Power consideration and caveats of Causal Inference Test (CIT) for mediation analysis

Tianjie Gu

A thesis
submitted in partial fulfillment of the
requirements for the degree of

Master of Science

University of Washington

2021

Committee:

James Y. Dai

Li Hsu

Program Authorized to Offer Degree:
Biostatistics - Public Health

University of Washington

Abstract

Power consideration and caveats of
Causal Inference Test (CIT)
for mediation analysis

Tianjie Gu

Chair of the Supervisory Committee:
James Y. Dai
Biostatistics

Causal Inference Test (CIT) has been proposed to infer the causality mediation from a genotype to a trait through a candidate mediator, for example certain gene expression. This thesis aims to study the performance of CIT under diverse settings including partial mediation where the temporal order of the candidate mediator and the trait cannot be determined. Through algebraic derivation and simulation, I show that there are scenarios of partial mediation that CIT will yield false negative results even for a large sample size, and a scenario where CIT will lead to false positives. In conclusion, CIT may not reliably detect the direction of the mediation.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGMENTS

First and foremost, I would also like to express my deepest appreciation to my advisor Dr. James Y. Dai, for being a wonderful advisor and mentor, for all the precious advice, encouragement and support throughout this journey. I would also like to thank my second reader Dr. Li Hsu for her valuable feedback.

I would like to extend my sincere thanks to Gitana Garofalo for her devotion to making the graduate program a warm family and for her relentless kind-hearted help.

Finally, I must express my very profound gratitude to my friends and my family for providing me with unfailing support and continuous encouragement throughout my years of study and through the process of researching and writing this thesis. This accomplishment would not have been possible without them.

DEDICATION

to my family

Chapter 1
INTRODUCTION

Through Genome-Wide Association Study(GWAS), thousands of single nucleotide polymorphism(SNP) have been reported to be associated with a wide range of phenotypes like disease, molecular measures, quantitative traits, and social and behavioral traits [3, 22, 27]. SNP is a single nucleotide difference at a specific position in the DNA sequence found in the human population. To its extreme, such a single base change in a protein coding region may alter amino acid sequence and lead to a dramatic physiological malfunction, like sickle cell disease, the most common severe monogenic disorders in the world [19]. Although most of SNPs don't directly change amino acid sequences, it still can affect many steps along the process from DNA to protein: mRNA expression, splicing, stability, translation, and gene expression [10, 21, 23]. The challenge remains to study the causality and pathway of the disease development, especially cancer, diabetes, cardiovascular disease, psychiatric issues, and other common diseases which are under influences of multiple genetic loci and environmental factors [10].

Because of the observational nature of many genetic association studies , data collection is typically cross-sectional; for instance, tissue collected at diagnosis for obtaining DNA sequence and DNA expression data. Although genotypes no doubt precede other traits, the temporal order of an intermediate variable (e.g., gene expression) and disease diagnosis cannot be determined. To infer the causality from a genotype to a clinical trait through a potential mediator like certain gene expression, Causal Inference Test (CIT) [15] was proposed and provides a quantitative measure of uncertainty in the form of a formal p-value and a call for causal direction. Using this method, genetic variants have been identified

to affect rheumatoid arthritis risk and metabolic traits in adipose tissue via altered DNA methylation [13, 28], and to be associated with antibody response to vaccination mediated by gene expression [7].

1.1 Notations and 5 Models

Let G denote a genotype marker at a specific locus. Let X be a potential intermediate and Y be an outcome of interest: Y could be molecular trait like DNA expression, metabolic or clinical traits. The order of the causal effect between X and Y is yet to be determined. In this case, G is the exposure and cannot be the result of X or Y. The alternative relationships between these three factors are: 1) X mediates the effect of G on Y (causal mediation model); 2) X is the consequence of Y which is under effect of G (reactive model); 3) the association between G and X is independent of the association between G and Y (independent model). Because of the complexity of the biological process, it is rare to have pure causal (or reactive) model where the effect of G on Y (or X) is completely mediated through X (or Y). Thus, for practical reasons, partial mediation model and partial reactive model should also be considered when studying the relation among genotype markers, intermediates and measured traits or molecular markers. More comprehensive models may also involve a hidden factor; or the trait is independent of genotype which is associated with intermediate. In this thesis, the discussion of the candidate models among the three variables is limited to the five model displayed in Figure (1.1).

1.2 Causal inference test (CIT)

Adapting from Schadt's model selection approach and Chen's Causality Equivalence Theorem, Millstein [15] used four mathematical conditions to define a causal mediation relationship, G acting on Y through X: 1) G and Y are associated, 2) G is associated with X conditional on Y, 3) X is associated with Y conditional on G, and 4) G is independent of Y

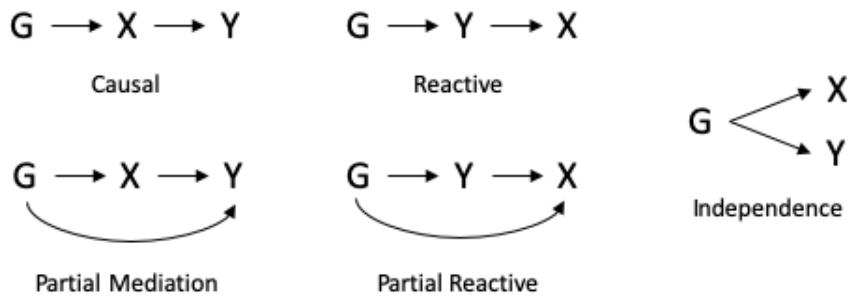


Figure 1.1: Possible relationships between an exposure, a candidate mediator and an outcome. Genotype as exposure is denoted by G, candidate intermediate is denoted by X, and Y denotes the outcome, i.e., the measured trait.

conditional on X (Figure 1.1). Each condition can be assessed by one statistical test, using the following linear regression models:

$$\mathbb{E}[Y_i|G_i] = \alpha_1 + \beta_1 G_i \tag{1.1}$$

$$\mathbb{E}[X_i|G_i, Y_i] = \alpha_2 + \beta_2 Y_i + \beta_3 G_i \tag{1.2}$$

$$\mathbb{E}[Y_i|G_i, X_i] = \alpha_3 + \beta_4 X_i + \beta_5 G_i \tag{1.3}$$

And the hypotheses of four component tests for each condition are:

$$1)\ H_0 : \beta_1 = 0, H_1 : \beta_1 \neq 0$$

$$2)\ H_0 : \beta_3 = 0, H_1 : \beta_3 \neq 0$$

$$3)\ H_0 : \beta_4 = 0, H_1 : \beta_4 \neq 0$$

$$4)\ H_0 : \beta_5 \neq 0, H_1 : \beta_5 = 0$$

Standard F test is used for condition 1 through 3, where each set of parameters equal to zero in the respective null hypothesis. For the test involving model (1.2) and (1.3), partial F test is applied. For condition 4, an equivalence test is conducted and null F^*-statistic distribution is constructed through bootstrap. Only causal model is expected to reject all of

four null hypotheses, while reactive and independent model would fail to reject at least one of them (Table 1.1).

