \frac{Pr(S^{(t_{m+1})}=j|S^{(t_m)}=l)}{\Delta t}$. The transition probabilities between time points can be calculated via matrix exponential for a given time interval Δt between time points. For the two-state case, direct calculation can be carried out using an analytic expression as:

$$P_{01}(\Delta t_m) = P(S^{(t_{m+1})} = 1|S^{(t_m)} = 0) = \frac{\lambda_{01}}{\lambda_{01}+\lambda_{10}}\{1 - e^{-\Delta t_m(\lambda_{01}+\lambda_{10})}\}$$

$$P_{10}(\Delta t_m) = P(S^{(t_{m+1})} = 0|S^{(t_m)} = 1) = \frac{\lambda_{10}}{\lambda_{01}+\lambda_{10}}\{1 - e^{-\Delta t_m(\lambda_{01}+\lambda_{10})}\}$$

Conditioning on the true severity, the observed severity,$O^{(t_m)}, O = \{0, 1\}, m = 1, 2, \ldots, n$, defined according to NIHSS is independent across the time points and has the misclassification probabilities defined as $\pi_{o|s} = Pr(O^{(t_m)} = l|S^{(t_m)} = j)$. We assume the misclassification probabilities remain constant across all time points. Together with the distribution of the true severity at $m = 1$, the observed data likelihood can be expressed as:

$$\begin{aligned} L(\mathbf{O}) &= L(O^{(t_1)}, O^{(t_2)}, \ldots, O^{(t_m)}) \\ &= \sum_{s^{(t_1)}} \sum_{s^{(t_2)}} \cdots \sum_{s^{(t_m)}} \pi(s^{(t_1)}) \prod_{m=1}^{n-1} P_{s^{(t_m)}s^{(t_{m+1})}}\{t_{m+1}, t_m|S = s^{(t_m)}\} \prod_{m=1}^{n} \pi_{o|s} \end{aligned}$$

The likelihood can be solved and the misclassification probabilities can be estimated via the Expectation Maximization (EM) - forward and backward algorithm developed by Baum et al. (Baum,

1970).

4.3.2 The Analytic model

Let Y denote the dichotomized primary outcome of interest. The primary analytic model has the form:

$$Logit\{Pr(Y = 1|Z, S^{(t_1)})\} = \beta_0 + \beta_z Z + \beta_s S^{(t_1)}$$

where Z represents the treatment assignment and is recorded without error. $S^{(t_1)}$ is the error-prone dichotomized prognostic covariate at baseline ($t_m = t_1$). In the case of misclassification, $S^{(t_1)}$ is not observed, but $O^{(t_1)}$, the misclassified version of $S^{(t_1)}$, is observed. Then the above equation becomes:

$$Logit\{Pr(Y = 1|Z, O^{(t_1)})\} = \beta_0^* + \beta_z^* Z + \beta_o^* O^{(t_1)}$$

We refer to the β^*s as naive estimators. With the estimated misclassification probabilities (Π) and a parametric approximation, $\beta^*(\Pi^\zeta) \approx F(1 + \zeta, \Gamma)$, where ζ is the amount of misclassification in addition to that already in the data and Γ is the extrapolation function, the MCSIMEX estimators or the corrected estimators can be extrapolated back to when $\zeta = -1$, representing no misclassification. Detailed information about the algorithm can be found in (Kuchenhoff et al., 2006).

4.4 Monte Carlo Simulation

Simulation is used to investigate the feasibility and performance of combining the two methods, the CTHMM and the MCSIMEX, in order to correctly estimate the treatment effect, accounting for covariate misclassification. As illustrated in our previous work, CTHMM can well estimate the misclassification probabilities for data subject to misclassification. For this paper, we first show the

performance of the MCSIMEX model regarding bias correction for the perfectly measured variable with respect to the accuracy of the estimated misclassification probabilities and the relative covariate effect to the treatment. Secondly, we examine the bias correction for the perfectly measured variable using the joint model.

4.4.1Performance of MCSIMEX model

Both treatment assignment Z and true stroke severity at baseline $S^{(t_1)}$ are generated form a Bernoulli distribution with $p_z = 0.5$ for Z and $p_{s^{(t_1)}} = 0.4$ for $S^{(t_1)}$ respectively. That is, $Pr(Z = 1) = Pr(Z = 0) = 0.5$ and $Pr(S^{(t_1)} = 1) = 0.4$. A misclassified version of $S^{(t_1)}$, denoted as $O^{(t_1)}$, is also generated with various misclassification rate(MCR) from 0 to 40%, assuming that the misclassification rates (MCRs) are the same for each level of the prognostic covariate(i.e. $\pi_{01} = Pr(O = 0|S = 1) = \pi_{10} = Pr(O = 1|S = 0)$). The response variable Y is generated from a Bernoulli distribution where the $Pr(Y = 1)$ is calculated based on the linear combination of Z, $S^{(t_1)}$ and their corresponding coefficients, which vary across levels of β_s from -0.5 to -2.5 while fixing β_z at 0.405. this resulted in a wide randg of absolute relative effect sizes of the covariate to that of the treatment (i.e. β_s/β_z) from around 1.2 to 6.2. After the data are generated, a total of 5 models are fit on either the true data (i.e. $Y, Z, S^{(t_1)}$) or the misclassified data (i.e. $Y, Z, O^{(t_1)}$) with or without MCSIMEX correction, listed as follows:

i) True model: adjusted for the true covariate value Z;

ii) Naïve model: adjusted for the misclassified covariate value $O^{(t_1)}$;

iii) True MCSIMEX model (MCSIMEX.true): apply the MCSIMEX correction using the true misclassification probability with which data were generated;

iv) Mis-specified MCSIMEX model1 (MCSIMEX.mis1): apply the MCSIMEX correction with mis-specified MCR at 30% less than the truth for all scenarios;

v) Mis-specified MCSIMEX model2 (MCSIMEX.mis2): apply the MCSIMEX correction with mis-specified MCR at 30% more than the truth for all scenarios.

For the MCSIMEX models, we used the quadratic extrapolation function to calculate the MCSIMEX estimators because it has better performance than the linear extrapolation function (Kuchenhoff et al., 2006). For each scenario described above, 1,000 datasets were simulated using Monte Carlo simulation method. The empirical percent bias and mean squared error of the estimated $\hat{\beta}_z$ and $\hat{\beta}_s$ obtained from the above models were then determined as:

$$\%Bias = \frac{1}{1000} \sum \frac{(\hat{\beta} - \beta_{true})}{\beta_{true}} \times 100\%$$

$$MSE = \frac{1}{1000} \sum (\hat{\beta} - \beta_{true})^2$$

where β_{true} is the true coefficient value.

