

Spring 5-2020

SAMPLE SIZE CALCULATIONS FOR LONGITUDINAL MEDIATION ANALYSIS WITH CONTINUOUS-TIME MARKOV CHAIN VARIABLES

ABIGAIL CHRISTINE SEDORY
UTHealth School of Public Health

Follow this and additional works at: https://digitalcommons.library.tmc.edu/uthsph_dissertsopen



Part of the [Community Psychology Commons](#), [Health Psychology Commons](#), and the [Public Health Commons](#)

Recommended Citation

SEDORY, ABIGAIL CHRISTINE, "SAMPLE SIZE CALCULATIONS FOR LONGITUDINAL MEDIATION ANALYSIS WITH CONTINUOUS-TIME MARKOV CHAIN VARIABLES" (2020). *UT School of Public Health Dissertations (Open Access)*. 116.

https://digitalcommons.library.tmc.edu/uthsph_dissertsopen/116

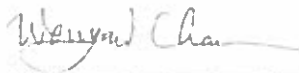
This is brought to you for free and open access by the School of Public Health at DigitalCommons@TMC. It has been accepted for inclusion in UT School of Public Health Dissertations (Open Access) by an authorized administrator of DigitalCommons@TMC. For more information, please contact digitalcommons@library.tmc.edu.

SAMPLE SIZE CALCULATIONS FOR LONGITUDINAL MEDIATION ANALYSIS
WITH CONTINUOUS-TIME MARKOV CHAIN VARIABLES

by

ABIGAIL CHRISTINE SEDORY, MS

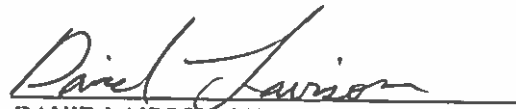
APPROVED:



WENYAW CHAN, PHD



MICHAEL SWARTZ, PHD



DAVID LAIRSON, PHD



DEAN, THE UNIVERSITY OF TEXAS
SCHOOL OF PUBLIC HEALTH

Copyright
by
Abigail Sedory, MS, PhD
2020

SAMPLE SIZE CALCULATIONS FOR LONGITUDINAL MEDIATION ANALYSIS
WITH CONTINUOUS-TIME MARKOV CHAIN VARIABLES

by

ABIGAIL CHRISTINE SEDORY
BS, Southwestern University, 2012
MS, UT Health, 2016

Presented to the Faculty of The University of Texas

School of Public Health

in Partial Fulfillment

of the Requirements

for the Degree of

DOCTOR OF PHILOSOPHY

THE UNIVERSITY OF TEXAS
SCHOOL OF PUBLIC HEALTH
Houston, Texas
May, 2020

ACKNOWLEDGEMENTS

First, I would like to thank my parents, Dr. Stephen and Mrs. Adline Sedory, for all of their love, inspiration, and support throughout my entire academic journey. Thank you to my sister, Sarah, and all of my extended family for always being there as source of strength and love. A special thank you goes out to all my colleagues, friends, and family friends who have helped me get through this entire process with their encouragement and support. Thank you to all the faculty and staff I have had the pleasure of working with, who aided in the progression of my education. A special thank you to my academic advisor, Dr. Swartz, who has been a prominent figure throughout both my Masters and Doctorate programs, not only for his support academically, but also emotionally. Thank you also to Dr. Chan for guiding me through the entire dissertation process and for always being available for anything I needed in both my graduate programs. Thank you to Dr. Lairson who served as minor advisor in Health Economics. I also would like to acknowledge the pandemic that progressed the completion of this dissertation. Finally, I would like to acknowledge the funding received for this program was through University of Texas Health Science Center at Houston School of Public Health, Training Program in Biostatistics National Institute of General Medical Sciences/NIH under Grant T32GM074902.

SAMPLE SIZE CALCULATIONS FOR LONGITUDINAL MEDIATION ANALYSIS WITH CONTINUOUS-TIME MARKOV CHAIN VARIABLES

Abigail Christine Sedory, MS, PhD
The University of Texas
School of Public Health, 2020

Dissertation Chair: Wenyaw Chan, PhD

Mediation is a type of analysis used to determine the causal mechanism linking a predictor and an outcome through a mediator variable. Various research has examined the inclusion of different variable types for the predictor, mediator, and outcome. However, no studies include the presence of a continuous-time Markov chain (CTMC) as any of the components in a mediating model. If researchers wanted to design a study with a CTMC in the mediating process, one of the first steps would be to determine the minimum number of subjects or observations needed to detect a significant mediating effect. Therefore, in this study, we used simulations to determine that minimum sample size to achieve 80% power for a longitudinal mediation analysis that includes a two-state CTMC as one of the variables in the mediating model. We examined three mediation models with the following variable types: 1) A CTMC outcome with a binary predictor and continuous mediator, 2) a CTMC mediator with a binary predictor and continuous outcome, and 3) a CTMC predictor with continuous mediator and outcome. We calculated the power in simulations where we varied the sample size and effect sizes used to calculate the overall mediating effect. We found that all models required minimum sample sizes that ranged from 100 to 500 observations.

TABLE OF CONTENTS

List of Tables	i
List of Figures	ii
Background	1
Mediation Analysis	1
Sample Size calculations.....	7
Markov Chains.....	8
Public Health Significance.....	9
Specific Aims.....	11
Journal Article 1	12
Title: Sample Size Analysis For Longitudinal Mediation Analysis With A Two-State Continuous-Time Markov Chain Outcome	12
To be submitted to Journal of Simulation.....	12
Introduction.....	12
Methods.....	14
Results.....	18
Discussion	19
References	21
Journal Article 2	25
Title: Sample Size Analysis For Longitudinal Mediation With A Two-State Continuous Time Markov Chain Mediator	25
To be submitted to Journal of Simulation.....	25
Introduction.....	25
Methods.....	27
Results.....	30
Discussion	31
References	33
Journal Article 3	39
Title: Sample Size Analysis for Longitudinal Mediation with a Two-State Continuous Time Markov Chain Predictor	39
To be submitted to Journal of Simulation.....	39
Introduction.....	39
Methods.....	41

Results.....	43
Discussion.....	44
References.....	46
Conclusion	51
References.....	52

LIST OF TABLES

Table 1.1. Power calculations from mediation analysis with CTMC outcome. The effect sizes of 0.26 and 0.39 were varied for effect a and b from mediation regression models.....	19
Table 2.1. Power calculations for the mediating effect when the effect size of X on M(t) was set to 0.26 for various sample sizes and transition effect size.....	31
Table 2.2. Power calculations for the mediating effect when the effect size of X on M(t) was set to 0.39 for various sample sizes and transition effect size.....	31
Table 3.1. Power calculations when effect size $b=0.15$. The total mediating effect (TME) and the specific effect sizes of the components of the mediating effect are recorded.....	43
Table 3.2. Power Calculations when effect size $b=0.26$. The total mediating effect (TME) and the specific effect sizes of the components of the mediating effect are recorded.....	44

LIST OF FIGURES

Figure 1a. Total Effect of X on Y	2
Figure 1b. Simple cross-sectional mediation model, where variable M mediates the effect of X on Y	2
Figure 2. Cross-lagged panel modelling for a longitudinal mediation analysis	4

BACKGROUND

Mediation Analysis

Mediation analysis has been on the rise in psychological research primarily due to the introduction of a mechanism (mediating process) used to explain the process of an organism's response to a stimulus (Hebb, 1958). The concept of mediation is the premise of social psychology theories, such as the Theory of Reason Action and Theory of Planned Behavior, which assumes not only can attitudes affect behavior, but also that attitudes affect intentions, which in turn can affect behavior; therefore, in this case, intentions serve as a mediator between attitudes and behavior (Fishbein & Ajzen, 1975). Researchers in prevention and treatment typically target a causally related mediating variable of an outcome, instead of targeting the outcome variable directly. For example, in some smoking cessation interventions, intention to smoke serves as a mediator between the intervention and reduction of tobacco use (Gonzalvez, Morales, Orgiles, Sussman, & Espada, 2018). The diverse use of mediation analysis includes studies of alcohol consumption, as well as policy interventions, and financial and market performance (Jones-Webb & Karriker-Jaffe, 2013; Keele, Tingley, & Yamamoto, 2015; Semrau & Sigmund, 2012; Voola, Casimir, Carlson, & Anushree Agnihotri, 2012). All of these studies are searching for a link of causality in their respective fields.

In these types of three-variable systems, the mediator serves as a causal link between the predictor and outcome variable. A mediating variable differs in from a moderating variable in the sense that a moderating variable differentiates effects, but does not provide

information on causality (Baron & Kenny, 1986). The simplest mediation model is the cross-sectional model illustrated in Figure 1.

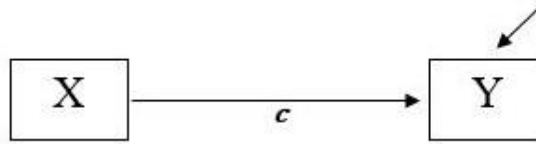


Figure 1a. Total Effect of X on Y

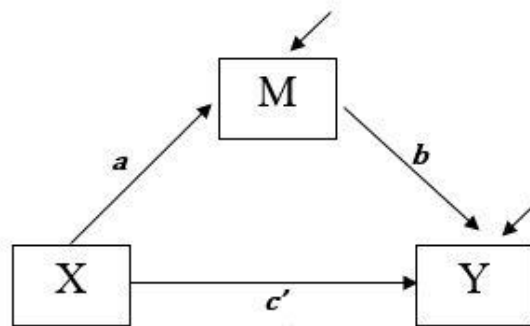


Figure 1b. Simple cross-sectional mediation model, where variable M mediates the effect of X on Y

The total effect of a predictor variable and outcome variable is illustrated in Figure 1a.

In the mediation model, the predictor variable (X) affects the mediator (M) which in turn affects the outcome (Y) (Figure 1b). The idea is that some variable, M, facilitates the effect of X on Y. In order to justify the presence of a mediating effect, Baron and Kenny suggested a causal-steps approach (Baron & Kenny, 1986). A mediating effect is present if the following four stipulations hold:

1. The total effect of X on Y must be significant (c)
2. The effect of X on M must be significant (a)
3. The effect of M on Y controlled for X must be significant (b)
4. The direct effect of X on Y adjusted for M must be non-significant (c')

When all variables are normally distributed, the mediation effect or indirect effect is the calculated difference between the total effect of X on Y (c) and the direct effect of X on Y (c'). However, the indirect effect is more commonly calculated by multiplying the effect of X on M (a) and the effect of M on Y (b), (i.e. $c' = ab$); this method is known as the product of coefficients method. The lone arrows pointing to M and Y represent additional explanatory variables that can be included in the model as controls.

Mathematically, the total effect of X on Y (c) is calculated through regression techniques. For example, using linear regression and treating all variables as continuous random variables, we would use the following equations to estimate the effects a , b and c above. The direct effect is modeled by:

$$Y_i = i_1 + cX_i + e_{1i}$$

The mediating effect is calculated using the following two equations:

$$Y_i = i_2 + c'X_i + bM_i + e_{2i}$$

$$M_i = i_3 + aX_i + e_{3i}$$

This type of cross-sectional analysis serves as a theoretical description of causality not occurring over time. Therefore, methods have been developed that utilize longitudinal data in order to detect a mediating effect and strengthen causal inference. The most common of these longitudinal methods include the cross-lagged panel model (CLPM), the latent growth

curve model (LGM), and the latent change score (LCS) model (MacKinnon, David, 2012; Newsom, Jones, & Hofer, 2013). The most popular longitudinal model is the CLPM (Figure 2).