	Null Hypotheses				
	$G \perp\!\!\!\perp Y$	$G \perp\!\!\!\perp X \mid Y$	$X \perp\!\!\!\perp Y \mid G$	$G \sim Y \mid X$	CIT
Causal Model	R	R	R	R	R
Reactive Model	R	FR	R	R	FR
Independent Model	R	R	FR	FR	FR

Table 1.1: Expected component test status of Causal, Reactive and Independent model. Under three basic models, rejection(R) or failure to reject(FR) of the null hypothesis of each component test.

To conduct component test 4, which is an equivalence test, generally the equivalence interval need to be specified, that is, a close neighborhood around the parameter that is considered to be compatible with the true null value [30]. In this case, the equivalence interval is set as the parameter space closer to zero than what would be expected under the independence model, which is defined by three conditions: 1) G is causal for X, 2) G is causal for Y, and 3) Y is independent of $X|G$ (Figure 1.1). CIT uses an ingenious trick to circumvent the standard equivalence test: it generates the null distribution of the test statistic conditional on the observed $Y \sim G$ relationship under the independence model (null model) via bootstrap. the significance of the observed statistic is assessed by whether it is significantly smaller than what expected under the null:

Step 1: The marginal effect of G on X can be estimated from linear regression $X_i = \alpha_4 + \beta_6 G_i + \epsilon_x$, where ϵ_x is the residuals. The residuals are randomly permuted and denoted as ϵ_x^*. Note that, because of the random permutation, ϵ_x^* is independent of $Y|G$. Therefore, $X_i^* = \alpha_4 + \beta_6 G_i + \epsilon_x^*$ is also independent of $Y|G$ and the triplet $\{G, X^*, Y\}$ satisfies the three conditions for the Independent model.

Step 2: Two nested models

$$Y_i = \alpha_3^* + \beta_4^* X_i^* + \beta_5^* G_i + \epsilon_i^*$$

$$Y_i = \alpha_3^* + \beta_4^* X_i^* + \epsilon_i^*$$

are fitted and test statistic F^* is computed under the null distribution.

Step 3: Step 1 and 2 would be repeated B times and an empirical distribution of F^* under the independent model would be constructed. With a large number of repeats B, the observed test statistic F can be compared directly with the empirical distribution of F^*. On the other hand, under limited permutation times, the null distribution can be approximated using the parametric property of F distribution and a small B sample of F^*. A method of moment estimator for the non-centrality parameter of the F distribution under the null can be calculated,

$$\hat{\lambda} = \frac{\overline{F}^* \nu_1 (\nu_2 - 2)}{\nu_2} - \nu_1$$

where ν_1 and ν_2 denote the degrees of freedom. The distribution is then transformed to a normal distribution,

$$Z^* = \Phi^{-1} \left(P \left[X < F^* \right] \right)$$

$$Z^* \sim N \left(0, \tau^2 \right)$$

where $X \sim F_{\nu_1, \nu_2, \hat{\lambda}}$, and the variance τ^2 can be estimated from the empirical distribution of Z^*. The same transformation is also done to the observed test statistic F,

$$Z = \Phi^{-1} \left(P \left[X < F \right] \right)$$

The transformed test statistic Z is tested against $N(0, \tau^2)$. Thus the p-value for the fourth condition equals $P[Z^* < Z]$.

The Causal model is concluded only if all of the four component tests are rejected. That is, the null hypothesis of the CIT is the union of the H_0 from the four component tests, and rejection region for the CIT is the intersection of four H_1. Thus, CIT can be seen as

an intersection-union test and the p-value for the CIT is the supremum of the four p-values for the component tests. As shown in Table (1.1), CIT is expected to distinguish causal model from reactive model and independent model as only true causal model is anticipated to result a significant CIT p-value.

Millstein also extend CIT into an informal model selection method to call the direction of mediation (Table 1.2). Switching the order of Y and X, the p-value can be computed for both the forward direction (G to X to Y) and reverse direction (G to Y to X) as shown in Figure 1.1. The four null hypotheses of reverse-direction component tests are: 1. G is independent of X; 2. G is independent of $Y|X$; 3. X is independent of $Y|G$; 4. G is associated with $X|Y$. If only one of the p-values is significant at 0.05, the corresponding model would be returned as the call. If both p-values are non-significant, the call is independent and if both are significant, the call is "no call". One caveat of the call for mediation direction is that CIT did not consider partial mediation and partial reactive model in its final call, so the independent call also includes other non-causal and non-reactive models, such as trait(Y) completely independent of genotype(G).

1.3 Aims of this thesis

In this thesis, I aim to study the performance of the CIT under diverse settings including partial mediation and partial reactive models. The CIT test is not designed to address partial mediation and its performance under the latter two scenarios is yet to be determined. In my view, complete mediation is nearly impossible and partial mediation is likely true in most biological systems, therefore the partial causal mediation model should be carefully studied and perhaps ruled as a causal mediation model. However, partial mediation model would not meet the fourth condition of the CIT test, i.e., G is independent of $Y|X$. Thus, the performance of CIT under partial mediation model is questionable. Note that the similarity

	Null Hypotheses										Model selection
	Forward direction					Reverse direction					
	$G \perp\!\!\!\perp Y$	$G \perp\!\!\!\perp X \mid Y$	$X \perp\!\!\!\perp Y \mid G$	$G \sim Y \mid X$	CIT	$G \perp\!\!\!\perp X$	$G \perp\!\!\!\perp Y \mid X$	$X \perp\!\!\!\perp Y \mid G$	$G \sim X \mid Y$	CIT	
Causal Model	R	R	R	R	R	R	FR	R	R	FR	Causal
Reactive Model	R	FR	R	FR	FR	R	R	R	R	R	Reactive
Independent Model	R	R	FR	FR	FR	R	R	FR	FR	FR	Independent

Table 1.2: Expected CIT and component tests result of Causal, Reactive and Independent models. Rejection(R) or failure to reject(FR) of the null hypothesis of each component test in both forward and reverse direction.

between partial mediation model and partial reactive model, switching the direction of the effect between X and Y. So it is also of interest to study whether CIT could correctly detect the direction of mediation under the two partial models, which are more common in real studies. In this thesis, I would use CIT p-value as the primary significance criterion, but also consider its model selection extension as it is now used in the literature [7]. In Chapter 2, the properties of CIT will be algebraically derived under different models including partial mediation and partial reactive models. Simulations will be performed in Chapter 3 to evaluate its performance, followed by a real data applied with CIT in Chapter 4. Finally, I will present concluding remarks that summarize this investigation in Chapter 5.

Chapter 2
THEORETICAL DERIVATION

In this chapter, I will examine the characteristics of CIT under different models using algebraic derivation. The special emphasis is its properties under true data-generating partial mediation and partial reactive models. In the original CIT paper, biallelic markers with co-dominance were considered [15]. Here we extend the notation of G to be any type of genotype data: binary as dominate/recessive alleles; or continuous between 0 and 1 as imputed genetic score. Without loss of generality, the intercepts in the models (α_j) are set to 0 as the predictors and the outcome can be centered at their means. Thus, the linear regression models used in CIT, model (1.1) to (1.3), can be simplified:

$$\mathbb{E}[Y_i|G_i] = \beta_1 G_i \tag{2.1}$$

$$\mathbb{E}[X_i|G_i, Y_i] = \beta_2 Y_i + \beta_3 G_i \tag{2.2}$$

$$\mathbb{E}[Y_i|G_i, X_i] = \beta_4 X_i + \beta_5 G_i \tag{2.3}$$

And the four component tests are:

$$1)\ H_0 : \beta_1 = 0, H_1 : \beta_1 \neq 0$$

$$2)\ H_0 : \beta_3 = 0, H_1 : \beta_3 \neq 0$$

$$3)\ H_0 : \beta_4 = 0, H_1 : \beta_4 \neq 0$$

$$4)\ H_0 : \beta_5 \neq 0, H_1 : \beta_5 = 0$$

Since the CIT p-value is the supremum of four p-values from the component tests, failure to reject even only one test would lead to a non-significant CIT test result. In this Chapter, I will study the power of CIT under causal and partial mediation models, and I would look

into each of four component tests and list scenarios it may result in false negative. Indeed, as I show later that it is possible under several scenarios including the partial models that CIT will never make the correct call no matter how large the sample size is.