Figure 18 presents the results for the impact of MCR for a given treatment effect ($\beta_z = 0.405$) on the estimated $\hat{\beta}$ coefficient of the treatment (upper panel) and of the covariate (lower panel). With the correctly specified MCSIMEX.true model, the absolute percent bias (left panel) increases as the MCR increases for both estimates (dashed black lines). However when the misclassification probabilities are mis-specified, the MCSIMEX model tends to correct less bias with smaller MCR (MCSIMEX.mis1: green lines) and overcorrect the bias (MCSIMEX.mis2: red lines) compared to the cases when the true MCR is used for the correction. Especially for relatively small MCR, estimates based on MCSIMEX.mis2 are greater than those based on the true models. The mean squared error of all the MCSIMEX estimators for the treatment effect (top right plot) is reasonably small, which indicates that the estimators are relatively close to the true parameter values.

Figure 19 presents the results for the impact of the covariate effect on the bias correction for a given misclassification rate (MCR = 0.2). Similar patterns are observed for both percent bias and MSE for the estimated coefficient of the treatment effect compared to the impact of varying misclassification rates. With smaller covariate effect on the outcome, percent bias of the treatment effect reduces with

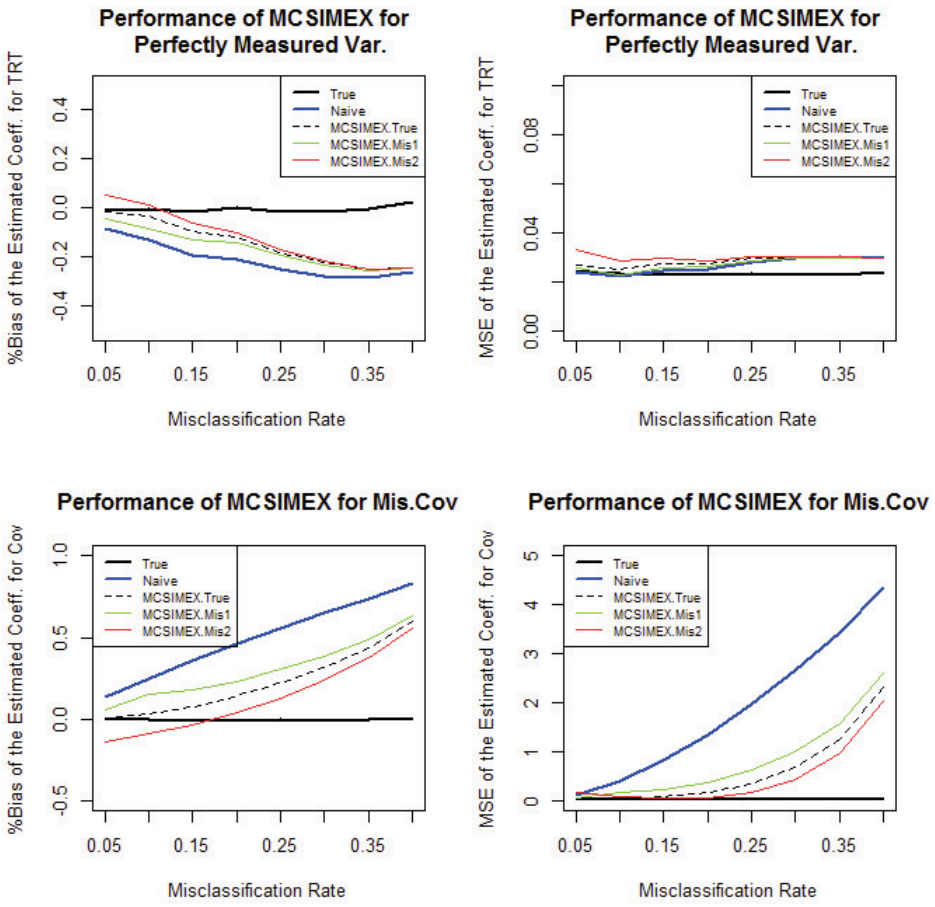

Figure 18: Impact of MCR on the performance of MCSIMEX

MCSIMEX correction. On the other hand, the covariate effect does not influence the magnitude of bias correction for the misclassified covariate itself. That is, the amount of correction for the misclassified covariate in terms of percent bias stays almost the same with increasing covariate effect.

4.4.2 Performance of the Joint Model

For this simulation, we evaluate the performance of combining the two models in terms of the bias correction, power and type I error with respect to the treatment effect. Suppose an adjusted model is pre-specified as the primary analysis for a randomized controlled clinical trial, designed