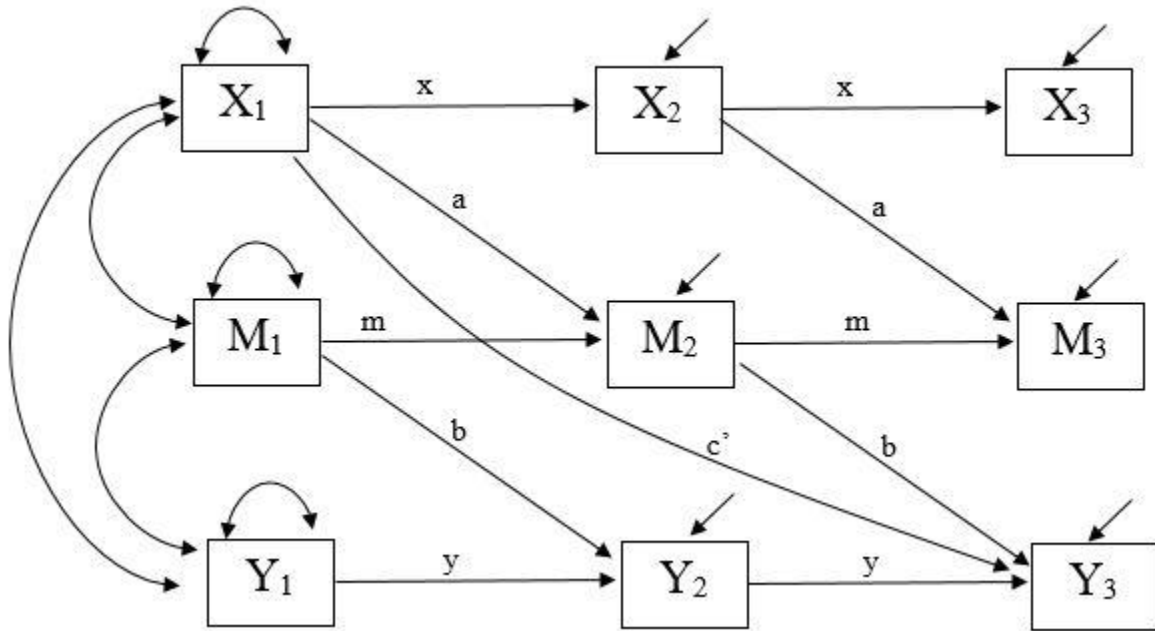


Figure 2. Cross-lagged panel modelling for a longitudinal mediation analysis

Panel methods assess causality in a sequential matter. Therefore, a predictor variable at time 1 (X_1) affects a mediator variable at time 2 (M_2) which in turn affects the outcome at time 3 (Y_3). The direct effect ($X_1 \rightarrow Y_3$) is again labelled as c' and the indirect effect is the product of a ($X_1 \rightarrow M_2$) and b ($M_2 \rightarrow Y_3$). Autoregressive effects are also included in the model. For example, the mediator variable assessed at time 1 for an individual may also have an effect on the mediator variable at time 2 and the outcome variable at time 2 may also affect the outcome at time 3. The inclusion of these autoregressive effects aid in identifying

whether the individual differences in a variable are stable over time (MacKinnon, David P., Fairchild, & Fritz, 2007). In order to interpret and generalize the results of the mediating (indirect) effects over time in a CLPM, not only is the assessment of stability important, but also stationarity (causal structure is stable) and equilibrium (cross-sectional variances and covariances are stable). CLPM often require a minimum of three repeated measures for each variable (data over three time points) in order to estimate a mediating effect. There are models (half-longitudinal and latent longitudinal models) that assess the mediating effect with data having observations from only two time points. However, having fewer repeated measures reduces the ability to assess the degree of stationarity in the half-longitudinal design and the degree of stability and equilibrium in the latent longitudinal design.

Often times in panel modelling for mediation analysis, data are analyzed treating time as discrete. However, the time elapsed between observations, commonly referred to as “lag”, does play a role in not only the effect sizes but the interpretation and generalizability as well (Cole & Maxwell, 2003; Gollob & Reichardt, 1991; Maxwell, Cole, & Mitchell, 2011). Assessing the effects at multiple lags and comparing models can strengthen the understanding and interpretation of parameter estimates (Selig, Preacher, & Little, 2012). Another more novel modelling approach to CLPM are continuous time models (CTM) in which the time between repeated measures is treated as continuous instead of discrete (Deboeck & Preacher, 2016; MacKinnon et al., 2007). By assuming continuous time interactions in mediation analyses, CTMs eliminate the dependency of lag in resulting effect sizes and estimates. CTMs provide lag independent estimates that not only have an

interpretation that more nearly matches the theory of the mediation process, but CTMs can be used in studies when the time between observations is not equally spaced.

Most studies in psychology treat the predictor, mediator, and outcome variables in the mediation process as continuous normally distributed variables, but researchers are branching out to include other types of variables, such as categorical variables. Recently, medicinal and epidemiological studies are including categorical variables for X, M, and/or Y. The use and development of mediation models with categorical mediators and outcomes has grown more abundantly since 2010 (Preacher 2015). The inclusion of categorical variables, specifically for M and Y, in mediation analysis utilizes Generalized Linear Models (GLMs) in order to calculate direct and indirect effects (Huang, Sivaganesan, Succop, & Goodman, 2004). Logistic or probit regression techniques are used for binary mediators and outcomes. In these models, studies show that interpreting the indirect and direct effect in terms of the odds ratio is beneficial as well (Buis, 2010; Vanderweele, T. J. & Vansteelandt, 2010). Although little effort has been placed towards investigating the use of mediation models with count data, Poisson regression techniques can be used when M or Y is a count variable (Coxe & MacKinnon, 2010; Imai, Keele, & Tingley, 2010a; Valeri & VanderWeele, 2013).

As previously mentioned, the mediating effect is often calculated as the product of the effect of X on M and the effect of M on Y. The three most common methods to test for the significance of a mediation effect in longitudinal studies are the joint significance test, normal approximation method, and test of b. In the joint significance test we are jointly testing whether the estimates of a and b are different from zero and assume significance if both hypotheses are rejected. In this method, we are assuming the estimates are each

asymptotically normal and not concerned with the actual distribution of the mediation effect estimates, the product of a and b , which could have a complex distribution. In the normal approximation method, or Sobel's test, we assume that the distribution of the product estimate of ab is asymptotically normal with mean ab and a variance derived using the multivariate delta method. With this variance, we can derive a confidence interval or conduct a significance test for the mediating effect. The last test commonly used is the test of b . In this method, researchers may be sure of an effect of X on M and can assume that a is different from zero. Therefore, we only need to test the significance of parameter estimate b from zero. However, in most cases, the distribution of the product estimate is not easily calculable and therefore a significance test or confidence intervals are more difficult to create. Bootstrapping techniques are commonly used in practice to find confidence intervals when analyzing data having estimates with unknown distributions (Tibshirani & Efron, 1993). Although this method is more applicable retrospectively in studies after data collection and not practical in the design stage, due to the complexity of the distribution of the product of two random variables, researchers may choose bootstrapping over any of the other methods previously mentioned.

Sample Size calculations

Researchers may be interested in designing an experiment that can investigate and test the presence and/or significance of a mediating effect. In this situation, it is important to determine the minimum sample size required to detect a mediation effect with a specified power. Several studies have demonstrated different ways to calculate the sample and power needed to detect a mediating or indirect effect (Matthew S. Fritz & David P. MacKinnon,

2007; Pan, Liu, Miao, & Yuan, 2018; Vittinghoff, E., Sen, & McCulloch, 2009; Vittinghoff, Eric & Neilands, 2015; Wang & Xue, 2016). In some studies, researchers developed a closed form formula for calculating sample size and power, but in a majority of sample size papers, researchers use bootstrapping to determine the optimal sample size (Matthew S. Fritz & David P. MacKinnon, 2007). Not only do these studies review different methods for calculating the mediating effect, but some also consider different types of regression such as logistic, Poisson, and Cox models (Vittinghoff et al., 2009). The sample size needed to detect an indirect effect for mediation models with continuous and binary predictors and mediators as well as binary, count, and survival outcomes has also been studied (Vittinghoff & Neilands, 2015). However, little research use a mediation pathway that treats any of the variables as a Markov Chain.

Markov Chains

Stochastic modelling and the use of Markov chains in various studies has been growing, particularly in the field of economics and epidemiology. For example, in clinical studies Markov Chains can help analyze time-to-event or survival data (Abner, Charnigo, & Kryscio, 2013). The understanding and prediction of stage of progression of a disease, such as liver disease, advances with the utilization of Markov chain design studies (Tada et al., 2018). Economic evaluation of health care programs can involve analysis with Markov chains as well (Larsen & Turkensteen, 2014; Sato & Zouain, 2010). Health promotion interventions can utilize a Markov chain model, for example, when analyzing the changes of stage in a trans-theoretical model (Ma, Chan, Tsai, Xiong, & Tilley, 2015; Ma, Chan, & Tilley, 2018; Mhoon, Chan, Del Junco, & Vernon, 2010). However, the use of Markov

chains in mediation analysis has not been widely studied. Aalen et al. used a stochastic modelling approach in mediation analysis (Aalen, Gran, Røysland, Stensrud, & Strohmaier, 2018). However, Aalen et al. treats the entire mediation process as a Markov chain as opposed to examining the mediating effect of a variable on the transition rates of a Markov outcome or whether the transition rates of a Markov chain serve as a mediator between two variables.

Public Health Significance

Mediation analysis uses statistics to determine the causal mechanism between two variables (Baron & Kenny, 1986; Matthew S. Fritz & David P. MacKinnon, 2007). Public health researchers can use mediation analysis to provide insight into the causal pathway of an intervention on its desired outcome by determining mediators and comparing the natural indirect and direct effects. With the use of mediation analysis, researchers can identify the true effect of an intervention on the outcome of interest, by determining whether the intervention directly targets the outcome or if a variable exists which serves as a link or mediator between the predictor and outcome. For example, Nguyen et al examined mediators of an alcohol intervention such as adolescent self-control, adolescent-reported parental rules about alcohol, parent-reported rules, adolescent attitudes about alcohol and parent attitudes about adolescent drinking (Nguyen, Webb-Vargas, Koning, & Stuart, 2016) . In this study, researchers were interested in determining how the intervention may affect a subject's drinking status at follow-up and whether the intervention effect is mediated by other variables. Thus by using mediation analysis we can make causal inferences on how an intervention effects a desired outcome. Having this knowledge is important because there

may be situations where an intervention may not directly affect the desired outcome fully. In this instance, we would be interested in determining how the intervention works and using those results to further and possibly improve the intervention in order to produce our desired outcome.

In addition to mediation analysis, the utilization of Continuous-time Markov chains is on the rise. Treating certain variables as Markov-chains can give us insight and information into the dynamic changes of the process as opposed examining whether or not a subject changed states. For example, suppose we are interested in an alcohol intervention for adolescents such as the study above, where the outcome of interest is whether a subject is a heavily weekly drinker or not. We could use logistic regression and examine the effect of the intervention and additional variables on the probability that a subject heavily drinks. However, if we treat the outcome as a Markov chain, we can further examine the effect of those factors on the transition rate, transition probability, and sojourn time in a specific state. For instance, we could determine the probability of a subject transitioning from a heavy weekly drinker to not a heavy drinker and the average time a subject remains in a heavy drinking state before transitioning.

Finally, as an initial step in any experimental design, we want to determine the sample size required in order to achieve a specified power for our statistical tests. By determining our desired sample size and power, we can then begin recruiting the required sample of that size. Unfortunately, most studies do not have an unlimited budget or resource supply, therefore thus having an idea of the number of subjects required can improve the efficiency of our experiment by reducing unnecessary costs and efforts. For this reason, in a

longitudinal study where we are following subjects prospectively over a time-period, we would want to follow the minimum number of people required to detect a difference or effect.