2.1 Causal mediation Model

Suppose G_i takes a Bernoulli distribution with probability p. Suppose the true data-generating model is that G_i causes changes in X_i which then leads to changes in Y_i, that is, a causal mediation model of G (exposure), X (mediator) and Y (outcome):

$$G_i \sim B(1, p)$$

$$X_i = \eta_1 G_i + \epsilon_{1i}$$

$$Y_i = \eta_3 X_i + \epsilon_{2i}$$

where the ϵ_{1i} and ϵ_{2i} represent independent random error, distributed as a Gaussian distribution with mean zero and variance τ_1^2 and τ_2^2. Later I consider the scenario that error distributions are arbitrary under regularity conditions.

2.1.1 Component test 1

Applying the iteration formulas to obtain the expected value and variance of Y condition on G,

$$\begin{aligned}
\mathbb{E}[Y|G] &= \mathbb{E}\left\{\mathbb{E}\left[Y|G,X\right]|\,G\right\} \\
&= \mathbb{E}\left\{\mathbb{E}\left[Y|X\right]|G\right\} \\
&= \mathbb{E}\left[\eta_3 X|\,G\right] \\
&= \eta_3 \mathbb{E}[X|G] \\
&= \eta_1\eta_3 G \\
\mathrm{Var}[Y|G] &= \mathbb{E}\left\{\mathrm{Var}\left[Y|G,X\right]|\,G\right\} + \mathrm{Var}\left\{\mathbb{E}\left[Y|G,X\right]|\,G\right\}
\end{aligned}$$

$$
\begin{aligned}
&= \mathbb{E}\left\{\operatorname{Var}\left[Y|X\right]|\, G\right\} + \operatorname{Var}\left\{\mathbb{E}\left[Y|X\right]|\, G\right\} \\
&= \tau_2^2 + \operatorname{Var}\left[\eta_3 X|G\right] \\
&= \tau_2^2 + \eta_3^2\tau_1^2
\end{aligned}
$$

Therefore, under the causal model and condition on G, Y has a normal distribution:

$$Y|G \sim \mathcal{N}\left(\eta_1\eta_3 G, \eta_3^2\tau_1^2 + \tau_2^2\right) \tag{2.4}$$

For model (2.1), F test is used to test the null hypothesis $\beta_1 = \eta_1\eta_3 = 0$. Under the alternative hypothesis, the F statistic follows a non-central F distribution with degree of freedom 1 and $n-2$ and non-centrality parameter λ. In simple regression problem, λ can be simplified [20, 29]:

$$\lambda = (n-1)\left(\frac{\eta_1^2\eta_3^2}{\eta_3^2\tau_1^2 + \tau_2^2}\right) SD_G^2 \tag{2.5}$$

where, SD_G^2 is the sample variance of G. The power of F test increases with the non-centrality parameter λ, that is, with sample size n, or with the "effect size" of the parameter scaled by the standard error of the random error. With the true underlying distribution (2.4), the latter equals to the ratio of $\eta_1\eta_3$ to $(\eta_3^2\tau_1^2 + \tau_2^2)^{1/2}$. Thus, besides small sample size, the test could have low power when ηs are small and/or τs are large. This test also implies that the CIT test always has lower power than the genetic association of G and Y. For complex diseases, effective sizes of SNPs based genetic association are typically small with limited power. This becomes a practical limitation of CIT in genetic studies.

2.1.2 Component test 2

Applying the iteration rule to the covariance of X and Y conditional on G,

$$
\begin{aligned}
\operatorname{Cov}\left[X_i, Y_i|G_i\right] &= \mathbb{E}\left[XY|G\right] - \mathbb{E}\left[X|G\right]\mathbb{E}\left[Y|G\right] \\
&= \mathbb{E}\left\{\mathbb{E}\left[XY|G, X\right]|\, G\right\} - \mathbb{E}\left[X|G\right]\mathbb{E}\left[Y|G\right] \\
&= \mathbb{E}\left\{X\mathbb{E}\left[Y|G, X\right]|\, G\right\} - \mathbb{E}\left[X|G\right]\mathbb{E}\left[Y|G\right]
\end{aligned}
$$

$$
\begin{aligned}
&= \mathbb{E}\left[\eta_3 X^2|G\right] - \mathbb{E}\left[X|G\right]\mathbb{E}\left[Y|G\right] \\
&= \eta_3\left(\mathrm{Var}\left[X|G\right] + \left[\mathbb{E}\left[X|G\right]\right]^2\right) - \mathbb{E}\left[X|G\right]\mathbb{E}\left[Y|G\right] \\
&= \eta_3\left(\tau_1^2 + \eta_1^2 G^2\right) - \eta_1^2\eta_3 G^2 \\
&= \eta_3\tau_1^2
\end{aligned}
$$

Thus, condition on G, the pair $\{X, Y\}$ have a bivariate normal distribution:

$$
\begin{pmatrix} X \\ Y \end{pmatrix} \Bigg| G \sim \mathcal{N}\left[\begin{pmatrix} \eta_1 G \\ \eta_1\eta_3 G \end{pmatrix}, \begin{pmatrix} \tau_1^2 & \eta_3\tau_1^2 \\ \eta_3\tau_1^2 & \eta_3^2\tau_1^2 + \tau_2^2 \end{pmatrix}\right]
$$

Then, the distribution of X condition on both G and Y:

$$
\begin{aligned}
X|Y, G &\sim \mathcal{N}\left(\eta_1 G + \frac{\eta_3\tau_1^2}{\eta_3^2\tau_1^2 + \tau_2^2}[Y - \eta_1\eta_3 G], \tau_1^2 - \frac{\eta_3^2\tau_1^4}{\eta_3^2\tau_1^2 + \tau_2^2}\right) \\
&\sim \mathcal{N}\left(\frac{\eta_1\tau_2^2}{\eta_3^2\tau_1^2 + \tau_2^2} G + \frac{\eta_3\tau_1^2}{\eta_3^2\tau_1^2 + \tau_2^2} Y, \frac{\tau_1^2\tau_2^2}{\eta_3^2\tau_1^2 + \tau_2^2}\right)
\end{aligned} \tag{2.6}
$$

For model (2.2), F test is used to test the null hypothesis $\beta_3 = \eta_1\tau_2^2/(\eta_3^2\tau_1^2 + \tau_2^2) = 0$. Under the alternative hypothesis, F statistic has a non-central F distribution with degree of freedom 1 and $n-3$. And the non-centrality parameter in multivariate regression equals [20, 29]:

$$
\lambda = (n-1)\left(\frac{\eta_1^2\tau_2^2}{\tau_1^2\left(\eta_3^2\tau_1^2 + \tau_2^2\right)}\right) SD_G^2\left(1 - R_{G,Y}^2\right)
$$

where, SD_G^2 is the sample variance of G and $R_{G,Y}^2$ is the value of R^2 for the OLS regression with response G and regressor Y, and the "effect size" of the parameter $(\eta_1\tau_2^2/(\eta_3^2\tau_1^2 + \tau_2^2))$ scaled by the standard error of the random error $(\tau_1\tau_2/\sqrt{\eta_3^2\tau_1^2 + \tau_2^2})$. Again, the power of the test would be low when λ is small, that is, small sample size, small value of $\eta_1^2\tau_2^2$ relative to $\tau_1^2\left(\eta_3^2\tau_1^2 + \tau_2^2\right)$, or small unexplained variability in G after Y.