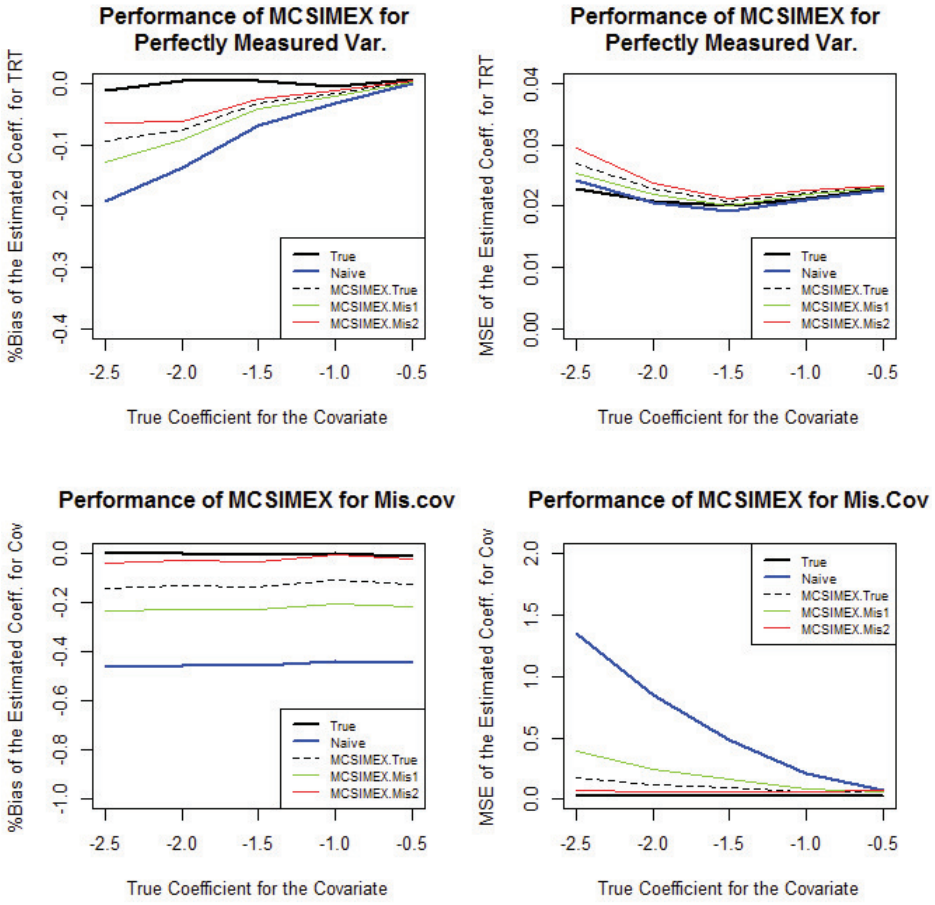

Figure 19: Impact of magnitude of covariate effect on the performance of MCSIMEX

to detect an absolute 10% difference in the treatments, with the success rate in the control group equal to 40%. This is the same assumption made in the IMS III trial (Broderick, 2013). A sample size of 1000 (500 subjects per group) is estimated to detect this unadjusted odds ratio of 1.5 for the treatment effect with 90% power at a 5% significance level. The prognostic covariate that the model is adjusted for is dichotomized and known to be subject to classification error, although the misclassification rates are unknown. However, additional measurements other than the baseline for the covariate are available at the time of the analysis. Five hundred trials are simulated using Monte Carlo simulation methods for each scenario described below. The empirical percent bias and mean squared error for the estimates of the treatment effect is calculated as in the previous section.

In each trial, repeated measurements of the covariate were simulated first with 5 unequally-spaced observations per subject, i.e. the true stroke severity that a subject belongs to at each time point t_m. The true severity $S^{(t_m)} = i, i = \{0, 1\}, t_m = \{1, 3, 5, 8, 12\}$ are assumed to follow a homogeneous continuous-time Markov chain and are generated via the transition intensities matrix Λ specified below:

$$\Lambda = \begin{pmatrix} -0.4 & 0.4 \\ 0.3 & -0.3 \end{pmatrix}$$

where the rows represent the current severity at time t_m and the columns are the severity at time t_{m+1}. That is, the rate of transition from state 0 to state 1 is 0.4 and from state 1 to state 0 is 0.3. Given the true severity, misclassified versions of the covariate at each time point are generated as described in the previous section, with varying misclassification magnitudes ranging from 0% to 40% . The outcome Y is also generated as described in section 3. A total of 3 final analytic models are compared:

i) True model: adjusted for the true covariate value;

ii) Naïve model: adjusted for the misclassified covariate value;

iii) CTM-MCSIMEX model: The joint model, which estimates the misclassification probabilities via CTHMM first and then applies the MCSIMEX algorithm with the estimated misclassification probabilities to the Naïve model.

Since in reality, it is unknown whether misclassification truly exists in the observed data, misclassification probabilities will also be estimated in the simulated true data and the corresponding MCSIMEX estimator is also obtained for the true data.

Table 7 presents the results of the impact of the magnitude of the covariate misclassification on the estimation for the treatment effect. With increasing misclassification rates of the covariate, percent bias of the point estimate in the naïve models increases greatly, from 6.27% in the mild

misclassification case to about 25% in the severe misclassification. The misclassification probabilities are estimated reasonably well via CTHMM; the MCSIMEX estimators partially reduce the bias. CTM-MCSIMEX works for mild and moderate misclassification, where $\geq 50\%$ bias reduction is achieved. However, when a large amount of misclassification (>30%) exists in the covariate, CTM-MCSIMEX estimators do not perform well. In addition, using MCSIMEX estimators do not improve the power of the hypothesis testing for the treatment effect. The power when using the MCSIMEX estimators is similar to that when using the Naïve models. In addition, type I error rates are also slightly inflated using CTM-MCSIMEX estimators compared to the true model and the naïve model.

Table 7: %Bias, MSE, Power and Type I Error for the estimated treatment effect

Estimators($\hat{\beta}_{trt}$)		True parameters: β_{trt}=0.405, β_{sev}=-2.4				
MCR*(Est.MCR)	Models	Mean(% Bias)	SE	MSE	Power	Type I Error
(0,0) (0.015,0.014)	True	0.4012(-0.94%)	0.1488	0.0222	0.798	0.051
	Naive	-	-	-	-	-
	CTM-MCSIMEX	0.4107(1.4%)	0.1517	0.0230	0.782	0.054
(0.05,0.05) (0.055,0.054)	True	0.4064(-0.10%)	0.1395	0.0195	0.794	0.049
	Naive	0.3796(-6.27%)	0.1350	0.0189	0.756	0.052
	CTM-MCSIMEX	0.4038(-0.30%)	0.1499	0.0225	0.754	0.056
(0.2,0.2) (0.190,0.190)	True	0.4011(-0.96%)	0.1489	0.0222	0.762	0.053
	Naive	0.3260(-19.51%)	0.1344	0.0243	0.658	0.050
	CTM-MCSIMEX	0.3597(-11.19%)	0.1559	0.0264	0.664	0.057
(0.3,0.3) (0.270,0.267)	True	0.4174(-3.05%)	0.1465	0.0216	0.790	0.051
	Naive	0.3162(-21.92%)	0.1246	0.0234	0.686	0.048
	CTM-MCSIMEX	0.3322(-17.96%)	0.1354	0.0236	0.676	0.058
(0.4,0.4) (0.383,0.381)	True	0.4140(2.22%)	0.1521	0.0232	0.782	0.054
	Naive	0.3014(-25.58%)	0.1298	0.0276	0.632	0.050
	CTM-MCSIMEX	0.3073(-24.12%)	0.1346	0.0277	0.652	0.056

*Assuming misclassificaiton probabilities are the same at both levels of the severity; Est.MCR: estimated misclassification probabilities from CTHMM, $(\hat{\pi}_{1|0}, \hat{\pi}_{0|1})$:
$\hat{\pi}_{0|1} = Pr(O^{(t_1)} = 0|S^{(t_1)} = 1), \hat{\pi}_{1|0} = Pr(O^{(t_1)} = 1|S^{(t_1)} = 0)$