In this study, we will examine various situations using mediation analysis with continuous time Markov chain variables and determine a desired sample size required for a specific power. By combining the use of mediation analysis and Markov chains, we can gain even further insight into the casual mechanisms of an intervention or a predictor on some outcome and use this knowledge to further public health research. Determining a required sample size prior to an experiment or observational study can minimize the waste of resources by saving both time and money.

Specific Aims

The following are the specific aims of this project:

1. Determine the sample size needed to find a mediating effect in a longitudinal mediation analysis with an outcome variable that is a continuous time Markov chain, a continuous mediator and binary predictor.
2. Determine the sample size needed to find a mediating effect in a longitudinal mediation analysis with a mediator variable that is a continuous time Markov chain and a binary predictor and outcome.
3. Determine the sample size needed to find a mediating effect in a longitudinal mediation analysis with a predictor that is a continuous-time Markov chain, a continuous mediator, and binary outcome.

Journal Article 1

Title: Sample Size Analysis For Longitudinal Mediation Analysis With A Two-State

Continuous-Time Markov Chain Outcome

To be submitted to Journal of Simulation

Introduction

Mediation analysis is the latest topic of interest in causal research, especially in treatment and prevention (Imai, Keele, & Tingley, 2010; Preacher, 2015; VanderWeele, 2016). Researchers in this area could find that an intervention does not directly affect the desired outcome, but instead discover a causal variable serving as a mediator between the intervention and outcome variable. Mediation analysis examines those indirect effects of a predictor variable on an outcome variable through a mediator variable as well as the direct effect of a predictor on the outcome. For example, in some smoking cessation interventions, the intervention did not directly target the reduction of tobacco use, but instead targeted mediating the individual's intention to smoke to reduce tobacco use (Gonzalvez, Morales, Orgiles, Sussman, & Espada, 2018). Not only is mediation analysis used to aid in identifying an intervention's target, but it is also useful to help explain the mechanism and connections between two variables. For example, other smoking cessation studies found that the link between smoking cessation and social disadvantage is mediated through momentary smoking context and daily stress (Jahnel, Ferguson, Shiffman, Thrul, & Schüz, 2018; Jahnel, Ferguson, Shiffman, & Schuz, 2019).

In addition to the growth of mediation analysis, the use of Markov chain models, particularly in smoking cessation studies, is growing (Killeen, 2011; Minard, Chan, Wetter, & Etzel, 2012). Researchers in smoking cessation studies not only use Markov chain models to examine transitions between abstinence, relapse and other states as a mechanism for understanding and intervention, but also as an aid in cost-benefit analysis (Baker et al., 2018). However, there is little research of using Markov Chain regression in mediation analysis.

Studies have looked at the mediating effect of several variables on an intervention and smoking cessation (Bandiera, Atem, Ma, Businelle, & Kendzor, 2016a; Hajek et al., 2018; Hoepfner, Hoepfner, & Abroms, 2017; Li et al., 2015). Suppose we wish to design an experiment that examines the mediating effect of these variables on the transition rate between smoking and abstinence, instead of the outcome of smoking cessation or tobacco intake. This information could give researchers a greater insight into the mechanisms involved in smoking cessation. The first step in designing such an experiment is to determine the minimum required sample size to achieve at least 80% or possibly 90% power. Sample size calculations for studies involving mediation analysis are limited to the inclusion of only specific types of variables, not including Markov Chain. In some studies, a closed form solution to calculate sample size is available, but many studies use a bootstrapping technique in order to calculate power and minimum sample size (Matthew S. Fritz & David P. MacKinnon, 2007; Pan, Liu, Miao, & Yuan, 2018; Schoemann, Boulton, & Short, 2017; Vittinghoff, E., Sen, & McCulloch, 2009; Vittinghoff, Eric & Neilands, 2015; Wang & Xue, 2016). To determine the significance of the mediating effect in order to calculate power,

researchers create a confidence interval based on bootstrapped simulations and check for the inclusion of zero. Therefore, in our study, we used similar techniques to determine the minimum required sample size for a mediation analysis involving a two-state continuous time Markov chain outcome with a binary predictor and non-time varying continuous mediator variable.

Methods

In a longitudinal mediation analysis, we treated the outcome variable, Y , as a two-state continuous time Markov Chain. The explanatory variable, X , we treated as a binary variable and the mediator variable, M , we treated as a non-time varying continuous random variable. Therefore, we will use a continuous-time Markov chain regression model of Y as part of our mediation model.

Continuous-time Markov Chain regression model

Let $\{Y(t): 0 \leq t \leq \infty\}$ be a stochastic process describing the state of a process at time t . The random variable $Y(t)$ is defined as a finite state continuous-time Markov chain if for all $t \geq 0$, $s \geq 0$, and $i, j \in S$, where S represents a discrete state space,

$$\begin{aligned} \Pr(Y(t+s) = j | Y(s) = i, Y(u) = y(u), 0 \leq u \leq s) \\ = \Pr(Y(t+s) = j | Y(s) = i) = p_{ij}(t) \quad i, j = 1, \dots, k \end{aligned}$$

where $p_{ij}(t)$ represents the transition probability for the period $t \geq 0$. This process obeys both the Markov property that given the present state, the future distribution is independent of the past and stationarity property for time-homogenous models such that the transition probabilities are independent of s in the above equation.

This process is completely defined by its transition rates. The transition rates can be arranged into a $k \times k$ transition matrix $Q(t)$ known as the infinitesimal matrix or transition rate matrix where

$$q_{ij}(t) = \lim_{\Delta_t \rightarrow 0} \frac{p_{ij}(t, t + \Delta_t) - p_{ij}(t, t)}{\Delta_t} \quad i \neq j$$

$$q_{ii} = - \sum_{j \neq i} q_{ij}(t) \quad i = 1, \dots, k$$

Therefore a two-state transition matrix, when $k=2$, can be represented as

$$Q = \begin{bmatrix} q_{11} & q_{12} \\ q_{21} & q_{22} \end{bmatrix}$$

Researchers are often interested in the evaluation and interpretation of hazard rates and any covariate effects on these rates. Treating q_{ij} as a hazard rate, we can incorporate covariates in the model and evaluate their effect through the use of a regression-type modelling by taking the log transformation of the hazard rates as defined by the following equation (Ma et. al)

$$\log(q_{ij}) = \alpha_{ij} + \mathbf{z} \boldsymbol{\beta}_{ij}^T$$

where $\mathbf{z} = (z_1, z_2, \dots, z_h)$ represents the covariate vector for h number of covariates, $\boldsymbol{\beta}_{ij}^T = (\beta_{ij}^1, \beta_{ij}^2, \dots, \beta_{ij}^h)^T$ represents the regression coefficients associated with the corresponding covariate vector \mathbf{z} , and α_{ij} represents the intercept of the log-transformed hazard rate for the transition from state i to state j .

Mediation Model with two-state CTMC outcome

The mediating effect was determined through a series of equations. Let $Y(t)$ be a two-state CTMC outcome variable with the following transition matrix

$$Q = \begin{bmatrix} -\lambda & \lambda \\ \mu & -\mu \end{bmatrix}$$

where λ represents the transition rate from state 1 to state 2, and μ represents the transition rate from state 2 to state 1.

Let X and M denote the independent and mediating variable respectfully for the i^{th} subject, $i = 1, 2, \dots, n$, where both X and M are time-invariant random variables.

The total effect of X on the hazard rates of $Y(t)$ can be modelled as follows:

$$\log(\lambda) = \gamma_{\lambda} + cX$$

$$\log(\mu) = \gamma_{\mu} + cX$$

where γ_{λ} and γ_{μ} are the random intercepts that differ between the hazard rate, and c is the fixed effect of X on the hazard rates of Y . We treated the direct effect as the same between the two transitions, but differed the intercepts.

We can model the effect of X on M as follows:

$$M = \beta_0 + aX + \varepsilon$$

where β_0 represents the overall average intercept term for all subjects and a represents the effect of the independent variable, X , on the mediator, M .

We can model the effect of X and M on the hazard rates by the following equations

$$\log(\lambda) = \gamma_{\lambda}^* + c'_{\lambda} X + bM$$

$$\log(\mu) = \gamma_{\mu}^* + c'_{\mu} X + bM$$

where c' is the direct effect of the predictor, X , on the hazard rates and b is the effect of the mediator, M , on the hazard rates when controlling for X .

Combining the above equations we get:

$$\log(\lambda) = \gamma_{\lambda}^* + c'X + b(\beta_{0i} + aX + \varepsilon) = \alpha_{\lambda}^* + c'X + abX$$

$$\log(\mu) = \gamma_{\mu}^* + c'X + b(\beta_{0i} + aX + \varepsilon) = \alpha_{\mu}^* + c'X + abX$$

where α_{λ}^* and α_{μ}^* are the combined intercept term and c' is the direct effect of X on the log of the transition rates of $Y(t)$. Using the product of coefficients method, the mediating effect on the log of the transition rates, or the hazard rate can be expressed as ab .

The equations of the variable effect on the transition rate can be expressed as

$$\lambda = e^{\alpha_{\lambda}^* + c'X + abX}$$

$$\mu = e^{\alpha_{\mu}^* + c'X + abX}$$

Therefore, the mediating effect on the transition rates of $Y(t)$ can be expressed as e^{ab} .

Simulation

For this study, we simulated data sets with varying number of subjects in order to determine ideal sample size required to detect a mediating effect with 80% power. In this case, we treated X as a binary variable, representing treatment or control, and M as a non-time varying continuous variable.

We first simulated the independent variable, X , as a Bernoulli variable with a probability of 0.5. We simulated the mediating variable for each observation using the mediation equation, $M = \beta_0 + aX + \varepsilon$. The error terms, $e_i, i = 1, 2, \dots$, were normally distributed with mean of 0 variance of 1 and the intercept term was set to 1.

Using the simulated X and M, we defined the transition rates for a Q matrix. We set the direct effect of X on the hazard rates (c') to be 0.001. We simulated a Markov chain with 10 transition time points and defined initial states with half starting in state 1 and half in state 2. We varied mediation effect sizes for each parameter estimate (\hat{a} and \hat{b}) these estimates ranged from medium (0.26) and halfway medium-large (0.39).

Once all variables were simulated, for a specified sample size we created 500 data sets of the same sample size. The sample sizes varied from 50, 100, 150, 200, 250, 300, 350, 400, 500 and 600. We calculated the power by determining the percentage of data sets out of the 500 that detect a significant mediating effect. To determine significance, we used 300 bootstrapped samples to calculate a 95% confidence interval around the mediating effect. If the confidence interval excluded zero, then the mediating effect was significant.

Results

Table 1.1 reports the power calculations from 500 simulations of samples with size 50, 100, 150, 200, 250, 300, 350, 400, 500 and 600. The effect a, represents the effect of the predictor variable, X, on the mediator, M. The effect b, represents the effect of the mediator, M, on the outcome, Y, when controlling for the predictor, X. The total mediating effect (TME) on the hazard rates is the product of the two effects (i.e. $TME=a*b$). When a is set to 0.26, the minimum required sample size to achieve 80% power is 500, for both effect sizes of b. When a is set to 0.39, the minimum required sample size to achieve 80% power is 250 for both effect sizes of b.

Sample size (n)	<u>Mediation Effect Size</u>			
	$a=0.26, b=0.26$	$a=0.26, b=0.39$	$a=0.39, b=0.26$	$a=0.39, b=0.39$
50	0.174	0.188	0.294	0.318
100	0.276	0.274	0.518	0.550
150	0.364	0.394	0.670	0.696
200	0.470	0.472	0.786	0.798
250	0.552	0.578	0.876	0.838
300	0.620	0.598	0.906	0.910
350	0.680	0.664	0.968	0.964
400	0.756	0.742	0.964	0.976
500	0.842	0.826	0.994	0.986
600	0.900	0.896	0.999	0.992

Table 1.1. Power calculations from mediation analysis with CTMC outcome. The effect sizes of 0.26 and 0.39 were varied for effect a and b from mediation regression models.