2.1.3 Component test 3

$$
Y|G, X \sim \mathcal{N}\left(\eta_3 X, \tau_2^2\right)
$$

Null hypothesis $\beta_4 = \eta_3 = 0$ is tested in the model (2.3). Under the alternative hypothesis, F statistic has a non-central F distribution with degree of freedom 1 and $n-3$ and the

non-centrality parameter

$$\lambda = (n-1)\left(\frac{\eta_3}{\tau_2}\right)^2 SD_G^2\left(1 - R_{X,G}^2\right)$$

where, SD_G^2 is the sample variance of G and $R_{X,G}^2$ is the value of R^2 for the OLS regression with response X and regressor G. Thus, the power of the test decreases with decreasing ratio of η_3 to τ_2, apart from smaller sample size and lower unexplained variability in X after G.

2.1.4 Component test 4

To test the hypothesis $\beta_5 \neq 0$ in model (2.3), X^* is simulated to construct triplets that follow an independent model :

$$\begin{pmatrix} X^* \\ Y \end{pmatrix} \Bigg| G \sim \mathcal{N}\left[\begin{pmatrix} \hat{\eta}_1 G \\ \eta_1\eta_3 G \end{pmatrix}, \begin{pmatrix} \xi^2 & 0 \\ 0 & \eta_3^2\tau_1^2 + \tau_2^2 \end{pmatrix}\right]$$

where $\hat{\eta}_1$ is β estimate in equation $X = \beta G + \epsilon$ and ξ^2 is the variance of $X^*|G$. Through bootstrap, the null distribution of F^* for the fourth component test is computed comparing two nested models:

$$Y_i = \alpha_3^* + \beta_4^* X_i^* + \beta_5^* G_1 + \epsilon_i^*$$

$$Y_i = \alpha_3^* + \beta_4^* X_i^* + \epsilon_i^*$$

that is, testing $\eta_1\eta_3 = 0$, which can be approximated by a non-central distribution with degree of freedom 1 and $n-3$, and the non-central parameter:

$$\lambda^* = (n-1)\left(\frac{\eta_1^2\eta_3^2}{\eta_3^2\tau_1^2 + \tau_2^2}\right) SD_G^2\left(1 - R_{G,X^*}^2\right)$$

where, SD_G^2 is the sample variance of G and R_{G,X^*}^2 is the value of R^2 for the OLS regression with response G and regressor X^*. After transforming to a normal distribution, the probability of simulated statistics being less than the observed test statistic is considered as the p-value for this test. With the causal mediation model, the observed F statistic is expected to follow a centered F distribution. The larger λ^* is, the smaller the area of left tail under

the curve, that is, more likely to reject the null hypothesis. This suggests that a small value of λ^* could lead to low power in test 4: small sample size, or small ratio of $\eta_1^2\eta_3^2$ to $\eta_3^2\tau_1^2 + \tau_2^2$, or small unexplained variability in G after X^*. Notice, the non-centrality parameter in this component test is a faction of the one in component test 1. Therefore, if a parameter space results low power in component test 1, it is likely to have a low power in the component test 4, too

2.1.5 CIT and model selection

Since the p-value of CIT is the supremum of all p-values from the four component tests, the power of CIT can be low even if only one of the component tests has a low power. As introduced earlier in Chapter 1, CIT can be fitted in two directions: $G \to X \to Y$ and $G \to Y \to X$. To obtain the p-value for the reverse direction $G \to Y \to X$, the second component test is to test the null hypothesis, G is independent of Y conditional on X, i.e., $\beta_3' = 0$ in the model $Y_i = \beta_2' X_i + \beta_3' G_i + \epsilon_{2i}'$. When the truth is the causal mediation model in the direction of $G \to X \to Y$, this null hypothesis would not be rejected, that is, the p-value under the reverse diction is expected to be rightfully above 0.05. Thus, as sample size increases, CIT will correctly call the forward mediation direction $G \to X \to Y$.

2.2 Partial Mediation Model

Suppose G_i takes a Bernoulli distribution with probability p. Suppose the true data-generating model is that G has effect on Y, which partially mediated by X:

$$\begin{aligned} G_i &\sim B(1, p) \\ X_i &= \eta_1 G_i + \epsilon_{1i} \\ Y_i &= \eta_2 G_i + \eta_3 X_i + \epsilon_{2i} \end{aligned} \tag{2.7}$$

where the ϵ_{1i} and ϵ_{2i} represent independent random error, distributed as a Gaussian distribution with mean zero and variance τ_1^2 and τ_2^2.

2.2.1 Component test 1

Applying iteration formulas to obtain the expected value and variance of Y condition on G,

$$\begin{aligned}
\mathbb{E}[Y|G] &= \mathbb{E}\left\{\mathbb{E}\left[Y|G,X\right]|\,G\right\} \\
&= \mathbb{E}\left[\eta_2 G + \eta_3 X|\,G\right] \\
&= \eta_2 G + \eta_3 \mathbb{E}[X|G] \\
&= \eta_2 G + \eta_1\eta_3 G \\
&= \left(\eta_2 + \eta_1\eta_3\right) G \\
\mathrm{Var}[Y|G] &= \mathbb{E}\left\{\mathrm{Var}\left[Y|G,X\right]|\,G\right\} + \mathrm{Var}\left\{\mathbb{E}\left[Y|G,X\right]|\,G\right\} \\
&= \mathbb{E}\left\{\mathrm{Var}\left[Y|X\right]|\,G\right\} + \mathrm{Var}\left\{\mathbb{E}\left[Y|X\right]|\,G\right\} \\
&= \tau_2^2 + \mathrm{Var}\left[\eta_3 X|G\right] \\
&= \tau_2^2 + \eta_3^2\tau_1^2
\end{aligned}$$

Therefore, under the partial mediation model, condition on G, Y has the following normal distribution:

$$Y|G \sim \mathcal{N}\left[\left(\eta_2 + \eta_1\eta_3\right) G, \eta_3^2\tau_1^2 + \tau_2^2\right] \tag{2.8}$$

Same as in the causal model, under the alternative hypothesis, the F statistic has a non-central F distribution with degree of freedom 1 and $n-2$. With (2.5) and (2.8), the non-centrality parameter:

$$\lambda = (n-1)\left(\frac{\left(\eta_2 + \eta_1\eta_3\right)^2}{\eta_3^2\tau_1^2 + \tau_2^2}\right) SD_G^2 \tag{2.9}$$

where, SD_G^2 is the sample variance of G. The power of the test decreases with decreasing λ, that is, small total effect of G on Y relative to the standard deviation of Y condition on G. Importantly, it is numerically possible that λ would equal to 0; that is when $\eta_2 = -\eta_1\eta_3$. In

this numerical incidence, component test 1 will have no power to reject no matter how large the sample size. Moreover, when tThe direct effect of G on Y is opposite to the mediated effect of G on Y through X, the power is low as the overall genetic effect will be small with the same size. This is a serious drawback of CIT that the test has low or no power to detect mediation when the direct effect and the indirect effect are in the opposite direction.