4.5 Application

We applied the joint modeling strategy to real data from the IMS III trial. A total of 521 subjects were included in the analysis, all of which had 4 repeated measurements of NIHSS at baseline, and 40 minutes, 24 hours and 5 days after symptom onset. The misclassification probabilities of the

Table 8: Estimated coefficients in the naive model and CTM-MCSIMEX model

Coefficients	Naive Model		CTM-MCSIMEX	
	Estimates	SE	Estimates	SE
Intercept	-0.0977	0.150	-0.0575	0.153
Treatment	0.0668	0.175	0.0633	0.178
Severity	-1.2660	0.196	-1.476	0.220

dichotomized NIHSS (stroke severity) were first estimated via CTHMM with the unit of the time in days, i.e. $\mathbf{t} = \{0, 0.03, 1, 5\}$. According to our previous work, a CTHMM with piecewise constant transition intensity fit the data better; therefore the resulting misclassification probabilities from such CTHMM were used for the bias correction using MCSIMEX model in the second stage. In the primary analysis, the treatment effect was assessed via Cochran-Mantel-Haenszel test, with adjustment for the baseline stroke severity defined according to NIHSS. To be consistent with the primary analysis specified in the Statistical Analysis Plan, only dichotomized NIHSS was adjusted for in the analytic model, in addition to the treatment assignment.

Based on CTHMM, about 3.8% (95%CI: [0.0302, 0.0423]) of the subjects with mild/moderate stroke were misclassified as a severe stroke. The misclassification probability for subjects with severe stroke was almost double that of those with mild/moderate (6.6%, 95%CI: [0.0554, 0.0782]). Table 8 presents the results for the naive model as well as CTM-MCSIMEX estimators. The primary analysis from the IMS III trial was not able to demonstrate the treatment efficacy; using the CTM-MCSIMEX estimator, we confirmed the finding with a smaller estimated coefficient for the treatment and slightly larger standard error of the estimate.

4.6 Discussion

In this paper, we demonstrated the feasibility and the performance of combining CTHMM and MCSIMEX algorithm to correct the estimation bias for the perfectly measured variable (e.g. treatment assignment) caused by covariate misclassification when no gold standard for the misclassified variable is available. The repeated measurements of the misclassified variable provide additional

information for the error process. For data with mild to moderate misclassification (e.g. $< 20\%$), bias correction for the perfect measured variable using this joint modeling strategy is reasonable provided that the estimated misclassification probabilities in the first stage are relatively accurate. However, the bias correction is attenuated significantly when misclassification probabilities in the covariate are greater than 30%. Large amount of bias remains uncorrected even with the CTM-MCSIMEX estimators. In such case, as far as our knowledge, no methods in the literature can provide a good bias correction. On the other hand, if the estimated misclassification probabilities are small, e.g. total error rate $(\pi_{1|0} + \pi_{0|1}) < 5\%$, joint modeling is not necessary. With such small amounts of misclassification of the covariate, the impact on the perfectly measured variable is small. In addition, if data are not misclassified, using CTHMM can still provide estimation of misclassification probabilities, although with small magnitude, which results in overcorrection of the parameter estimation for the perfect measured variable via MCSIMEX model as demonstrated in our simulation.

While the point estimates are improved using joint modeling, the power of the hypothesis testing for the treatment effect is also impacted and is similar to the naïve model without correction. This may partly be due to the simulation procedure in the MCSIMEX algorithm introducing extra variability into the simulated data, resulting in large standard errors while making inference. In addition, the combined approach also leads to a slightly inflated type I error probability. Since the misclassification probabilities used in the MCSIMEX algorithm are estimated from the first stage, the variances of the estimated misclassification probabilities are not fully considered in the second stage even with the increased standard error estimation in the MCSIMEX algorithm. Thus, further research is needed to improve the variance estimation for the parameter estimates, which could incorporate the variability of the estimation for the misclassification probabilities using repeated measurements.

In the simulations, we only demonstrate simple cases where no additional covariates were adjusted for in either CTHMM or MCSIMEX model. However, the joint modeling strategy is flexible. If there is external information about misclassification probabilities, it could be incorporated into

the CTHMM directly. Likelihood ratio tests could be applied to compare if such adjustment is needed. Other information, such as death, could also be included in the CTHMM while estimating the misclassification probabilities (Jackson et al., 2003; Jackson and Sharples, 2002), which may potentially improve the estimation for the misclassification probabilities given the certainty of the death information.

There are some limitations in this study. The assumption that the misclassification probabilities are the same across all time points is strong. It may be the case that the misclassification probabilities decrease over time; that is, the evaluation of the disease severity for individual subjects is improved over time. Therefore, a time varying misclassification matrix may be more appropriate in this case. This will be addressed in future works. Secondly, we only evaluate the performance for the two-state cases; further exploration is needed for the scenario with more states.

In summary, the joint analysis strategy can be applied to adjust the estimation bias of the perfectly measured variable caused by covariate misclassification when repeated measurements of the covariate are available. With mild to moderate misclassification in the covariate, the performance of CTM-MCSIMEX estimator in terms of bias correction is reasonable. With the flexibility of the joint modelling strategy and ease of implementation, it can be applied to more complicate modelling such as survival analysis.

5 Summary and Conclusions

5.1 Conclusions

This work mainly focuses on a common practical issue of covariate misclassification under covariate-adaptive randomization. However, the emphasis surrounds the impact on both the type I error and power of the hypothesis testing and bias of the estimate for the perfectly measured variable (e.g. treatment effect). To understand such impact, we first demonstrate theoretically and through simulation that under covariate-adaptive randomization, type I error can be maintained if the same misclassified covariate is adjusted for both during randomization procedures and in the analysis. However, adjusting for the misclassified covariate in the analysis will lead to power loss and biased estimation for the perfectly measured variable using generalized linear regression. The magnitude of power loss and bias is not ignorable, especially when the covariate effect on the outcome is large relative to the treatment effect, or the misclassification rate is relatively high (Original paper I). Secondly, in order to identify the misclassification error in the covariate, we illustrate the performance of the CTHMM with a focus on estimating misclassification probabilities when a repeatedly measured error-prone covariate is available. We propose to use AIC, which is calculated based on a modified likelihood of estimated parameters given data, to directly compare between models with and without misclassification. We show through simulation that the AIC approach behaves much better, in terms of identifying the correct model for the misclassified data, than the Pearson-type goodness-of-fit test in terms of identifying the correct model for the misclassified data, which is the only formal test currently in the literature (Original manuscript II). Further, we propose a two-stage analysis strategy, which combines the CTHMM and MCSIMEX models to correct the estimation bias for the treatment effect. It provides at least 50% bias correction when there is mild to moderate misclassification in the covariate (e.g. $< 20\%$). The performance is comparable to other methods in the literature. In addition, the two-stage modeling strategy is flexible in terms of dealing with complicated analysis model and easy to implement. Estimates based on CTM-MCSIMEX may provide some insight into the treatment effect when no misclassification error can

be assumed and be used for future study design (Original manuscript III).