Discussion

Although this study has its limitations, it can serve as a preliminary step into the designation of mediation analysis studies involving CTMC outcomes. In this study, we examined a few varying effect sizes. We did notice that the size of the effect of the predictor, X , on the mediator, M , or the effect a , seemed to have more of an influence on the sample size required to reach 80% power. When effect a was equal to 0.26, the required sample size was around 500 regardless of the effect size b , or the effect of M on the outcome hazard rate when controlling for X . When we increased the effect size of a to 0.39, the required sample size reduced by half to around 250 in order to achieve 80% power regardless of the effect

size of b . Future studies could examine this effect and run simulations with varying effect sizes and differences as well as direction.

Compared to other longitudinal studies with similar effect sizes (Pan et al., 2018), our calculated sample size was larger. However, in those studies, the number of observations varied from one to five and they only included continuous variables. Due to the complexity of the model, the larger sample size makes sense. In this study, when simulating the Markov Chain, we set the number of transition points to ten. Due to convergence issues with Markov Chain regression, we required a larger number of time points than other longitudinal studies (Pan et al., 2018). Future studies could vary the transition number and length and examine its effect on power and desired sample size.

In many longitudinal studies, there are only a few follow-ups due to various reasons, such as cost or patient retention. In a study with only a few transition points, it would be difficult to draw conclusions about the transition rates. However, with technological advances, there are now studies that allow for the acquisition of more data. For example, in some smoking cessation research, a wide range of data is collected through the use of Ecological Momentary Assessment (EMA) (Bandiera, Atem, Ma, Businelle, & Kendzor, 2016b; Jahnel et al., 2018). Therefore, we could design an experiment to examine the mediating effect on transition rates between smoking and abstinence, daily or weekly, by using EMA data. In this study, we simulated the mediator variable as a non-time varying continuous random variable. Suppose we are interested in the mediating effect of post-quit stress between some category or intervention and smoking cessation transitions. With the results of this study, we could design an experiment that examines the mediating effect of

average stress at a specified time points. In future research, we could examine the required sample size needed to examine a time-varying variable, such as daily stress over a certain time-period. We also hope to examine the sample size and power with the inclusion of different covariates, as well as larger state CTMC models.

References

- Baker, C. L., Ding, Y., Ferrufino, C. P., Kowal, S., Tan, J., & Subedi, P. (2018). A cost-benefit analysis of smoking cessation prescription coverage from a US payer perspective.(Original Research). *ClinicoEconomics and Outcomes Research*, 10, 359. doi:10.2147/CEOR.S165576
- Bandiera, F. C., Atem, F., Ma, P., Businelle, M. S., & Kendzor, D. E. (2016). Post-quit stress mediates the relation between social support and smoking cessation among socioeconomically disadvantaged adults. *Drug and Alcohol Dependence*, 163, 71-76. doi:10.1016/j.drugalcdep.2016.03.023
- Gonzalvez, M. T., Morales, A., Orgiles, M., Sussman, S., & Espada, J. P. (2018). Role of smoking intention in tobacco use reduction: A mediation analysis of an effective classroom-based prevention/cessation intervention for adolescents. *Addictive Behaviors*, 84, 186-192. doi:S0306-4603(18)30317-4 [pii]

- Hajek, P., Lewis, S., Munafo, M., Lindson, N., Coleman, T., & Aveyard, P. (2018). Mediators of the effect of nicotine pre-treatment on quitting smoking. *Addiction (Abingdon, England)*, 113(12), 2280. doi:10.1111/add.14401
- Hoepfner, B. B., Hoepfner, S. S., & Abrams, L. C. (2017). How do text-messaging smoking cessation interventions confer benefit? A multiple mediation analysis of Text2Quit. *Addiction*, 112(4), 673-682. doi:10.1111/add.13685
- Imai, K., Keele, L., & Tingley, D. (2010). A general approach to causal mediation analysis. *Psychological Methods*, 15(4), 309-334. doi:10.1037/a0020761
- Jahnel, T., Ferguson, S. G., Shiffman, S., & Schuz, B. (2019). Daily stress as link between disadvantage and smoking: An ecological momentary assessment study.(report). *BMC Public Health*, 19(1) doi:10.1186/s12889-019-7631-2
- Jahnel, T., Ferguson, S. G., Shiffman, S., Thrul, J., & Schüz, B. (2018). Momentary smoking context as a mediator of the relationship between SES and smoking. *Addictive Behaviors*, 83, 136-141. doi:10.1016/j.addbeh.2017.12.014
- Killeen, P. R. (2011). Markov model of smoking cessation. *Proceedings of the National Academy of Sciences of the United States of America*, 108, 15549. doi:10.1073/pnas.1011277108

Li, S., Fang, L., Zhou, Y., Pan, L., Yang, X., Li, H., Jia, C. (2015). Mediation of smoking abstinence self-efficacy on the association of nicotine dependence with smoking cessation. *European Journal of Public Health*, 25(2), 200.
doi:10.1093/eurpub/cku183

Matthew S. Fritz, & David P. MacKinnon. (2007). Required sample size to detect the mediated effect. *Psychological Science*, 18(3), 233-239. doi:10.1111/j.1467-9280.2007.01882.x

Minard, C. G., Chan, W., Wetter, D. W., & Etzel, C. J. (2012). Trends in smoking cessation: A markov model approach. *Journal of Applied Statistics*, 39(1), 113-127.
doi:10.1080/02664763.2011.578619

Pan, H., Liu, S., Miao, D., & Yuan, Y. (2018). Sample size determination for mediation analysis of longitudinal data. *BMC Medical Research Methodology*, 18(1), 32-11.
doi:10.1186/s12874-018-0473-2

Preacher, K. J. (2015). Advances in mediation analysis: A survey and synthesis of new developments. *Annual Review of Psychology*, 66(1), 825-852. doi:10.1146/annurev-psych-010814-015258

Schoemann, A. M., Boulton, A. J., & Short, S. D. (2017). Determining power and sample size for simple and complex mediation models. *Social Psychological and Personality Science*, 8(4), 379-386. doi:10.1177/1948550617715068

- VanderWeele, T. J. (2016). Mediation analysis: A practitioner's guide. *Annual Review of Public Health*, 37(1), 17-32. doi:10.1146/annurev-publhealth-032315-021402
- Vittinghoff, E., Sen, S., & McCulloch, C. E. (2009). Sample size calculations for evaluating mediation. *Statistics in Medicine*, 28(4), 541-557. doi:10.1002/sim.3491
- Vittinghoff, E., & Neilands, T. (2015). Sample size for joint testing of indirect effects. *Prevention Science*, 16(8), 1128-1135. doi:10.1007/s11121-014-0528-5
- Wang, C., & Xue, X. (2016). Power and sample size calculations for evaluating mediation effects in longitudinal studies. *Statistical Methods in Medical Research*, 25(2), 686-705. doi:10.1177/0962280212465163

JOURNAL ARTICLE 2

*Title: Sample Size Analysis For Longitudinal Mediation With A Two-State Continuous
Time Markov Chain Mediator*

To be submitted to Journal of Simulation

Introduction

Mediation analysis is a method that researchers can use to determine the causal pathway between a predictor, X, and outcome, Y (Imai, Keele, & Tingley, 2010; Judd & Kenny, 1981; MacKinnon, David, 2012; Preacher, 2015; VanderWeele, Tyler J., 2016). There may exist an indirect effect of a predictor on an outcome through a mediator, M, that explains the overall predictor-outcome relationship either fully or partially ($X \rightarrow M \rightarrow Y$); and if we control for that mediating effect, the direct effect between a predictor and outcome either ceases to exist or reduces substantially (Baron & Kenny, 1986). In this type of analysis, we are interested in determining or understanding the causal mechanism that links a predictor and an outcome. Researchers use regression techniques to determine this overall mediating effect or indirect effect. Current research has examined mediation with different types of variables and regression, such as linear, logistic, or Poisson regression for both cross-sectional and longitudinal data (MacKinnon, David P. & Fairchild, 2009; Preacher, 2015; Vanderweele, T. J. & Vansteelandt, 2010; VanderWeele, Tyler J. & Tchetgen Tchetgen, 2017). However, no research has examined mediation with a continuous-time Markov chain variable. With the rise of longitudinal data, some researchers use Markov-chain variables to examine the transition rates between stages in disease status, health

behaviors, or in smoking cessation studies (Baker et al., 2018; Jahnel, Ferguson, Shiffman, Thrul, & Schüz, 2018; Killeen, 2011; Ma, Chan, Tsai, Xiong, & Tilley, 2015; Ma, Chan, & Tilley, 2018; Mhoon, Chan, Del Junco, & Vernon, 2010; Tada et al., 2018). Suppose we are interested in developing and studying a weight-maintenance intervention program among smokers. Studies have shown that there is a link between smoking and weight loss (Murphy, Rohsenow, Johnson, & Wing, 2018; Seeley & Sandoval, 2011) . We may be interested in designing a study to examine the causal mechanism between our intervention, smoking cessation, and weight. Treating smoking cessation as a two-state continuous time Markov chain, we can examine the mediating effect of the transition rates between smoking and abstinence on the pathway between the intervention and weight. One of the first steps in study design is determining the minimum sample size required to achieve a desired power to detect a statistically significant effect. Researchers have examined sample size requirements for mediation analysis for different types of studies (Matthew S. Fritz & David P. MacKinnon, 2007; Pan, Liu, Miao, & Yuan, 2018; Schoemann, Boulton, & Short, 2017; Vittinghoff, E., Sen, & McCulloch, 2009; Vittinghoff, Eric & Neilands, 2015; Wang & Xue, 2016). In some of these studies, researchers developed a closed form formula for the calculation of sample size and power; however, in most studies, they use simulations to determine sample size requirements. In addition to power calculations with simulations, researchers use bootstrapping techniques in order to determine the distribution and significance of the mediating effect (MacKinnon, David P., Fairchild, & Fritz, 2007; Matthew S. Fritz & David P. MacKinnon, 2007). Therefore, in this study, we used simulations and bootstrapping to examine the sample size requirements to achieve 80%

power for a longitudinal mediation analysis with a two-state continuous-time Markov chain mediator variable. We treated the predictor as a binary random variable, mimicking intervention groups, and treated the outcome as a non-time varying continuous random variable, mimicking a measurement, such as weight, that would be taken at the end of an intervention time-period.

Methods

In a longitudinal mediation analysis we treated the mediator variable, M , as a two-state continuous time Markov Chain. The explanatory variable, X , was a binary random variable and the outcome, Y , was a non-time-varying continuous random variable.

The total effect of X on Y can be modelled as

$$Y = \alpha + cX$$

where α represents the intercept term, and c represents the total effect of X on the outcome Y .

The random variable $M(t)$ is a two-state continuous time Markov chain with the following transition rate matrix defined as,

$$Q = \begin{bmatrix} -\lambda & \lambda \\ \mu & -\mu \end{bmatrix}$$

where λ is the transition rate from state 1 to state 2, and μ is the transition rate from state 2 to state 1. Using regression-type modelling by taking the log transformation of the transition rate, the effect of X on the hazard rates of $M(t)$ can be modelled as:

$$\log(\lambda) = \gamma_{\lambda}^* + aX$$

$$\log(\mu) = \gamma_{\mu}^* + aX$$

where a is the fixed effect of X on both the hazard rates of $M(t)$ and γ_{λ}^* and γ_{μ}^* are the intercept terms (Ma et al., 2018) .