2.2.2 Component test 2

Applying the iteration rule to the covariance of X and Y conditional on G,

$$\begin{aligned}
\mathrm{Cov}\left[X_i, Y_i|G_i\right] &= \mathbb{E}\left[XY|G\right] - \mathbb{E}\left[X|G\right]\mathbb{E}\left[Y|G\right] \\
&= \mathbb{E}\left\{\mathbb{E}\left[XY|G, X\right]|\, G\right\} - \mathbb{E}\left[X|G\right]\mathbb{E}\left[Y|G\right] \\
&= \mathbb{E}\left\{X\mathbb{E}\left[Y|G, X\right]|\, G\right\} - \mathbb{E}\left[X|G\right]\mathbb{E}\left[Y|G\right] \\
&= \mathbb{E}\left[X\left(\eta_2 G + \eta_3 X\right)|G\right] - \mathbb{E}\left[X|G\right]\mathbb{E}\left[Y|G\right] \\
&= \eta_2 G\mathbb{E}\left[X|G\right] + \eta_3\left(\mathrm{Var}\left[X|G\right] + \left[\mathbb{E}\left[X|G\right]\right]^2\right) - \mathbb{E}\left[X|G\right]\mathbb{E}\left[Y|G\right] \\
&= \eta_1\eta_2 G^2 + \eta_3\left(\tau_1^2 + \eta_1^2 G^2\right) - \left(\eta_1 G\right)\left(\eta_2 G + \eta_1\eta_3 G\right) \\
&= \eta_3\tau_1^2
\end{aligned}$$

Therefore, condition on G, the pair $\{X, Y\}$ have a bivariate normal distribution:

$$\left.\begin{pmatrix} X \\ Y \end{pmatrix}\right| G \sim \mathcal{N}\left[\begin{pmatrix} \eta_1 G \\ \left(\eta_2 + \eta_1\eta_3\right) G \end{pmatrix}, \begin{pmatrix} \tau_1^2 & \eta_3\tau_1^2 \\ \eta_3\tau_1^2 & \eta_3^2\tau_1^2 + \tau_2^2 \end{pmatrix}\right]$$

Then, the distribution of X condition on both G and Y:

$$\begin{aligned}
X|Y, G &\sim \mathcal{N}\left(\eta_1 G + \frac{\eta_3\tau_1^2}{\eta_3^2\tau_1^2 + \tau_2^2}[Y - \left(\eta_2 + \eta_1\eta_3\right) G], \tau_1^2 - \frac{\eta_3^2\tau_1^4}{\eta_3^2\tau_1^2 + \tau_2^2}\right) \\
&\sim \mathcal{N}\left(\frac{\eta_1\tau_2^2 - \eta_2\eta_3\tau_1^2}{\eta_3^2\tau_1^2 + \tau_2^2} G + \frac{\eta_3\tau_1^2}{\eta_3^2\tau_1^2 + \tau_2^2} Y, \frac{\tau_1^2\tau_2^2}{\eta_3^2\tau_1^2 + \tau_2^2}\right) \qquad (2.10)
\end{aligned}$$

Under the alternative hypothesis, the F statistic has a non-central F distribution with degree of freedom 1 and $n - 3$. With underlying distribution (2.10), the non-centrality parameter

can be written as:

$$\lambda = (n-1)\, SD_G^2 \left(1 - R_{G,Y}^2\right) \left(\frac{\left(\eta_1 \tau_2^2 - \eta_2 \eta_3 \tau_1^2\right)^2}{\tau_1^2 \tau_2^2 \left(\eta_3^2 \tau_1^2 + \tau_2^2\right)} \right)$$

where, SD_G^2 is the sample variance of G and $R_{G,Y}^2$ is the value of R^2 for the OLS regression with response G and regressor Y. Again, the power decreases with decreasing λ. In special case, λ can equal to 0 when

$$\eta_1 \tau_2^2 = \eta_2 \eta_3 \tau_1^2$$
$$\frac{\eta_1}{\eta_2 \eta_3} = \frac{\tau_1^2}{\tau_2^2}$$

If this numerical coincidence happens and no matter how large the sample is, CIT would have no power at all to detect the causal relation between $\{G, X, Y\}$ even the true model is partial mediation model. While numerical coincidence rarely happens, in realty the power for the second component test can be low so that CIT will likely fail to reject. This could be another drawback of CIT test that will fail to detect partial mediation.

2.2.3 Component test 3

As in the set-up of the partial mediation model:

$$Y|G, X \sim \mathcal{N}\left(\eta_2 G + \eta_3 X, \tau_2^2\right)$$

Similar to the component test 3 under causal model, the non-centrality parameter

$$\lambda = (n-1) \left(\frac{\eta_3}{\tau_2}\right)^2 SD_G^2 \left(1 - R_{X,G}^2\right)$$

where, SD_G^2 is the sample variance of G and $R_{X,G}^2$ is the value of R^2 for the OLS regression with response X and regressor G. Thus, apart from smaller sample size and lower unexplained variability in X after G, the power of the test decreases with decreasing ratio of effect of X on Y to the standard error of random error deviation of $Y|G, X$ (η_3/τ_2).

2.2.4 Component test 4

To test the hypothesis $\beta_5 \neq 0$ in model (2.3), X^* is simulated to construct an independent model with $\{G, X^*, Y\}$:

$$\begin{pmatrix} X^* \\ Y \end{pmatrix} \Bigg| G \sim \mathcal{N}\left[\begin{pmatrix} \hat{\eta}_1 G \\ (\eta_2 + \eta_1\eta_3)G \end{pmatrix}, \begin{pmatrix} \xi^2 & 0 \\ 0 & \eta_3^2\tau_1^2 + \tau_2^2 \end{pmatrix} \right]$$

where $\hat{\eta}_1$ is β estimate in equation $X = \beta G + \epsilon$ and ξ^2 is the variance of $X^*|G$. Through bootstrap, the distribution of F^* is computed comparing two nested models:

$$Y_i = \alpha_3^* + \beta_4^* X_i^* + \beta_5^* G_1 + \epsilon_i^*$$

$$Y_i = \alpha_3^* + \beta_4^* X_i^* + \epsilon_i^*$$

which should close to a non-central distribution with degree of freedom 1 and $n-3$, and non-central parameter:

$$\lambda^* = (n-1)\left(\frac{(\eta_2 + \eta_1\eta_3)^2}{\eta_3^2\tau_1^2 + \tau_2^2}\right) SD_G^2 \left(1 - R_{G,X^*}^2\right)$$

where, SD_G^2 is the sample variance of G and R_{G,X^*}^2 is the value of R^2 for the OLS regression with response G and regressor X^*. After transforming to normal distribution, the probability of less than the observed test statistic is considered as the p-value for this test.