5.2 Strengths & Limitations

This work mainly focuses on the estimation of the perfectly measured variable when there is covariate misclassification. It simultaneously takes into consideration the disease development as well as the uncertainty of diagnosis for disease severity. It maximizes the utility of the available information which makes more biological sense, and also accounts for real practice, both clinically and statistically. To the best of our knowledge, no study has emphasized the importance of and methods related to estimation of the perfectly measured variable under misclassification of prognostic covariates.

There are several limitations of this study. First of all, we are not able to deal with covariates that are not ordinal, like race or cancer type, since those variables do not transit between types (states). For simplicity, we only consider two-state scenarios with no additional covariate adjustment in either the CTHMM or the MCSIMEX model in the simulation. In addition, we assume the misclassification probabilities are the same across all time points. This may not be true in real practices; the misclassification probabilities may reduce after the disease has stabilized, and physicians may become more familiar with patient status over time.

5.3 Future Directions

Although we have demonstrated the performance of the CTHMM with a focus on estimating misclassification probabilities and the feasibility of combining the CTHMM and the MCSIMEX models for correcting the bias for the perfectly measured variable, there is still much work to do. First of all, it is reasonable to assume that the misclassification probabilities vary over time. Therefore, a time varying misclassification matrix should be considered in the analysis. This can be done either by adding a random effect or imposing known correlation structures to the misclassification probabilities. The performance of such a model needs to be evaluated, including parameter iden-

tification, point estimation and model diagnosis. In addition, more complicated scenarios such as three or more states, additional covariates on the misclassification probabilities or the outcome, death should be considered and incorporated into analyses. Because those scenarios may provide additional information for parameter estimation, therefore are worth further studying in depth.

6 References

Jr. Adams, H. P., P. H. Davis, E. C. Leira, K. C. Chang, B. H. Bendixen, W. R. Clarke, R. F. Woolson, and M. D. Hansen. Baseline nih stroke scale score strongly predicts outcome after stroke: A report of the trial of org 10172 in acute stroke treatment (toast). *Neurology*, 53(1):126–31, 1999.

R. Aguirre-Hernandez and V. T. Farewell. A pearson-type goodness-of-fit test for stationary and time-continuous markov regression models. *Stat Med*, 21(13):1899–911, 2002. ISSN 0277-6715 (Print) 0277-6715. doi: 10.1002/sim.1152. Aguirre-Hernandez, R Farewell, V T Journal Article England Stat Med. 2002 Jul 15;21(13):1899-911.

Yonghong An, A Texas, Yingyao Hu, and Matt Shum. Identifiability and inference of hidden markov models. *Technical report*, 2013.

S. Atagi, M. Kawahara, A. Yokoyama, H. Okamoto, N. Yamamoto, Y. Ohe, T. Sawa, S. Ishikura, T. Shibata, H. Fukuda, N. Saijo, and T. Tamura. Thoracic radiotherapy with or without daily low-dose carboplatin in elderly patients with non-small-cell lung cancer: a randomised, controlled, phase 3 trial by the japan clinical oncology group (jcog0301). *Lancet Oncol*, 13(7):671–8, 2012.

N. Bartolomeo, P. Trerotoli, and G. Serio. Progression of liver cirrhosis to hcc: an application of hidden markov model. *BMC Med Res Methodol*, 11:38, 2011. ISSN 1471-2288 (Electronic) 1471-2288 (Linking). doi: 10.1186/1471-2288-11-38. Bartolomeo, Nicola Trerotoli, Paolo Serio, Gabriella England BMC Med Res Methodol. 2011 Apr 4;11:38. doi: 10.1186/1471-2288-11-38.

Leonard E. Baum and Ted Petrie. Statistical inference for probabilistic functions of finite state markov chains. *The Annals of Mathematical Statistics*, 37(6):1554–1563, 1966. ISSN 00034851. doi: 10.2307/2238772. URL `http://www.jstor.org/stable/2238772http://www.jstor.org/stable/pdfplus/2238772.pdf?acceptTC=true`.

Ted; Soules George; Weiss Norman. Baum, Leonard E.; Petrie. A maximization technique occurring in the statistical analysis of probabilistic functions of markov chains. *The Annals of Mathematical Statistics*, 41(1), 1970. doi: 10.1214/aoms/1177697196.

Melissa Dowd Begg and Stephen Lagakos. Loss in efficiency caused by omitting covariates and misspecifying exposure in logistic regression models. *Journal of the American Statistical Association*, 88(421):166–170, 1993.

Sandra Blitz, Joanna Baxter, Janet Raboud, Sharon Walmsley, Anita Rachlis, Fiona Smaill, Alex Ferenczy, François Coutlée, Catherine Hankins, Deborah Money, and for the Canadian Women's HIV Study Group. Evaluation of hiv and highly active antiretroviral therapy on the natural history of human papillomavirus infection and cervical cytopathologic findings in hiv-positive and high-risk hiv-negative women. *Journal of Infectious Diseases*, 208(3):454–462, 2013. doi: 10.1093/infdis/jit181. URL `http://jid.oxfordjournals.org/content/208/3/454.abstract`.

Kristy Elizabeth Boyer. The dialogue hmm project. URL `http://people.engr.ncsu.edu/keboyer/dialoguehmm.html`.

D. E. Briggs, R. A. Felberg, M. D. Malkoff, P. Bratina, and J. C. Grotta. Should mild or moderate stroke patients be admitted to an intensive care unit? *Stroke*, 32(4):871–6, 2001.

Joseph Broderick. Interventional management of stroke trial (ims iii): A phase iii clinical trial examining whether a combined intravenous (iv) and intra-arterial (ia) approach to recanalization is superior to standard iv rt-pa (activase®) alone, 2013. URL `http://clinicaltrials.gov/show/NCT00359424NLMIdentifier:NCT00359424.`

J. P. Buonaccorsi, P. Laake, and M. B. Veierod. On the effect of misclassification on bias of perfectly measured covariates in regression. *Biometrics*, 61(3):831–6, 2005.

A. Bureau, S. Shiboski, and J. P. Hughes. Applications of continuous time hidden markov models to the study of misclassified disease outcomes. *Stat Med*, 22(3):441–62, 2003.