Therefore, the transition rates of $M(t)$ can be modelled as

$$\lambda = e^{\gamma_{\lambda}^* + aX}$$

$$\mu = e^{\gamma_{\mu}^* + aX}$$

The total effect of X and the transition rates of $M(t)$ on the continuous outcome variable, Y , is modelled by the following linear equation:

$$\begin{aligned} Y &= \alpha^* + c'X + b_1\lambda + b_2\mu + \varepsilon \\ &= \alpha^* + c'X + b_1e^{\gamma_{\lambda}^* + aX} + b_2e^{\gamma_{\mu}^* + aX} + \varepsilon \\ &= \alpha^* + c'X + b_1e^{\gamma_{\lambda}^*}e^{aX} + b_2e^{\gamma_{\mu}^*}e^{aX} + \varepsilon \end{aligned}$$

where c' is the direct effect of X on the outcome, Y , when controlling for $M(t)$; b_1, b_2 are the overall effects of the transition rates from $M(t)$ on the outcome, Y ; ε is the error term for each individual.

To interpret and determine the mediating effect, we used a first order Taylor polynomial as a linear approximation for e^{aX} , thus resulting in the following equation

$$\begin{aligned} Y &= \alpha^* + c'X + b_1e^{\gamma_{\lambda}^*}(1 + aX) + b_2e^{\gamma_{\mu}^*}(1 + aX) \\ &= \beta_0^* + c'X + a(b_1e^{\gamma_{\lambda}^*} + b_2e^{\gamma_{\mu}^*})X \end{aligned}$$

where β^* represents the combined intercept term of $\alpha^* + b_1 e^{\gamma_\lambda^*} + b_2 e^{\gamma_\mu^*}$, c' is the direct effect of X on Y, and $a(b_1 e^{\gamma_\lambda^*} + b_2 e^{\gamma_\mu^*})$ is the total mediating effect Y.

Simulation

We first simulated X as a binary variable with a success probability of 0.5. We then simulated the Markov chain variable, M(t) by defining the transition matrix using the transition rate equations above by setting different intercept terms between the two transition rates ($\gamma_\lambda^* = -0.2$ and $\gamma_\mu^* = -0.8$) and differing the effect size, a , for each simulation. We then simulated a two-state Markov chain with 20 transition time points. Using continuous-time Markov chain regression, we calculated individual transition rates from the simulated Markov chain for both λ and μ . We simulated the outcome variable, Y, using the linear regression equation including X, $\vec{\lambda}$ and $\vec{\mu}$. The direct effect of X controlling for M(t) was set to 0.0001 and the error terms, $\varepsilon_i, i = 1, 2, \dots$, were normally distributed with mean of 0 and variance of 1. We varied the effect size of the mediating effect by changing the parameter value for a, b_1, b_2 . The effect size for a (the effect of X on M(t)) was varied between 0.26 and 0.39. The effect sizes for b_1 (the effect of the transition rate λ on the outcome Y when controlling for X) and b_2 (the effect of the transition rate μ on the outcome Y when controlling for X) were varied by 0.15 and 0.26.

We simulated 500 data sets with assigned parameters, sample size, and calculated the power. To determine the significance of the mediating effect, we constructed a confidence interval based on a bootstrapping of 300 resamples. If the confidence interval contained

zero, then we concluded a non-significant mediating effect. We calculated the power after running the mediation model by counting the number of data sets that resulted in a significant mediation effect. We started the sample sizes at 50 and increased by 50 or 25 until we reached 80% power.

Results

Table 2.1 records the power calculations when the effect of X on M(t) was set to 0.26 with varying effect sizes for the effect of the transition rates from M(t) on the continuous outcome, Y, when controlling for binary X. For each of these simulations, the sample size that achieved at least 80% power was 250.

Table 2.2 reports the power calculations when the effect of X on M(t) was set to 0.39. For these simulations, a sample size of 125 achieved 80% power. We found that when increasing by 50 like in Table 1, that a sample size of 150 achieved beyond the desired 80% and reached 90%. Therefore, we increased by 25 to determine a smaller sample size that still achieved our desired power.

Sample size (n)	<u>Mediation Effect Size Parameters ($a=0.26$)</u>			
	$b1=0.26$ $b2=0.26$	$b1=0.26$ $b2=0.15$	$b1=0.15$ $b2=0.26$	$b1=0.15$ $b2=0.15$
50	0.278	0.276	0.268	0.236
100	0.476	0.476	0.444	0.478
150	0.628	0.628	0.634	0.590
200	0.766	0.766	0.758	0.778
250	0.824	0.814	0.818	0.818
300	0.904	0.884	0.890	0.904

Table 2.1. Power calculations for the mediating effect when the effect size of X on M(t) was set to 0.26 for various sample sizes and transition effect size

Sample size (n)	<u>Mediation Effect Size Parameters ($a=0.39$)</u>			
	$b1=0.26$ $b2=0.26$	$b1=0.26$ $b2=0.15$	$b1=0.15$ $b2=0.26$	$b1=0.15$ $b2=0.15$
50	0.520	0.456	0.446	0.420
100	0.780	0.738	0.740	0.780
125	0.842	0.842	0.830	0.836
150	0.924	0.914	0.904	0.890

Table 2.2. Power calculations for the mediating effect when the effect size of X on M(t) was set to 0.39 for various sample sizes and transition effect size

Discussion

This study can serve as a preliminary resource in designing a study to investigate the mediating effect of a two-state continuous-time Markov chain mediator variable, M(t), between a binary predictor, X, and continuous outcome variable, Y. When the effect of the predictor on the transition rates of the mediator (a) was set to 0.26, the sample size required

to achieve 80% power was 250 regardless of the effect sizes of the transition rates on the outcome adjusted for the predictor (b_1, b_2). In the same way, when the predictor effect on the mediator transition rates (a) was set to 0.39, the sample size required to achieve 80% power was 125 for all other varying effect sizes, b_1, b_2 . This result could be because of the additional multiplicative factors on the effects of the transition rates on the outcome controlled for the predictor. The formula used to calculate the over mediating effect was $a(b_1 e^{\gamma_\lambda^*} + b_2 e^{\gamma_\mu^*})$. Since γ_λ^* and γ_μ^* were set to negative values, this would decrease the overall effect size of the combined effect of both transition rates. Therefore, such a small change in those effect sizes would not have as great of an impact in the overall mediating effect product. Future studies could examine more variations in effect sizes, not just for effects a, b_1, b_2 , but also vary the intercept terms γ_λ^* and γ_μ^* .

One issue we encountered during simulation was the non-convergence of some continuous-time Markov-chain regressions. We based the intercept terms on values that reduced the percentage of convergence issues. Another element of the simulations we controlled for was the number of transition time points. We set the number of time points to 20 because that reduced the percentage of convergence issues. Future studies could examine the convergence issues in further detail and vary the number of time point transitions. Another limitation is the exclusion of additional covariates. In other longitudinal mediation analyses and Markov chain analyses, researchers include covariates that are pertinent to the study at hand (Abner, Charnigo, & Kryscio, 2013; Cole & Maxwell, 2003; Selig, Preacher, & Little, 2012; VanderWeele & Tchetgen Tchetgen, 2017; Wang & Xue, 2016). We could

also examine the sample size requirement for different variable types of predictors and outcomes and possibly include a time-varying outcome instead of only a non-time varying continuous outcome.

Although this study has its limitations, it can serve as an initial step in designing an experiment that expands on current mediation analysis research. We can also use these results to determine retrospectively if a completed study has enough observations to detect a potential mediating effect. We can use statistics to support or reject a conceptually potential mediation model that involves the inclusion of a Markov-chain mediator model. For example, we may have an intervention that targets weight loss or gain in smokers. We know that smoking affects weight change (Kim et al., 2012; Murphy et al., 2018; Seeley & Sandoval, 2011). If we also know that this intervention has an effect on smoking cessation, we can use this type of model to determine the presence or absence of a mediating effect. Researchers could collect data on whether a person had smoked or abstained at a specific time points using Ecological Momentary Assessment or other methods and treat these results as a two-state Markov chain (Bandiera, Atem, Ma, Businelle, & Kendzor, 2016; Jahnel et al., 2018).

References

Abner, E. L., Charnigo, R. J., & Kryscio, R. J. (2013). Markov chains and semi-markov models in time-to-event analysis. *Journal of Biometrics & Biostatistics, Suppl 1*, 19522. doi:10.4172/2155-6180.S1-e001

- Baker, C. L., Ding, Y., Ferrufino, C. P., Kowal, S., Tan, J., & Subedi, P. (2018). A cost-benefit analysis of smoking cessation prescription coverage from a US payer perspective.(ORIGINAL RESEARCH). *ClinicoEconomics and Outcomes Research*, 10, 359. doi:10.2147/CEOR.S165576
- Bandiera, F. C., Atem, F., Ma, P., Businelle, M. S., & Kendzor, D. E. (2016). Post-quit stress mediates the relation between social support and smoking cessation among socioeconomically disadvantaged adults. *Drug and Alcohol Dependence*, 163, 71-76. doi:10.1016/j.drugalcdep.2016.03.023
- Baron, R. M., & Kenny, D. A. (1986). The moderator–mediator variable distinction in social psychological research: Conceptual, strategic, and statistical considerations. *Journal of Personality and Social Psychology*, 51(6), 1173.
- Cole, D. A., & Maxwell, S. E. (2003). Testing mediational models with longitudinal data: Questions and tips in the use of structural equation modeling. *Journal of Abnormal Psychology*, 112(4), 558.
- Imai, K., Keele, L., & Tingley, D. (2010). A general approach to causal mediation analysis. *Psychological Methods*, 15(4), 309-334. doi:10.1037/a0020761
- Jahnel, T., Ferguson, S. G., Shiffman, S., Thrul, J., & Schüz, B. (2018). Momentary smoking context as a mediator of the relationship between SES and smoking. *Addictive Behaviors*, 83, 136-141. doi:10.1016/j.addbeh.2017.12.014

- Judd, C. M., & Kenny, D. A. (1981). Process analysis: Estimating mediation in treatment evaluations. *Evaluation Review*, 5(5), 602-619. doi:10.1177/0193841X8100500502
- Killeen, P. R. (2011). Markov model of smoking cessation. *Proceedings of the National Academy of Sciences of the United States of America*, 108, 15549.
doi:10.1073/pnas.1011277108
- Kim, J. H., Shim, K. W., Yoon, Y. S., Lee, S. Y., Kim, S. S., Oh, S. W., & Sun, Q. (2012). Cigarette smoking increases abdominal and visceral obesity but not overall fatness: An observational study (cigarette smoking and visceral obesity).7(9), e45815.
doi:10.1371/journal.pone.0045815
- Ma, J., Chan, W., & Tilley, B. C. (2018). Continuous time markov chain approaches for analyzing transtheoretical models of health behavioral change: A case study and comparison of model estimations. *Statistical Methods in Medical Research*, 27(2), 593-607. doi:10.1177/0962280216639859
- Ma, J., Chan, W., Tsai, C., Xiong, M., & Tilley, B. C. (2015). Analysis of transtheoretical model of health behavioral changes in a nutrition intervention study--a continuous time markov chain model with bayesian approach. *Statistics in Medicine*, 34(27), 3577-3589. doi:10.1002/sim.6571
- MacKinnon, D. (2012). *Introduction to statistical mediation analysis* Routledge.

MacKinnon, D. P., & Fairchild, A. J. (2009). Current directions in mediation analysis.