When the truth is the partial mediation model,

$$Y|G, X \sim \mathcal{N}\left(\eta_2 G + \eta_3 X, \tau_2^2\right)$$

the observed F statistic is expect to follow a F distribution with same degree of freedom, but the non-central parameter is

$$\lambda = (n-1)\left(\frac{\eta_2^2}{\tau_2^2}\right) SD_G^2 \left(1 - R_{G,X}^2\right)$$

where, SD_G^2 is the sample variance of G and $R_{G,X}^2$ is the value of R^2 for the OLS regression with response G and regressor X. Compared to λ^*, the smaller the λ, the more likely the left tail of the observed F statistic under the computed F^* distribution is below 0.05, therefore rejecting the 4th component test. On the other hand, if λ is close to λ^*, it is less likely to

result a significantly small p-value. That is, with large sample size, the power of test 4 would be low if

$$\frac{(\eta_2 + \eta_1\eta_3)^2}{\eta_3^2\tau_1^2 + \tau_2^2} = \frac{\eta_2^2}{\tau_2^2}$$

The special condition discussed in the component test 1, $\eta_2 = -\eta_1\eta_3$, would also lead to a low or no power in test 4. In this case, it is implied that the β_5 equals 0 under the constructed independent model. The null distribution of the test statistic under independent model is centered F distribution ($\lambda^* = 0$) while the observed F statistic follows a non-central F distribution. The area under the null distribution left to the observed test statistic is impossible to be below 0.05.

2.2.5 CIT and model selection

If CIT for the forward direction $G \to X \to Y$ fails to reject component test 3, then this would result in misclassifyingmake the partial mediation model appear to be theas independent model becauseas the conditional independence condition tested by the component test 3 is the equivalent (switchingchanging the order of Y and X) under the forward direction and under the reverse direction. In addition, the component test 2 in the forward direction $G \to X \to Y$ is opposite to the component test 4 in the reverse direction $G \to Y \to X$. Thus, if component test 2 in the forward direction is the only test has a p-value above 0.05, it is likely that component test 4 in the reverse direction along with other component tests is below 0.05. Consequently, the final call for the direction of mediation is reactive, which is opposite to the truth.

2.3 Reactive Model

Suppose G_i is under Bernoulli distribution with probability p, and the true data-generating model is that its effect on X is mediated completely through Y. Thus, this is a reactive

model of G, X and Y, where Y is a mediator :

$$\begin{aligned} G_i &\sim B(1,p) \\ Y_i &= \eta_4 G_i + \epsilon_{3i} \\ X_i &= \eta_6 Y_i + \epsilon_{4i} \end{aligned} \tag{2.11}$$

where the ϵ_{3i} and ϵ_{4i} represent independent random errors, both of which are a Gaussian distribution with mean zero and variance τ_3^2 and τ_4^2 respectively.

The null hypothesis of component test 2, $\beta_4 = 0$ in $X_i = \beta_3 Y_i + \beta_4 G_i + \varepsilon_i$, could not be rejected, as the true β_4 is equal to 0 in the reactive model. Since the final p-value for the CIT test is the maximum of all 4 p-values, under the true reactive model, the type 1 error of CIT will be guarded by component test 2, and is expected to 0.05.

Testing the reactive model with component tests in the reverse direction is equivalent to test the causal model in the forward direction. Thus, as sample size increases, CIT will correctly call the reverse mediation direction $G \to Y \to X$.

2.4 Partial Reactive Model

Suppose G_i is under Bernoulli distribution with probability p, and and suppose its effect on X is partially mediated by Y :

$$\begin{aligned} G_i &\sim B(1,p) \\ Y_i &= \eta_4 G_i + \epsilon_{3i} \\ X_i &= \eta_5 G_i + \eta_6 Y_i + \epsilon_{4i} \end{aligned} \tag{2.12}$$

where the ϵ_{3i} and ϵ_{4i} represent independent random errors, both of which are a Gaussian distribution with mean zero and variance τ_3^2 and τ_4^2 respectively.

2.4.1 Component test 1

As in the set-up of the partial reactive model:

$$Y|G \sim \mathcal{N}\left(\eta_4 G, \tau_3^2\right)$$

For model (2.1), F test is used to test the null hypothesis $\beta_1 = \eta_4 = 0$. At large sample size, this null hypothesis is expected to be rejected as the true value of η_4 is non-zero in the partial reactive model.

2.4.2 Component test 2

The distribution of X condition on (G, Y) is given as:

$$X|G, Y \sim \mathcal{N}\left(\eta_5 G + \eta_6 Y, \tau_4^2\right)$$

Again, the null hypothesis, $\beta_3 = \eta_5 = 0$ for model (2.2) would be rejected as the true value of the parameter is non-zero.

2.4.3 Component test 3

Applying the iteration rule to obtain the expected value and variance of X condition on G, and covariance of X and Y conditional on G,

$$\begin{aligned}
\mathbb{E}[X|G] &= \mathbb{E}\left\{\mathbb{E}\left[X|G, Y\right]|\, G\right\} \\
&= \mathbb{E}\left[\eta_5 G + \eta_6 Y|\, G\right] \\
&= \left(\eta_5 + \eta_4\eta_6\right) G \\
\mathrm{Var}[X|G] &= \mathbb{E}\left\{\mathrm{Var}\left[X|G, Y\right]|\, G\right\} + \mathrm{Var}\left\{\mathbb{E}\left[X|G, Y\right]|\, G\right\} \\
&= \mathbb{E}\left\{\mathrm{Var}\left[X|Y\right]|\, G\right\} + \mathrm{Var}\left\{\mathbb{E}\left[X|Y\right]|\, G\right\} \\
&= \tau_4^2 + \mathrm{Var}\left[\eta_6 X|G\right] \\
&= \tau_4^2 + \eta_6^2\tau_3^2 \\
\mathrm{Cov}\left[X, Y|G_i\right] &= \mathbb{E}\left[XY|G\right] - \mathbb{E}\left[X|G\right]\mathbb{E}\left[Y|G\right]
\end{aligned}$$

$$
\begin{aligned}
&= \mathbb{E}\left\{\mathbb{E}\left[XY|G,X\right]|\,G\right\} - \mathbb{E}\left[X|G\right]\mathbb{E}\left[Y|G\right] \\
&= \mathbb{E}\left\{Y\mathbb{E}\left[X|G,Y\right]|\,G\right\} - \mathbb{E}\left[X|G\right]\mathbb{E}\left[Y|G\right] \\
&= \eta_6\tau_3^2
\end{aligned}
$$

Therefore, condition on G, the pair $\{X, Y\}$ have a bivariate normal distribution:

$$
\left.\begin{pmatrix} X \\ Y \end{pmatrix}\right| G \sim \mathcal{N}\left[\begin{pmatrix} (\eta_5 + \eta_4\eta_6)\,G \\ \eta_4 G \end{pmatrix}, \begin{pmatrix} \tau_4^2 + \eta_6^2\tau_3^2 & \eta_6\tau_3^2 \\ \eta_6\tau_3^2 & \tau_3^2 \end{pmatrix}\right]
$$

Then, the distribution of Y condition on bother G and X:

$$
Y|X,G \sim \mathcal{N}\left(\frac{\eta_4\tau_4^2 - \eta_5\eta_6\tau_3^2}{\tau_4^2 + \eta_6^2\tau_3^2}G + \frac{\eta_6\tau_3^2}{\tau_4^2 + \eta_6^2\tau_3^2}X, \frac{\tau_3^2\tau_4^2}{\tau_4^2 + \eta_6^2\tau_3^2}\right) \tag{2.13}
$$

For model (2.3), F test is used to test the null hypothesis $\beta_4 = \eta_6\tau_3^2/(\tau_4^2 + \eta_6^2\tau_3^2) = 0$. Thus, as long as $\eta_6\tau_3^2 \neq 0$, the null hypothesis is expected to be rejected at large sample size.