T. C. Buter, A. van den Hout, F. E. Matthews, J. P. Larsen, C. Brayne, and D. Aarsland. Dementia and survival in parkinson disease: A 12-year population study. *Neurology*, 70(13): 1017–1022, 2008.

Jeffrey S. Buzas. Measurement error and misclassification in statistics and epidemiology: Impacts and bayesian adjustments by p. gustafson. *Biometrics*, 62(1):307–308, 2006.

R.J. Carroll, D. Ruppert, L.A. Stefanski, and C.M. Crainiceanu. *Measurement Error in Nonlinear Models: A Modern Perspective, Second Edition*. Taylor & Francis, 2006.

T. C. Chalmers, Jr. Smith, H., B. Blackburn, B. Silverman, B. Schroeder, D. Reitman, and A. Ambroz. A method for assessing the quality of a randomized control trial. *Control Clin Trials*, 2(1):31–49, 1981.

Zhijian Chen, Grace Y. Yi, and Changbao Wu. Marginal analysis of longitudinal ordinal data with misclassification in both response and covariates. *Biometrical Journal*, 56(1): 69–85, 2014.

J. R. Cook and L. A. Stefanski. Simulation-extrapolation estimation in parametric measurement error models. *Journal of the American Statistical Association*, 89(428):1314–1328, 1994.

L Costas, C Infante-Rivard, JP Zock, M Van Tongeren, P Boffetta, A Cusson, C Robles, D Casabonne, Y Benavente, and N Becker. Occupational exposure to endocrine disruptors and lymphoma risk in a multi-centric european study. *British journal of cancer*, 2015. ISSN 0007-0920.

R. De Haan, J. Horn, M. Limburg, J. Van Der Meulen, and P. Bossuyt. A comparison of five stroke scales with measures of disability, handicap, and quality of life. *Stroke*, 24 (8):1178–81, 1993.

Peter Diggle, Patrick Heagerty, Kung-Yee Liang, and Scott Zeger. *Analysis of longitudinal data*. Oxford University Press, 2002. ISBN 0191664324.

Bradley Efron. Forcing a sequential experiment to be balanced. *Biometrika*, 58(3):403–417, 1971.

R. J. Ellis, S. Letendre, F. Vaida, R. Haubrich, R. K. Heaton, N. Sacktor, D. B. Clifford, B. M. Best, S. May, A. Umlauf, M. Cherner, C. Sanders, C. Ballard, D. M. Simpson, C. Jay, and J. A. McCutchan. Randomized trial of central nervous system-targeted antiretrovirals for hiv-associated neurocognitive disorder. *Clin Infect Dis*, 58(7): 1015–22, 2014.

M. Ersek, N. Polissar, A. D. Pen, A. Jablonski, K. Herr, and M. B. Neradilek. Addressing methodological challenges in implementing the nursing home pain management algorithm randomized controlled trial. *Clin Trials*, 9(5):634–44, 2012.

L. Fan, S. D. Yeatts, B. J. Wolf, L. A. McClure, M. Selim, and Y. Y. Palesch. The impact of covariate misclassification using generalized linear regression under covariate-adaptive randomization. *Stat Methods Med Res*, 2015. ISSN 0962-2802. doi: 10.1177/0962280215616405.

U. Fischer, M. Arnold, K. Nedeltchev, C. Brekenfeld, P. Ballinari, L. Remonda, G. Schroth, and H. P. Mattle. Nihss score and arteriographic findings in acute ischemic stroke. *Stroke*, 36(10):2121–5, 2005.

K. M. Flegal, C. Brownie, and J. D. Haas. The effects of exposure misclassification on estimates of relative risk. *Am J Epidemiol*, 123(4):736–51, 1986.

M. R. Frankel, L. B. Morgenstern, T. Kwiatkowski, M. Lu, B. C. Tilley, J. P. Broderick, R. Libman, S. R. Levine, and T. Brott. Predicting prognosis after stroke: a placebo group analysis from the national institute of neurological disorders and stroke rt-pa stroke trial. *Neurology*, 55(7):952–9, 2000.

Hironori Fujisawa and Shizue Izumi. Inference about misclassification probabilities from repeated binary responses. *Biometrics*, 56(3):706–711, 2000.

M. H. Gail, S. Wieand, and S. Piantadosi. Biased estimates of treatment effect in randomized experiments with nonlinear regressions and omitted covariates. *Biometrika*, 71(3): 431, 1984.

R. E. Gangnon, K. E. Lee, B. E. Klein, S. K. Iyengar, T. A. Sivakumaran, and R. Klein. Misclassification can explain most apparent regression of age-related macular degeneration: results from multistate models with misclassification. *Invest Ophthalmol Vis Sci*, 55(3):1780–6, 2014.

M. Maria Glymour, Lisa F. Berkman, Karen A. Ertel, Martha E. Fay, Thomas A. Glass, and Karen L. Furie. Lesion characteristics, nih stroke scale, and functional recovery after stroke. *American journal of physical medicine & rehabilitation / Association of Academic Physiatrists*, 86(9):725–733, 2007.

Steven M. Green. Cheerio, laddie! bidding farewell to the glasgow coma scale. *Annals of Emergency Medicine*, 58(5):427–430, 2011.

S. Greenland. The effect of misclassification in the presence of covariates. *Am J Epidemiol*, 112 (4):564–9, 1980.

J. K. Harrison, K. S. McArthur, and T. J. Quinn. Assessment scales in stroke: clinimetric and clinical considerations. *Clin Interv Aging*, 8:201–11, 2013.

Iris M Heid, Claudia Lamina, Helmut Küchenhoff, Guido Fischer, Norman Klopp, Melanie Kolz, Harald Grallert, Caren Vollmert, Stefanie Wagner, and Cornelia Huth. Estimating the single nucleotide polymorphism genotype misclassification from routine double measurements in a large epidemiologic sample. *American journal of epidemiology*, 168(8):878–889, 2008. ISSN 0002-9262.

Daniel J Hopkins and Gary King. A method of automated nonparametric content analysis for social science. *American Journal of Political Science*, 54(1):229–247, 2010. ISSN 1540-5907.

Yanqing; Hu and Feifang Hu. Asymptotic properties of covariate-adaptive randomization. *The Annals of Statistics*, Vol. 40,(No. 3,):1794–1815, 2012.

C. H. Jackson and L. D. Sharples. Hidden markov models for the onset and progression of bronchiolitis obliterans syndrome in lung transplant recipients. *Stat Med*, 21(1): 113–28, 2002.