Current Directions in Psychological Science, 18(1), 16-20. Retrieved from

<http://www.jstor.org.utsph.idm.oclc.org/stable/20695987>

MacKinnon, D. P., Fairchild, A. J., & Fritz, M. S. (2007). Mediation analysis. *Annual Review of Psychology*, 58(1), 593-614. doi:10.1146/annurev.psych.58.110405.085542

Matthew S. Fritz, & David P. MacKinnon. (2007). Required sample size to detect the mediated effect. *Psychological Science*, 18(3), 233-239. doi:10.1111/j.1467-9280.2007.01882.x

Mhoon, K. B., Chan, W., Del Junco, D.,J., & Vernon, S. W. (2010). A continuous-time markov chain approach analyzing the stages of change construct from a health promotion intervention. *JP Journal of Biostatistics*, 4(3), 213-226. Retrieved from <https://www.ncbi.nlm.nih.gov/pubmed/23504410>

Murphy, C. M., Rohsenow, D. J., Johnson, K. C., & Wing, R. R. (2018). Smoking and weight loss among smokers with overweight and obesity in look AHEAD. *Health Psychology : Official Journal of the Division of Health Psychology, American Psychological Association*, 37(5), 399. doi:10.1037/hea0000607

Pan, H., Liu, S., Miao, D., & Yuan, Y. (2018). Sample size determination for mediation analysis of longitudinal data. *BMC Medical Research Methodology*, 18(1), 32-11. doi:10.1186/s12874-018-0473-2

- Preacher, K. J. (2015). Advances in mediation analysis: A survey and synthesis of new developments. *Annual Review of Psychology*, 66(1), 825-852. doi:10.1146/annurev-psych-010814-015258
- Schoemann, A. M., Boulton, A. J., & Short, S. D. (2017). Determining power and sample size for simple and complex mediation models. *Social Psychological and Personality Science*, 8(4), 379-386. doi:10.1177/1948550617715068
- Seeley, R., & Sandoval, D. (2011). Weight loss through smoking. *Nature*, 475(7355), 176-7. doi:10.1038/475176a
- Selig, J. P., Preacher, K. J., & Little, T. D. (2012). Modeling time-dependent association in longitudinal data: A lag as moderator approach. *Multivariate Behavioral Research*, 47(5), 697-716.
- Tada, T., Kumada, T., Toyoda, H., Ohisa, M., Akita, T., & Tanaka, J. (2018). Long-term natural history of liver disease in patients with chronic hepatitis B virus infection: An analysis using the markov chain model. *Journal of Gastroenterology*, 53(11), 1196-1205. doi:10.1007/s00535-018-1467-x
- Vanderweele, T. J., & Vansteelandt, S. (2010). Odds ratios for mediation analysis for a dichotomous outcome. *American Journal of Epidemiology*, 172(12), 1339-1348. doi:10.1093/aje/kwq332 [doi]

- VanderWeele, T. J. (2016). Mediation analysis: A practitioner's guide. *Annual Review of Public Health*, 37(1), 17-32. doi:10.1146/annurev-publhealth-032315-021402
- VanderWeele, T. J., & Tchetgen Tchetgen, E. J. (2017). Mediation analysis with time varying exposures and mediators. *Journal of the Royal Statistical Society. Series B, Statistical Methodology*, 79(3), 917-938. doi:10.1111/rssb.12194
- Vittinghoff, E., Sen, S., & McCulloch, C. E. (2009). Sample size calculations for evaluating mediation. *Statistics in Medicine*, 28(4), 541-557. doi:10.1002/sim.3491
- Vittinghoff, E., & Neilands, T. (2015). Sample size for joint testing of indirect effects. *Prevention Science*, 16(8), 1128-1135. doi:10.1007/s11121-014-0528-5
- Wang, C., & Xue, X. (2016). Power and sample size calculations for evaluating mediation effects in longitudinal studies. *Statistical Methods in Medical Research*, 25(2), 686-705. doi:10.1177/0962280212465163

Journal Article 3

Title: Sample Size Analysis for Longitudinal Mediation with a Two-State Continuous Time

Markov Chain Predictor

To be submitted to Journal of Simulation

Introduction

Mediation Analysis has increased in popularity in the field of causal inference. In this type of analysis, the idea is that there is a known predictor (X) that affects some outcome (Y) but the effect of this predictor is “mediated”, either fully or partially, through the effect of a different variable (M) (Imai, Keele, & Tingley, 2010b; VanderWeele, Tyler J., 2016; Xinshu Zhao, John G. Lynch, & Qimei Chen, 2010). The diverse use of mediation analysis includes studies of alcohol consumption, policy interventions, financial and market performance, and smoking interventions (Gonzalvez et al., 2018; Jones-Webb & Karriker-Jaffe, 2013; Keele et al., 2015; Semrau & Sigmund, 2012; Voola et al., 2012). A mediation model is a causal model in which researches use statistics to assess a mediating effect. In this causal model researchers know that X causes Y, but in order to find the indirect, or mediating effect, they must also determine that X causes M and M causes Y. In a partial mediating model, the direct effect of X on Y when controlling for M is significantly reduced whereas in a fully mediated model, the direct effect is reduced to zero. A series of regression techniques is used to determine the possible existence and significance of the mediating effect (Baron & Kenny, 1986). Current research includes the use of linear, logistic, probit, and Poisson regression (Preacher, 2015). Previous mediation analysis studies include continuous, binary,

categorical, and count type variables for both cross-sectional and longitudinal studies, but none include the presence of a Continuous-time Markov Chain predictor.

Suppose we are interested in designing a study to determine the presence of a mediating effect of caffeine on smoking transition's effect on obesity, since studies have shown connections between smoking, caffeine consumption, and factors relating to obesity (Bjørngaard et al., 2017; Kim et al., 2012; Nurwanti & Bai, 2018). Specifically, we are interested in how caffeine consumption mediates the effect of a subject transitioning between smoking and abstinence on weight. We could collect data over a series of time points and treat the smoking variable as a two-state continuous time Markov chain, in which one state represents smoking and the other state is abstinence. One of the first steps in designing this type of experiment would be to determine the minimum sample size required to achieve our desired power. When it comes to sample size calculations in mediation analysis, there are not many closed form formulas, so the use of bootstrapping is often used (Matthew S. Fritz & David P. MacKinnon, 2007). There are several papers that have done these types of sample size calculations for various types of variables and effect sizes, but as previously mentioned, there are no studies involving Markov chains (Matthew S. Fritz & David P. MacKinnon, 2007; Pan et al., 2018; Schoemann, Boulton, & Short, 2017; Vittinghoff et al., 2009; Vittinghoff & Neilands, 2015; Wang & Xue, 2016). The purpose of this study is to provide sample size information to achieve 80% power that can serve as a starting point for the design of a study that includes a two-state Continuous-time Markov Chain as the predictor variable in a longitudinal mediation analysis.

Methods

In a longitudinal mediation analysis we treated the predictor variable, $X(t)$, as a two-state continuous time Markov Chain and both the outcome variable, Y , and the mediator variable, M , as continuous random variables. We considered both M and Y as non-time varying variables representing measurements taken once at a follow-up time. In this case, the idea is that we have already calculated the transition rates for the predictor variable from previous data collected over a specific number of time points.

We classified $X(t)$ as a two state CTMC with the following transition rate matrix:

$$Q = \begin{bmatrix} -\lambda & \lambda \\ \mu & -\mu \end{bmatrix}$$

where λ is the transition rate from state 1 to state 2 and μ is the transition rate from state 2 to state 1.

The overall total effect of the transition rates on the outcome Y is

$$Y = \beta_{01} + c_1\lambda + c_2\mu$$

Treating M as a continuous variable, the effect of the transition rates on the mediator is modelled as follows:

$$M = \beta_{02} + a_1\lambda + a_2\mu$$

The effect of the transition rates controlling for M is:

$$Y = \beta_{03} + c'\lambda + c'\mu + bM$$

Combining the two equations we get

$$\begin{aligned} Y &= \beta_{03} + c'\lambda + c'\mu + b(\beta_{02} + a_1\lambda + a_2\mu) \\ &= \beta_{03}^* + c'\lambda + c'\mu + a_1b\lambda + a_2b\mu \end{aligned}$$

where β_{03}^* represents the combined intercept value, c' represents the direct effect of the transition rates of $X(t)$ on Y and a_1b and a_2b are the mediating effects of the two transition rates on Y . The total mediating effect on Y can be expressed as: $b(a_1 + a_2)$. To determine the significance of the mediating effect, we used bootstrapping of sample size 300 to construct a 95% confidence interval around the mediating effect estimate. If the confidence interval contained the value of zero, then the mediating effect was non-significant. We simulated 500 data sets with a specified sample size and determined the power by calculating the percentage of significant tests out of the 500 simulated data sets.

Simulation

To simulate the Markov chain variable $X(t)$, we set the transition rates to $\lambda = 0.8$ and $\mu = 1.2$ initially. We assigned a 0.5 probability for the initial state, mimicking the idea that half the subjects start in state 1 and half in state 2. We simulated a Markov chain over ten time points. After simulating individual transition states over ten time points, individual simulated transition rates were determined through a continuous-time Markov chain regression with the “msm” package in R. We used the new simulated transition rates as the predictors in the regression equations. We then simulated the mediator variable and outcome variable using the regression equations above. We set the direct effect of the transition rates on the outcome when controlling for the mediator (c') as 0.001. We varied the mediating effect sizes using values 0.07, 0.15 and 0.26 for a_1 and a_2 and 0.15 and 0.26 for b , representing small and medium effect sizes based on previous studies ((Matthew S. Fritz & David P. MacKinnon,

2007). We calculated the total mediating effect (TME) using the product of coefficients method as defined above and reported it in the results table.

Results

The power calculations from the varying effect and sample sizes are reported in the tables below. In order to achieve at least 80% power a sample size of at least 350 is required if we are interested in detecting a mediating effect of 0.021 or 0.045 (Table 3.1). In both of these simulations, the effect of M on Y adjusted for X(t) is 0.15. When the effect of M on Y adjusted for X(t) was 0.26, the minimum required sample size to achieve 80% power was 150, regardless of the effect size between the transition rates and mediator variable (Table 3.2).

Sample size (n)	Mediation Effect Size when $b=0.15$	
	$TME=0.021$ $a1, a2=0.07$	$TME=0.045$ $a1,a2=0.15$
50	0.100	0.198
100	0.278	0.318
150	0.414	0.466
200	0.582	0.552
250	0.662	0.704
300	0.736	0.718
350	0.816	0.814
400	0.854	0.854
500	0.898	0.926

Table 3.1. Power calculations when effect size $b=0.15$. The total mediating effect (TME) and the specific effect sizes of the components of the mediating effect are recorded.

Sample size (n)	<u>Mediation Effect Size when $b=0.26$</u>			
	$TME= 0.0364$ $a1=0.07$ $a2=0.07$	$TME= 0.078$ $a1=0.15$ $a2=0.15$	$TME= 0.106$ $a1=0.26,$ $a2=0.15$	$TME= 0.106$ $a1=0.15$ $a2=0.26$
50	0.224	0.428	0.434	0.434
100	0.658	0.732	0.734	0.734
150	0.874	0.888	0.888	0.888
200	0.966	0.942	0.942	0.942

Table 3.2. Power Calculations when effect size $b=0.26$. The total mediating effect (TME) and the specific effect sizes of the components of the mediating effect are recorded.

Discussion

The results of this paper can serve as a starting point in designing an experiment to test for the mediating effect of a continuous variable between a two-state continuous time Markov chain and a continuous outcome. However, there are some limitations to these results. For example, we did not vary the direction of the mediating effect. In some studies, such as smoking and obesity, smoking tends to have a negative effect on weight (Kim et al., 2012; Li et al., 2015). Several longitudinal studies of mediation analysis initially evaluate the sample size using positive effect sizes (Pan et al., 2018; Schoemann et al., 2017; Wang & Xue, 2016). In future research, we could examine the sample size requirements for negative mediating effects. We could also examine more combinations of effect sizes and reduced the direct effect more.

It is interesting to note that in one instance the total mediating effect size was 0.0364, but only required a sample size of 150, whereas the data sets simulated from a larger effect size of 0.045 required 350. A smaller effect size should require a larger sample size to

achieve a higher power, but this was not the case. There seems to be a relation between the power and sample size and the effect of M on Y when adjusting for X (effect b). In the instance where the total mediating effect was 0.0364, the effect $b=0.26$ and required a lower sample size, suggesting that this effect may have more of an effect on power.