2.4.4 Component test 4

To test the hypothesis $\beta_5 \neq 0$ in model (2.3), X^* is simulated to construct an independent model with $\{G, X^*, Y\}$:

$$
\left.\begin{pmatrix} X^* \\ Y \end{pmatrix}\right| G \sim \mathcal{N}\left[\begin{pmatrix} \hat{\eta}_x G \\ \eta_4 G \end{pmatrix}, \begin{pmatrix} \xi^2 & 0 \\ 0 & \tau_3^2 \end{pmatrix}\right]
$$

where $\hat{\eta}_x$ is β estimate in equation $X = \beta G + \epsilon$ and ξ^2 is the variance of $X^*|G$. Through bootstrap, the distribution of F^* is computed comparing two nested models:

$$
\begin{aligned}
Y_i &= \alpha_3^* + \beta_4^* X_i^* + \beta_5^* G_1 + \epsilon_i^* \\
Y_i &= \alpha_3^* + \beta_4^* X_i^* + \epsilon_i^*
\end{aligned}
$$

which should close to a non-central distribution with degree of freedom 1 and $n-3$, and non-central parameter:

$$
\lambda^* = (n-1)\left(\frac{\eta_4^2}{\tau_3^2}\right) SD_G^2\left(1 - R_{G,X^*}^2\right)
$$

where, SD_G^2 is the sample variance of G and R^2_{G,X^*} is the value of R^2 for the OLS regression with response G and regressor X^*.

With the underlying conditional distribution of Y condition on G and X (2.13), the observed F statistic is expect to follow a F distribution with same degree of freedom, but the non-central parameter is

$$\lambda = (n-1)\left(\frac{\left(\eta_4\tau_4^2 - \eta_5\eta_6\tau_3^2\right)^2}{\tau_3^2\tau_4^2\left(\tau_4^2 + \eta_6^2\tau_3^2\right)}\right) SD_G^2\left(1 - R^2_{G,X}\right)$$

where, SD_G^2 is the sample variance of G and $R^2_{G,X}$ is the value of R^2 for the OLS regression with response G and regressor X.

The p-value is the area under the computed F^* distribution smaller than the observed F statistic. It is more likely to result in a significant p-value when λ^* is much larger than λ. One of the special case would be $\lambda = 0$, that is, $\eta_4\tau_4^2 = \eta_5\eta_6\tau_3^2$. Notice, this implies that $\beta_5 = 0$, which is the alternative hypothesis of the component test 4. In this case, the component test 4 could also be rejected. Since the first three component test are expected to be rejected under the set-up of the partial reactive model, the rejection of the fourth component test would lead to a significant CIT p-value. Thus, when $\eta_4\tau_4^2 = \eta_5\eta_6\tau_3^2$, CIT would result a significant p-value in the forward direction, and the final call for the direction of mediation is causal, which is opposite to the truth.

2.5 Independent Model

Suppose G_i is under Bernoulli distribution with probability p, and its effect on X is independent of its effect on Y. Thus, this is an independent model of G, X and Y:

$$\begin{aligned} G_i &\sim B(1,p) \\ X_i &= \eta_7 G_i + \epsilon_{5i} \\ Y_i &= \eta_8 G_i + \epsilon_{6i} \end{aligned} \tag{2.14}$$

where the ϵ_{5i} and ϵ_{6i} represent independent random errors, both of which are a Gaussian distribution with mean zero and variance τ_5^2 and τ_6^2 respectively.

Similar to the reactive model, the null hypothesis of component test 3, $\beta_4 = 0$ in $Y_i = \beta_4 X_i + \beta_5 G_i + \varepsilon_i$, could not be rejected, since the true β_4 is equal to 0 in the true data-generating independent model. Thus, under independent model, the type 1 error of CIT is guarded by component test 3, and is expected to 0.05. When used as a model selection method for the direction of mediation, since the component test 3 is the equivalent in the forward mediation direction and in the reverse mediation direction, both non-significant p-values would let CIT correctly call the final causal direction to be the independent.

2.6 Generalize to Non-normal error model

In previous derivation for the second component test, I assume the normality of the error distribution and use the properties of a bivariate normal distribution. Here I present alternative derivation when the error distribution is an arbitrary distribution that satisfies regularity conditions of a proper OLS estimator. This is derived for the causal mediation model as below. Other model settings can be derived similarly. Suppose G_i is under Bernoulli distribution with probability p, and causes X which then leads Y, that is, a causal model of G (exposure), X (mediator) and Y (outcome):

$$G_i \sim B(1, p)$$

$$X_i = \eta_1 G_i + \epsilon_{1i}$$

$$Y_i = \eta_3 X_i + \epsilon_{2i}$$

where the ϵ_j represent independent random error, which follows an arbitrary distribution with mean zero and variance τ_j^2.

Due to centering and no intercept, the OLS estimator of β_3 in linear regression (2.2) can

be explicitly written as:

$$\begin{aligned}
\hat{\beta}_3 &= \begin{pmatrix} 0 & 1 \end{pmatrix} \left[\begin{pmatrix} \mathbf{Y}' \\ \mathbf{G}' \end{pmatrix} \begin{pmatrix} \mathbf{Y} & \mathbf{G} \end{pmatrix} \right]^{-1} \begin{pmatrix} \mathbf{Y}' \\ \mathbf{G}' \end{pmatrix} \mathbf{X} \\
&= \begin{pmatrix} 0 & 1 \end{pmatrix} \begin{pmatrix} \mathbf{Y}'\mathbf{Y} & \mathbf{G}'\mathbf{Y} \\ \mathbf{Y}'\mathbf{G} & \mathbf{G}'\mathbf{G} \end{pmatrix}^{-1} \begin{pmatrix} \mathbf{Y}'\mathbf{X} \\ \mathbf{G}'\mathbf{X} \end{pmatrix} \\
&= \begin{pmatrix} 0 & 1 \end{pmatrix} \frac{1}{(\mathbf{Y}'\mathbf{Y})(\mathbf{G}'\mathbf{G}) - (\mathbf{Y}'\mathbf{G})(\mathbf{G}'\mathbf{Y})} \begin{pmatrix} \mathbf{G}'\mathbf{G} & -\mathbf{G}'\mathbf{Y} \\ -\mathbf{Y}'\mathbf{G} & \mathbf{Y}'\mathbf{Y} \end{pmatrix} \begin{pmatrix} \mathbf{Y}'\mathbf{X} \\ \mathbf{G}'\mathbf{X} \end{pmatrix} \\
&= \frac{-(\mathbf{Y}'\mathbf{G})(\mathbf{Y}'\mathbf{X}) + (\mathbf{G}'\mathbf{X})(\mathbf{Y}'\mathbf{Y})}{(\mathbf{Y}'\mathbf{Y})(\mathbf{G}'\mathbf{G}) - (\mathbf{Y}'\mathbf{G})(\mathbf{G}'\mathbf{Y})} \\
&= \frac{-(\sum_{i=1}^n G_i Y_i)(\sum_{i=1}^n X_i Y_i) + (\sum_{i=1}^n G_i X_i)(\sum_{i=1}^n Y_i^2)}{(\sum_{i=1}^n Y_i^2)(\sum_{i=1}^n G_i^2) - (\sum_{i=1}^n G_i Y_i)^2} \\
&= \frac{-\bar{G}\bar{Y}\bar{X}\bar{Y} + \bar{G}\bar{X}\bar{Y}^2}{\bar{Y}^2\bar{G}^2 - (\bar{G}\bar{Y})^2}
\end{aligned}$$