Christopher H. Jackson, Linda D. Sharples, Simon G. Thompson, Stephen W. Duffy, and Elisabeth Couto. Multistate markov models for disease progression with classification error. *Journal of the Royal Statistical Society: Series D (The Statistician)*, 52(2):193–209, 2003.

Frederick Jelinek. *Statistical Methods for Speech Recognition*. MIT Press, 1997.

P. Jewell John, M. Neuhaus; Nicholas. A geometric approach to assess bias due to omitted covariates in generalized linear models. *Biometrika*, 80(4):807–15, 1993.

D. N. Krag, S. J. Anderson, T. B. Julian, A. M. Brown, S. P. Harlow, J. P. Costantino, T. Ashikaga, D. L. Weaver, E. P. Mamounas, L. M. Jalovec, T. G. Frazier, R. D. Noyes, A. Robidoux, H. M. Scarth, and N. Wolmark. Sentinel-lymph-node resection compared with conventional axillary-lymph-node dissection in clinically node-negative patients with breast cancer: overall survival findings from the nsabp b-32 randomised phase 3 trial. *Lancet Oncol*, 11(10):927–33, 2010.

H. Kuchenhoff, S. M. Mwalili, and E. Lesaffre. A general method for dealing with misclassification in regression: the misclassification simex. *Biometrics*, 62(1):85–96, 2006.

G. Kundt. Comparative evaluation of balancing properties of stratified randomization procedures. *Methods Inf Med*, 48(2):129–34, 2009.

Claudia Lamina, Helmut Kuchenhoff, Jenny Chang-Claude, Bernhard Paulweber, H Wichmann, Thomas Illig, Margret R Hoehe, Florian Kronenberg, and Iris M Heid. Haplotype misclassification resulting from statistical reconstruction and genotype error, and

its impact on association estimates. *Annals of human genetics*, 74(5):452–462, 2010. ISSN 1469-1809.

J. M. Lange and V. N. Minin. Fitting and interpreting continuous-time latent markov models for panel data. *Stat Med*, 32(26):4581–95, 2013. ISSN 1097-0258 (Electronic) 0277-6715 (Linking). doi: 10.1002/sim.5861. URL `http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3795797/pdf/nihms494444.pdf`.

Brian G. Leroux. Maximum-likelihood estimation for hidden markov models. *Stochastic Processes and their Applications*, 40(1):127–143, 1992. ISSN 0304-4149. doi: http://dx.doi.org/10.1016/0304-4149(92)90141-C. URL `http://www.sciencedirect.com/science/article/pii/030441499290141C`.

Tong Li. Robust and consistent estimation of nonlinear errors-in-variables models. *Journal of Econometrics*, 110(1):1–26, 2002.

Tong Li and Cheng Hsiao. Robust estimation of generalized linear models with measurement errors. *Journal of Econometrics*, 118(1–2):51–65, 2004.

S. Luo, W. Chan, M. A. Detry, P. J. Massman, and R. S. Doody. Binomial regression with a misclassified covariate and outcome. *Stat Methods Med Res*, 2012.

R. H. Lyles and J. Lin. Sensitivity analysis for misclassification in logistic regression via likelihood methods and predictive value weighting. *Stat Med*, 29(22):2297–309, 2010.

Riccardo E Marioni, Ardo van den Hout, Michael J Valenzuela, Carol Brayne, and Fiona E Matthews. Active cognitive lifestyle associates with cognitive recovery and a reduced risk of cognitive decline. *Journal of Alzheimer's Disease*, 28(1):223, 2012. ISSN 1387-2877.

J. P. Matts and J. M. Lachin. Properties of permuted-block randomization in clinical trials. *Control Clin Trials*, 9(4):327–44, 1988.

Aurélie Mayet, Stéphane Legleye, Bruno Falissard, and Nearkasen Chau. Cannabis use stages as predictors of subsequent initiation with other illicit drugs among french adolescents: Use of a multi-state model. *Addictive Behaviors*, 37(2):160–166, 2012.

D. Moher, S. Hopewell, K. F. Schulz, V. Montori, P. C. Gotzsche, P. J. Devereaux, D. Elbourne, M. Egger, and D. G. Altman. Consort 2010 explanation and elaboration: updated guidelines for reporting parallel group randomised trials. *Bmj*, 340:c869, 2010.

K. W. Muir, C. J. Weir, G. D. Murray, C. Povey, and K. R. Lees. Comparison of neurological scales and scoring systems for acute stroke prognosis. *Stroke*, 27(10):1817–20, 1996.

Sam Norton, Fiona E Matthews, and Carol Brayne. A commentary on studies presenting projections of the future prevalence of dementia. *BMC Public Health*, 13(1):1, 2013. ISSN 1471-2458.

W. Pan, X. Lin, and D. Zeng. Structural inference in transition measurement error models for longitudinal data. *Biometrics*, 62(2):402–12, 2006.

Wenqin Pan, Donglin Zeng, and Xihong Lin. Estimation in semiparametric transition measurement error models for longitudinal data. *Biometrics*, 65(3):728–736, 2009.

T. Petrie. Probabilistic functions of finite state markov chains. *The Annals of Mathematical Statistics*, 40(1):97–115, 1969. ISSN 00034851. URL `http://www.jstor.org/stable/2239201`.

Lawrence R. Rabiner. *A tutorial on hidden Markov models and selected applications in speech recognition*, pages 267–296. Morgan Kaufmann Publishers Inc., 1990.

L.R. Rabiner and B.H. Juang. *Fundamentals of Speech Recognition*. Pearson Education, 1993.

S. J. Reade-Christopher and L. L. Kupper. Effects of exposure misclassification on regression analyses of epidemiologic follow-up study data. *Biometrics*, 47(2):535–48, 1991.

Laurence D. Robinson and Nicholas P. Jewell. Some surprising results about covariate adjustment in logistic regression models. *International Statistical Review / Revue Internationale de Statistique*, 59(2):227–240, 1991.

Lindsay BG Roeder K, Carroll RJ. Semiparametric mixture approach to case-control studies with errors in covariables. *Journal of the American Statistical Association*, 91:722–32, 1996.

R. J. Rosychuk and M. E. Thompson. Bias correction of two-state latent markov process parameter estimates under misclassification. *Stat Med*, 22(12):2035–55, 2003. ISSN 0277-6715 (Print) 0277-6715. doi: 10.1002/sim.1473. Rosychuk, Rhonda J Thompson, Mary E Journal Article Research Support, Non-U.S. Gov't England Stat Med. 2003 Jun 30;22(12):2035-55.