We also came across issues with convergence when calculating individual transition rates through Continuous time Markov chain regression and the regression tended to overestimate the simulated parameters possibly leading to an inflation of power. In order to aid the convergence issue, we increased the number of time points to 10. In future studies, we could differ the number of transition time points and calculate the required sample size. We could also differ the transition rate size. In this study, we only used a transition rate of 0.8 from state 1 to state 2 and a rate of 1.2 for transition from state 2 to state 1. In future studies, we aim to vary the transition rates.

Although there are limitations to this study, it still can serve as a helpful resource when wanting to find a mediating effect with a two-state continuous-time Markov chain predictor. Studies have examined the sample size requirements for various types of variables (MacKinnon et al., 2007; MacKinnon, David P. & Fairchild, 2009; Matthew S. Fritz & David P. MacKinnon, 2007; Preacher, 2015). However, along with mediation Markov chain usage is also a progressing area in statistics. This novel idea of conducting a mediation analysis using transition rate variables could prove informative and influential, particularly in the area of smoking research. Since researchers have already begun to study the transition rates of smoking to cessation using Markov-chains, the next step would be to examine how these

transition effects could be mediated by another variable(Killeen, 2011; Koslovsky et al., 2018; Ma et al., 2018; Minard, Chan, Wetter, & Etzel, 2012a).

References

- Baron, R. M., & Kenny, D. A. (1986). The moderator–mediator variable distinction in social psychological research: Conceptual, strategic, and statistical considerations. *Journal of Personality and Social Psychology*, 51(6), 1173.
- Bjørngaard, J.,H., Nordestgaard, A. T., Taylor, A. E., Treur, J. L., Gabrielsen, M. E., Munafò, M.,R., . . . Davey Smith, G. (2017). Heavier smoking increases coffee consumption: Findings from a mendelian randomization analysis. *International Journal of Epidemiology*, 46(6), 1958. doi:10.1093/ije/dyx147
- Gonzalvez, M. T., Morales, A., Orgiles, M., Sussman, S., & Espada, J. P. (2018). Role of smoking intention in tobacco use reduction: A mediation analysis of an effective classroom-based prevention/cessation intervention for adolescents. *Addictive Behaviors*, 84, 186-192. doi:S0306-4603(18)30317-4 [pii]
- Imai, K., Keele, L., & Tingley, D. (2010). A general approach to causal mediation analysis. *Psychological Methods*, 15(4), 309-334. doi:10.1037/a0020761

Jones-Webb, R., & Karriker-Jaffe, K. (2013). Neighborhood disadvantage, high alcohol content beverage consumption, drinking norms, and drinking consequences: A mediation analysis. *Journal of Urban Health*, 90(4), 667-684. doi:10.1007/s11524-013-9786-y

Keele, L., Tingley, D., & Yamamoto, T. (2015). Identifying mechanisms behind policy interventions via causal mediation analysis. *Journal of Policy Analysis and Management*, 34(4), 937-963. doi:10.1002/pam.21853

Killeen, P. R. (2011). Markov model of smoking cessation. *Proceedings of the National Academy of Sciences of the United States of America*, 108, 15549. doi:10.1073/pnas.1011277108

Kim, J. H., Shim, K. W., Yoon, Y. S., Lee, S. Y., Kim, S. S., Oh, S. W., & Sun, Q. (2012). Cigarette smoking increases abdominal and visceral obesity but not overall fatness: An observational study (cigarette smoking and visceral obesity).7(9), e45815. doi:10.1371/journal.pone.0045815

Koslovsky, M. D., Swartz, M. D., Chan, W., Leon-Novelo, L., Wilkinson, A. V., Kendzor, D. E., & Businelle, M. S. (2018). Bayesian variable selection for multistate markov models with interval-censored data in an ecological momentary assessment study of smoking cessation. *Biometrics*, 74(2), 636-644. doi:10.1111/biom.12792

- Li, S., Fang, L., Zhou, Y., Pan, L., Yang, X., Li, H., . . . Jia, C. (2015). Mediation of smoking abstinence self-efficacy on the association of nicotine dependence with smoking cessation. *European Journal of Public Health*, 25(2), 200. doi:10.1093/eurpub/cku183
- Ma, J., Chan, W., & Tilley, B. C. (2018). Continuous time markov chain approaches for analyzing transtheoretical models of health behavioral change: A case study and comparison of model estimations. *Statistical Methods in Medical Research*, 27(2), 593-607. doi:10.1177/0962280216639859
- MacKinnon, D. P., & Fairchild, A. J. (2009). Current directions in mediation analysis. *Current Directions in Psychological Science*, 18(1), 16-20. Retrieved from <http://www.jstor.org.utsph.idm.oclc.org/stable/20695987>
- MacKinnon, D. P., Fairchild, A. J., & Fritz, M. S. (2007). Mediation analysis. *Annual Review of Psychology*, 58(1), 593-614. doi:10.1146/annurev.psych.58.110405.085542
- Matthew S. Fritz, & David P. MacKinnon. (2007). Required sample size to detect the mediated effect. *Psychological Science*, 18(3), 233-239. doi:10.1111/j.1467-9280.2007.01882.x
- Minard, C. G., Chan, W., Wetter, D. W., & Etzel, C. J. (2012a). Trends in smoking cessation: A markov model approach. *Journal of Applied Statistics*, 39(1), 113-127. doi:10.1080/02664763.2011.578619

- Nurwanti, E., & Bai, C. (2018). A population-based study of sedentary behavior, coffee and caffeine intake, education level associated with obesity risk among young adult. *Proceedings of the Nutrition Society*, 77(OCE4), E184.
doi:10.1017/S0029665118001908
- Pan, H., Liu, S., Miao, D., & Yuan, Y. (2018). Sample size determination for mediation analysis of longitudinal data. *BMC Medical Research Methodology*, 18(1), 32-11.
doi:10.1186/s12874-018-0473-2
- Preacher, K. J. (2015). Advances in mediation analysis: A survey and synthesis of new developments. *Annual Review of Psychology*, 66(1), 825-852. doi:10.1146/annurev-psych-010814-015258
- Schoemann, A. M., Boulton, A. J., & Short, S. D. (2017). Determining power and sample size for simple and complex mediation models. *Social Psychological and Personality Science*, 8(4), 379-386. doi:10.1177/1948550617715068
- Semrau, T., & Sigmund, S. (2012). Networking ability and the financial performance of new ventures: A mediation analysis among younger and more mature firms. *Strategic Entrepreneurship Journal*, 6(4), 335-354. doi:10.1002/sej.1146
- VanderWeele, T. J. (2016). Mediation analysis: A practitioner's guide. *Annual Review of Public Health*, 37(1), 17-32. doi:10.1146/annurev-publhealth-032315-021402

Vittinghoff, E., Sen, S., & McCulloch, C. E. (2009). Sample size calculations for evaluating mediation. *Statistics in Medicine*, 28(4), 541-557. doi:10.1002/sim.3491

Vittinghoff, E., & Neilands, T. (2015). Sample size for joint testing of indirect effects. *Prevention Science*, 16(8), 1128-1135. doi:10.1007/s11121-014-0528-5

Voola, R., Casimir, G., Carlson, J., & Anushree Agnihotri, M. (2012). The effects of market orientation, technological opportunism, and e-business adoption on performance: A moderated mediation analysis. *Australasian Marketing Journal (AMJ)*, 20(2), 136-146. doi://doi.org/10.1016/j.ausmj.2011.10.001

Wang, C., & Xue, X. (2016). Power and sample size calculations for evaluating mediation effects in longitudinal studies. *Statistical Methods in Medical Research*, 25(2), 686-705. doi:10.1177/0962280212465163

Xinshu Zhao, John G. Lynch, & Qimei Chen. (2010). Reconsidering baron and kenny: Myths and truths about mediation analysis. *Journal of Consumer Research*, 37(2), 197-206. doi:10.1086/651257

CONCLUSION

All three of these studies can serve as a starting point in designing a mediation analysis study that consists of a two-state continuous time Markov chain (CTMC) as one of the elements in the mediation model. We examined a CTMC outcome with binary predictor and continuous mediator, a CTMC mediator with binary predictor and continuous outcome, and a CTMC predictor with continuous predictor and mediator. Not only have we determined minimum sample size to achieve 80% power for certain situations, we also have provided formulas to calculate the total mediating effect in these situations. Because no other mediation studies have included Markov chain regression of transition rates in the set of mediating equations used to calculate the indirect and direct effects, there is ample room for expansion in this field.

In general, these studies have similar limitations. In each study, we set the number of transition time points in the Markov chain and the number of stages to two; we did not examine any other variations due to convergence issue with the Markov chain regression. Further research could examine these issues and try different number of transition points and states to see how the sample size requirements and power calculations change. We could also examine more variations in effect sizes based on a specific research interest and effect sizes typically seen in that research. Finally, in future work we could include additional covariates that may serve as adjusters to aid in truly capturing the mediating effect and causal pathway in a study. Although these studies have their limitations, they serve as a preliminary step in causal research involving mediation with continuous-time Markov chain variables.

REFERENCES

- Aalen, O. O., Gran, J. M., Røysland, K., Stensrud, M. J., & Strohmaier, S. (2018). Feedback and mediation in causal inference illustrated by stochastic process models. *Scandinavian Journal of Statistics*, 45(1), 62-86. doi:10.1111/sjos.12286
- Abner, E. L., Charnigo, R. J., & Kryscio, R. J. (2013). Markov chains and semi-markov models in time-to-event analysis. *Journal of Biometrics & Biostatistics, Suppl 1*, 19522. doi:10.4172/2155-6180.S1-e001
- Baker, C. L., Ding, Y., Ferrufino, C. P., Kowal, S., Tan, J., & Subedi, P. (2018). A cost-benefit analysis of smoking cessation prescription coverage from a US payer perspective.(ORIGINAL RESEARCH). *ClinicoEconomics and Outcomes Research*, 10, 359. doi:10.2147/CEOR.S165576
- Bandiera, F. C., Atem, F., Ma, P., Businelle, M. S., & Kendzor, D. E. (2016). Post-quit stress mediates the relation between social support and smoking cessation among socioeconomically disadvantaged adults. *Drug and Alcohol Dependence*, 163, 71-76. doi:10.1016/j.drugalcdep.2016.03.023
- Baron, R. M., & Kenny, D. A. (1986). The moderator–mediator variable distinction in social psychological research: Conceptual, strategic, and statistical considerations. *Journal of Personality and Social Psychology*, 51(6), 1173.

- Bjørngaard, J.,H., Nordestgaard, A. T., Taylor, A. E., Treur, J. L., Gabrielsen, M. E., Munafò, M.,R., Davey Smith, G. (2017). Heavier smoking increases coffee consumption: Findings from a mendelian randomization analysis. *International Journal of Epidemiology*, 46(6), 1958. doi:10.1093/ije/dyx147
- Buis, M. L. (2010). Direct and indirect effects in a logit model. *The Stata Journal*, 10(1), 11.
- Cole, D. A., & Maxwell, S. E. (2003). Testing mediational models with longitudinal data: Questions and tips in the use of structural equation modeling. *Journal of Abnormal Psychology*, 112(4), 558.
- Coxe, S., & MacKinnon, D. P. (2010). Mediation analysis of poisson distributed count outcomes. *Multivariate Behavioral Research*, 45(6), 1022.
- Deboeck, P. R., & Preacher, K. J. (2016). No need to be discrete: A method for continuous time mediation analysis. *Structural Equation Modeling: A Multidisciplinary Journal*, 23(1), 61-75. doi:10.1080/10705511.2014.973960
- Fishbein, M.,A., & Ajzen, I. (1975). *Belief, attitude, intention and behaviour: An introduction to theory and research*
- Gollob, H. F., & Reichardt, C. S. (1991). Interpreting and estimating indirect effects assuming time lags really matter.