Rhonda J. Rosychuk and Shofiqul Islam. Parameter estimation in a model for misclassified markov data - a bayesian approach. *Comput. Stat. Data Anal.*, 53(11):3805–3816, 2009. ISSN 0167-9473. doi: 10.1016/j.csda.2009.04.002.

Thompson ME. Rosychuk RJ. Parameter identifiability issues in a model for possibly misclassified binary alternating processes. *American Statistical Association 1998 Proceedings of the Biometrics Section 1999;*, pages 86–91, 2004.

Daniel W. Schafer. Semiparametric maximum likelihood for measurement error model regression. *Biometrics*, 57(1):53–61, 2001.

D. Schlegel, S. J. Kolb, J. M. Luciano, J. M. Tovar, B. L. Cucchiara, D. S. Liebeskind, and S. E. Kasner. Utility of the nih stroke scale as a predictor of hospital disposition. *Stroke*, 34(1):134–7, 2003.

D. J. Schlegel, D. Tanne, A. M. Demchuk, S. R. Levine, and S. E. Kasner. Prediction of hospital disposition after thrombolysis for acute ischemic stroke using the national institutes of health stroke scale. *Arch Neurol*, 61(7):1061–4, 2004.

J. Shao and X. Yu. Validity of tests under covariate-adaptive biased coin randomization and generalized linear models. *Biometrics*, 69(4):960–9, 2013.

Michelle Shardell and Gregory E. Hicks. Statistical analysis with missing exposure data measured by proxy respondents: a misclassification problem within a missing-data problem. *Statistics in Medicine*, pages n/a–n/a, 2014.

C. Sherrington, S. R. Lord, C. M. Vogler, J. C. Close, K. Howard, C. M. Dean, G. Z. Heller, L. Clemson, S. D. O'Rourke, E. Ramsay, E. Barraclough, R. D. Herbert, and R. G. Cumming. A post-hospital home exercise program improved mobility but increased falls in older people: a randomised controlled trial. *PLoS One*, 9(9):e104412, 2014.

S. Sinha and Y. Ma. Semiparametric analysis of linear transformation models with covariate measurement errors. *Biometrics*, 70(1):21–32, 2014.

Elizabeth H Slate and Dipankar Bandyopadhyay. An investigation of the mc-simex method with application to measurement error in periodontal outcomes. *Statistics in medicine*, 28(28):3523–3538, 2009. ISSN 1097-0258.

Richard L. Smith. Properties of biased coin designs in sequential clinical trials. pages 1018–1034, 1984.

Donna Spiegelman, Bernard Rosner, and Roger Logan. Estimation and inference for logistic regression with covariate misclassification and measurement error in main study/validation study designs. *Journal of the American Statistical Association*, 95(449):51–61, 2000.

Michael J. Sweeting, Daniela De Angelis, Keith R. Neal, Mary E. Ramsay, William L. Irving, Mark Wright, Lisa Brant, and Helen E. Harris. Estimated progression rates in three united kingdom hepatitis c cohorts differed according to method of recruitment. *Journal of Clinical Epidemiology*, 59(2):144–152, 2006.

Norah A. Terrault, Kelly Im, Ross Boylan, Peter Bacchetti, David E. Kleiner, Robert J. Fontana, Jay H. Hoofnagle, and Steven H. Belle. Fibrosis progression in african americans

and caucasian americans with chronic hepatitis c. *Clinical Gastroenterology and Hepatology*, 6(12):1403–1411, 2008.

Peter F. Thall and J. Kyle Wathen. Covariate-adjusted adaptive randomization in a sarcoma trial with multi-stage treatments. *Statistics in Medicine*, 24(13):1947–1964, 2005.

A. C. Titman and L. D. Sharples. Model diagnostics for multi-state models. *Stat Methods Med Res*, 19(6):621–51, 2010.

Andrew C. Titman and Linda D. Sharples. A general goodness-of-fit test for markov and hidden markov models. *Statistics in Medicine*, 27(12):2177–2195, 2008. ISSN 1097-0258. doi: 10.1002/sim.3033.

R. Toorawa, M. Adena, M. Donovan, S. Jones, and J. Conlon. Use of simulation to compare the performance of minimization with stratified blocked randomization. *Pharm Stat*, 8(4):264–78, 2009.

Ardo Van Den Hout, Carol Jagger, and Fiona E. Matthews. Estimating life expectancy in health and ill health by using a hidden markov model. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 58(4):449–465, 2009.

C. Y. Wang and Margaret Sullivan Pepe. Expected estimating equations to accommodate covariate measurement error. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, 62(3):509–524, 2000.

C. Y. Wang and X. Song. Expected estimating equations via em for proportional hazards regression with covariate misclassification. *Biostatistics*, 14(2):351–65, 2013. ISSN 1465-4644. doi: 10.1093/biostatistics/kxs046. URL `http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3590925/pdf/kxs046.pdf`.

Naisyin Wang, R. J. Carroll, and Kung-Yee Liang. Quasilikelihood estimation in measurement error models with correlated replicates. *Biometrics*, 52(2):401–411, 1996.

Wang S. Wang CY. Semiparametric methods in logistic regression with measurement error. *Statistica Sinica*, 7:1103–20, 1997.

C. J. Weir and K. R. Lees. Comparison of stratification and adaptive methods for treatment allocation in an acute stroke clinical trial. *Stat Med*, 22(5):705–26, 2003.

I. White, C. Frost, and S. Tokunaga. Correcting for measurement error in binary and continuous variables using replicates. *Stat Med*, 20(22):3441–57, 2001.

David Wright. Phase 3 clinical trial to determine if progesterone along with standard medical care for brain injury is more effective at limiting the amount of damage cause by a traumatic brain injury than standard medical care alone., 2014. URL `http://clinicaltrials.gov/show/NCT00822900,NLMIdentifier:NCT00822900.`

M. L. Ybarra, J. S. Holtrop, T. L. Prescott, M. H. Rahbar, and D. Strong. Pilot rct results of stop my smoking usa: a text messaging-based smoking cessation program for young adults. *Nicotine Tob Res*, 15(8):1388–99, 2013.

Byung-Jun Yoon. Hidden markov models and their applications in biological sequence analysis. *Current Genomics*, 10(6):402–415, 2009.

M. Zelen. The randomization and stratification of patients to clinical trials. *J Chronic Dis*, 27 (7-8):365–75, 1974.