- Gonzalvez, M. T., Morales, A., Orgiles, M., Sussman, S., & Espada, J. P. (2018). Role of smoking intention in tobacco use reduction: A mediation analysis of an effective classroom-based prevention/cessation intervention for adolescents. *Addictive Behaviors, 84*, 186-192. doi:S0306-4603(18)30317-4 [pii]
- Hajek, P., Lewis, S., Munafo, M., Lindson, N., Coleman, T., & Aveyard, P. (2018). Mediators of the effect of nicotine pre-treatment on quitting smoking. *Addiction (Abingdon, England), 113*(12), 2280. doi:10.1111/add.14401
- Hebb, D. O. (1958). *Textbook of psychology* (1st ed.) Philadelphia: Saunders.
- Hoepfner, B. B., Hoepfner, S. S., & Abrams, L. C. (2017). How do text-messaging smoking cessation interventions confer benefit? A multiple mediation analysis of Text2Quit. *Addiction, 112*(4), 673-682. doi:10.1111/add.13685
- Huang, B., Sivaganesan, S., Succop, P., & Goodman, E. (2004). Statistical assessment of mediational effects for logistic mediational models. *Statistics in Medicine, 23*(17), 2713-2728.
- Imai, K., Keele, L., & Tingley, D. (2010b). A general approach to causal mediation analysis. *Psychological Methods, 15*(4), 309-334. doi:10.1037/a0020761
- Jahnel, T., Ferguson, S. G., Shiffman, S., & Schuz, B. (2019). Daily stress as link between disadvantage and smoking: An ecological momentary assessment study.(report). *BMC Public Health, 19*(1) doi:10.1186/s12889-019-7631-2

- Jahnel, T., Ferguson, S. G., Shiffman, S., Thrul, J., & Schüz, B. (2018). Momentary smoking context as a mediator of the relationship between SES and smoking. *Addictive Behaviors, 83*, 136-141. doi:10.1016/j.addbeh.2017.12.014
- Jones-Webb, R., & Karriker-Jaffe, K. (2013). Neighborhood disadvantage, high alcohol content beverage consumption, drinking norms, and drinking consequences: A mediation analysis. *Journal of Urban Health, 90*(4), 667-684. doi:10.1007/s11524-013-9786-y
- Judd, C. M., & Kenny, D. A. (1981). Process analysis: Estimating mediation in treatment evaluations. *Evaluation Review, 5*(5), 602-619. doi:10.1177/0193841X8100500502
- Keele, L., Tingley, D., & Yamamoto, T. (2015). Identifying mechanisms behind policy interventions via causal mediation analysis. *Journal of Policy Analysis and Management, 34*(4), 937-963. doi:10.1002/pam.21853
- Killeen, P. R. (2011). Markov model of smoking cessation. *Proceedings of the National Academy of Sciences of the United States of America, 108*, 15549. doi:10.1073/pnas.1011277108
- Kim, J. H., Shim, K. W., Yoon, Y. S., Lee, S. Y., Kim, S. S., Oh, S. W., & Sun, Q. (2012). Cigarette smoking increases abdominal and visceral obesity but not overall fatness: An observational study (cigarette smoking and visceral obesity). *7*(9), e45815. doi:10.1371/journal.pone.0045815

- Koslovsky, M. D., Swartz, M. D., Chan, W., Leon-Novelo, L., Wilkinson, A. V., Kendzor, D. E., & Businelle, M. S. (2018). Bayesian variable selection for multistate markov models with interval-censored data in an ecological momentary assessment study of smoking cessation. *Biometrics*, 74(2), 636-644. doi:10.1111/biom.12792
- Larsen, C., & Turkensteen, M. (2014). A vendor managed inventory model using continuous approximations for route length estimates and markov chain modeling for cost estimates. *International Journal of Production Economics; the International Society for Inventory Research*, 2012, 157, 120-132. doi://doi.org/10.1016/j.ijpe.2014.08.001
- Li, S., Fang, L., Zhou, Y., Pan, L., Yang, X., Li, H., . . . Jia, C. (2015). Mediation of smoking abstinence self-efficacy on the association of nicotine dependence with smoking cessation. *European Journal of Public Health*, 25(2), 200. doi:10.1093/eurpub/cku183
- Ma, J., Chan, W., & Tilley, B. C. (2018). Continuous time markov chain approaches for analyzing transtheoretical models of health behavioral change: A case study and comparison of model estimations. *Statistical Methods in Medical Research*, 27(2), 593-607. doi:10.1177/0962280216639859
- Ma, J., Chan, W., Tsai, C., Xiong, M., & Tilley, B. C. (2015). Analysis of transtheoretical model of health behavioral changes in a nutrition intervention study--a continuous time markov chain model with bayesian approach. *Statistics in Medicine*, 34(27), 3577-3589. doi:10.1002/sim.6571

- MacKinnon, D. (2012). *Introduction to statistical mediation analysis* Routledge.
- MacKinnon, D. P., & Fairchild, A. J. (2009). Current directions in mediation analysis. *Current Directions in Psychological Science*, 18(1), 16-20. Retrieved from <http://www.jstor.org.utsph.idm.oclc.org/stable/20695987>
- MacKinnon, D. P., Fairchild, A. J., & Fritz, M. S. (2007). Mediation analysis. *Annual Review of Psychology*, 58(1), 593-614. doi:10.1146/annurev.psych.58.110405.085542
- Matthew S. Fritz, & David P. MacKinnon. (2007). Required sample size to detect the mediated effect. *Psychological Science*, 18(3), 233-239. doi:10.1111/j.1467-9280.2007.01882.x
- Maxwell, S. E., Cole, D. A., & Mitchell, M. A. (2011). Bias in cross-sectional analyses of longitudinal mediation: Partial and complete mediation under an autoregressive model. *Multivariate Behavioral Research*, 46(5), 816-841.
- Mhoon, K. B., Chan, W., Del Junco, D.,J., & Vernon, S. W. (2010). A continuous-time markov chain approach analyzing the stages of change construct from a health promotion intervention. *JP Journal of Biostatistics*, 4(3), 213-226. Retrieved from <https://www.ncbi.nlm.nih.gov/pubmed/23504410>
- Minard, C. G., Chan, W., Wetter, D. W., & Etzel, C. J. (2012). Trends in smoking cessation: A markov model approach. *Journal of Applied Statistics*, 39(1), 113-127. doi:10.1080/02664763.2011.578619

- Murphy, C. M., Rohsenow, D. J., Johnson, K. C., & Wing, R. R. (2018). Smoking and weight loss among smokers with overweight and obesity in look AHEAD. *Health Psychology : Official Journal of the Division of Health Psychology, American Psychological Association*, 37(5), 399. doi:10.1037/hea0000607
- Newsom, J. T., Jones, R. N., & Hofer, S. M. (2013). *Longitudinal data analysis: A practical guide for researchers in aging, health, and social sciences* Routledge.
- Nguyen, T. Q., Webb-Vargas, Y., Koning, I. M., & Stuart, E. A. (2016). Causal mediation analysis with a binary outcome and multiple continuous or ordinal mediators: Simulations and application to an alcohol intervention. *Structural Equation Modeling: A Multidisciplinary Journal*, 23(3), 368-383.
doi:10.1080/10705511.2015.1062730
- Nurwanti, E., & Bai, C. (2018). A population-based study of sedentary behavior, coffee and caffeine intake, education level associated with obesity risk among young adult. *Proceedings of the Nutrition Society*, 77(OCE4), E184.
doi:10.1017/S0029665118001908
- Pan, H., Liu, S., Miao, D., & Yuan, Y. (2018). Sample size determination for mediation analysis of longitudinal data. *BMC Medical Research Methodology*, 18(1), 32-11.
doi:10.1186/s12874-018-0473-2

- Preacher, K. J. (2015). Advances in mediation analysis: A survey and synthesis of new developments. *Annual Review of Psychology*, 66(1), 825-852. doi:10.1146/annurev-psych-010814-015258
- Sato, R. C., & Zouain, D. (2010). Markov models in health care. *Einstein (São Paulo)*, 8, 376-379.
- Schoemann, A. M., Boulton, A. J., & Short, S. D. (2017). Determining power and sample size for simple and complex mediation models. *Social Psychological and Personality Science*, 8(4), 379-386. doi:10.1177/1948550617715068
- Seeley, R., & Sandoval, D. (2011). Weight loss through smoking. *Nature*, 475(7355), 176-7. doi:10.1038/475176a
- Selig, J. P., Preacher, K. J., & Little, T. D. (2012). Modeling time-dependent association in longitudinal data: A lag as moderator approach. *Multivariate Behavioral Research*, 47(5), 697-716.
- Selig, J. P., Preacher, K. J., & Little, T. D. (2012). Modeling time-dependent association in longitudinal data: A lag as moderator approach. *Multivariate Behavioral Research*, 47(5), 697-716.
- Semrau, T., & Sigmund, S. (2012). Networking ability and the financial performance of new ventures: A mediation analysis among younger and more mature firms. *Strategic Entrepreneurship Journal*, 6(4), 335-354. doi:10.1002/sej.1146

- Tada, T., Kumada, T., Toyoda, H., Ohisa, M., Akita, T., & Tanaka, J. (2018). Long-term natural history of liver disease in patients with chronic hepatitis B virus infection: An analysis using the markov chain model. *Journal of Gastroenterology*, 53(11), 1196-1205. doi:10.1007/s00535-018-1467-x
- Tibshirani, R. J., & Efron, B. (1993). An introduction to the bootstrap. *Monographs on Statistics and Applied Probability*, 57, 1-436.
- Valeri, L., & VanderWeele, T. J. (2013). Mediation analysis allowing for exposure–mediator interactions and causal interpretation: Theoretical assumptions and implementation with SAS and SPSS macros. *Psychological Methods*, 18(2), 137.
- VanderWeele, T. J. (2016). Mediation analysis: A practitioner's guide. *Annual Review of Public Health*, 37(1), 17-32. doi:10.1146/annurev-publhealth-032315-021402
- VanderWeele, T. J., & Tchetgen Tchetgen, E. J. (2017). Mediation analysis with time varying exposures and mediators. *Journal of the Royal Statistical Society. Series B, Statistical Methodology*, 79(3), 917-938. doi:10.1111/rssb.12194
- Vanderweele, T. J., & Vansteelandt, S. (2010). Odds ratios for mediation analysis for a dichotomous outcome. *American Journal of Epidemiology*, 172(12), 1339-1348. doi:10.1093/aje/kwq332 [doi]
- Vittinghoff, E., & Neilands, T. (2015). Sample size for joint testing of indirect effects. *Prevention Science*, 16(8), 1128-1135. doi:10.1007/s11121-014-0528-5

Vittinghoff, E., & Neilands, T. (2015). Sample size for joint testing of indirect effects.

Prevention Science, 16(8), 1128-1135. doi:10.1007/s11121-014-0528-5

Vittinghoff, E., Sen, S., & McCulloch, C. E. (2009). Sample size calculations for evaluating mediation. *Statistics in Medicine*, 28(4), 541-557. doi:10.1002/sim.3491

Voola, R., Casimir, G., Carlson, J., & Anushree Agnihotri, M. (2012). The effects of market orientation, technological opportunism, and e-business adoption on performance: A moderated mediation analysis. *Australasian Marketing Journal (AMJ)*, 20(2), 136-146. doi://doi.org/10.1016/j.ausmj.2011.10.001

Wang, C., & Xue, X. (2016). Power and sample size calculations for evaluating mediation effects in longitudinal studies. *Statistical Methods in Medical Research*, 25(2), 686-705. doi:10.1177/0962280212465163

Xinshu Zhao, John G. Lynch, & Qimei Chen. (2010). Reconsidering baron and kenny: Myths and truths about mediation analysis. *Journal of Consumer Research*, 37(2), 197-206. doi:10.1086/651257