Based on the strong law of large numbers,

$$\begin{aligned}
&\overline{GX} \to \mathbb{E}[G_i X_i], \overline{GY} \to \mathbb{E}[G_i Y_i], \overline{XY} \to \mathbb{E}[X_i Y_i], \\
&\qquad \bar{G}^2 \to \mathbb{E}[G_i^2], \bar{Y}^2 \to \mathbb{E}[Y_i^2], \qquad\qquad \text{as } n \to \infty.
\end{aligned}$$

Thus, with the delta-method and under large sample size, $n \to \infty$,

$$\hat{\beta}_3 \to \frac{-\mathbb{E}[G_i Y_i]\mathbb{E}[X_i Y_i] + \mathbb{E}[G_i X_i]\mathbb{E}[Y_i^2]}{\mathbb{E}[Y_i^2]\mathbb{E}[G_i^2] - \left(\mathbb{E}[G_i Y_i]\right)^2}$$

With

$$\begin{aligned}
\mathbb{E}[G_i Y_i] &= \mathbb{E}[\eta_1 \eta_3 G_i^2] = \eta_1 \eta_3 \mathbb{E}[G_i^2] \\
\mathbb{E}[X_i Y_i] &= \mathbb{E}\{\mathbb{E}[X_i Y_i | X_i]\} = \mathbb{E}\{X_i \mathbb{E}[Y_i | X_i]\} = \eta_3 \mathbb{E}[X_i^2] \\
&= \eta_3 \, \mathbb{E}\{\mathbb{E}[X_i^2 | G_i]\} = \eta_3 \mathbb{E}\{\mathrm{Var}\left[X|G\right] + \left[\mathbb{E}\left[X|G\right]\right]^2\} \\
&= \eta_3 \left(\tau_1^2 + \eta_1^2 \mathbb{E}[G_i^2]\right) = \eta_3 \tau_1^2 + \eta_1^2 \eta_3 \mathbb{E}[G_i^2] \\
\mathbb{E}[G_i X_i] &= \mathbb{E}\{\mathbb{E}[G_i X_i | G_i]\} = \mathbb{E}[\eta_1 G_i^2] = \eta_1 \mathbb{E}[G_i^2]
\end{aligned}$$

$$\begin{aligned}\mathbb{E}[Y_i^2] = \ & \mathbb{E}\{\mathbb{E}[Y_i^2|G_i]\} = \mathbb{E}\{\mathrm{Var}\left[Y|G\right] + \left[\mathbb{E}\left[Y|G\right]\right]^2\} \\ & = \mathbb{E}\{\tau_2^2 + \eta_3^2\tau_1^2 + \left[\eta_1\eta_3 G_i\right]^2\} \\ & = \tau_2^2 + \eta_3^2\tau_1^2 + \eta_1^2\eta_3^2\mathbb{E}[G_i^2]\end{aligned}$$

Thus,

$$\begin{aligned}\hat{\beta}_3 \to & \frac{-\eta_1\eta_3\mathbb{E}[G_i^2]\left(\eta_3\tau_1^2 + \eta_1^2\eta_3\mathbb{E}[G_i^2]\right) + \eta_1\mathbb{E}[G_i^2]\left(\tau_2^2 + \eta_3^2\tau_1^2 + \eta_1^2\eta_3^2\mathbb{E}[G_i^2]\right)}{\left(\tau_2^2 + \eta_3^2\tau_1^2 + \eta_1^2\eta_3^2\mathbb{E}[G_i^2]\right)\mathbb{E}[G_i^2] - \left(\eta_1\eta_3\mathbb{E}[G_i^2]\right)^2} \\ = & \frac{\eta_1\tau_2^2}{\eta_3^2\tau_1^2 + \tau_2^2}\end{aligned}$$

This shows that OLS estimator is a consistent estimator, that is, it converges to the true parameter in large samples and $\beta_3 = \eta_1\tau_2^2/(\eta_3^2\tau_1^2 + \tau_2^2)$. This is the same as what I got by using bivariate normal distribution in section 2.1.2. If the null hypothesis is true, F statistic converges to chi-square distribution with 1 degree of freedom with large sample size. Under local alternative hypothesis, F statistic converges to chi-square distribution with 1 degree of freedom and non-centrality parameter

$$\lambda = \left(\frac{\eta_1^2\eta_3^2}{\eta_3^2\tau_1^2 + \tau_2^2}\right)^2 \sigma_2^{-2}\frac{(\mathbf{Y}'\mathbf{Y})(\mathbf{G}'\mathbf{G}) - (\mathbf{Y}'\mathbf{G})(\mathbf{G}'\mathbf{Y})}{\mathbf{Y}'\mathbf{Y}} \tag{2.15}$$

where σ_2^2 is the variance of X conditioned on (G, Y).

Chapter 3
SIMULATION

In this chapter I evaluate the power and type I error for CIT under diverse data-generating models. For each scenario being considered, 1000 simulated datasets were generated. For all scenario, the probability of the dominant allele is set to be 0.2 ($p = 0.2$). The desired power of the test is set to be 0.8 with type I error of 0.05.

3.1 Causal Model

$$G_i \sim B(1, p)$$
$$X_i \sim \mathcal{N}\left(\eta_1 G_i, \tau_1^2\right)$$
$$Y_i \sim \mathcal{N}\left(\eta_3 X_i, \tau_2^2\right)$$

The causal model is generated as above with the normal disturbance assumption. η_1 is denoted as the effect of G on X and τ_1^2 as the variance of X conditioned on G, where the effect of X on Y is η_3 with conditional variance of $Y|X$ equals τ_2^2. There is no direct effect of G on Y - the effect of G on Y is completely mediated by X. Under different sets of parameters, the power of CIT along with the rejection percent of each component test is presented in Figure 3.1. Both component test 1 and 4 are influenced by all the parameters and the pattern is similar. The decrease in effect of G on X (η_1) and the decrease in effect of X on Y (η_3) have similar influence on the rejection percent of test 1 and 4. The increase in τ_2^2 would result in a much lower rejection rate of the two tests, comparing to the same increase in τ_1^2. The rejection percent of component test 3 is least affected by a smaller effect size or a larger variability, specifically, only by parameters related to Y (η_3 and τ_2^2). Similarly, the

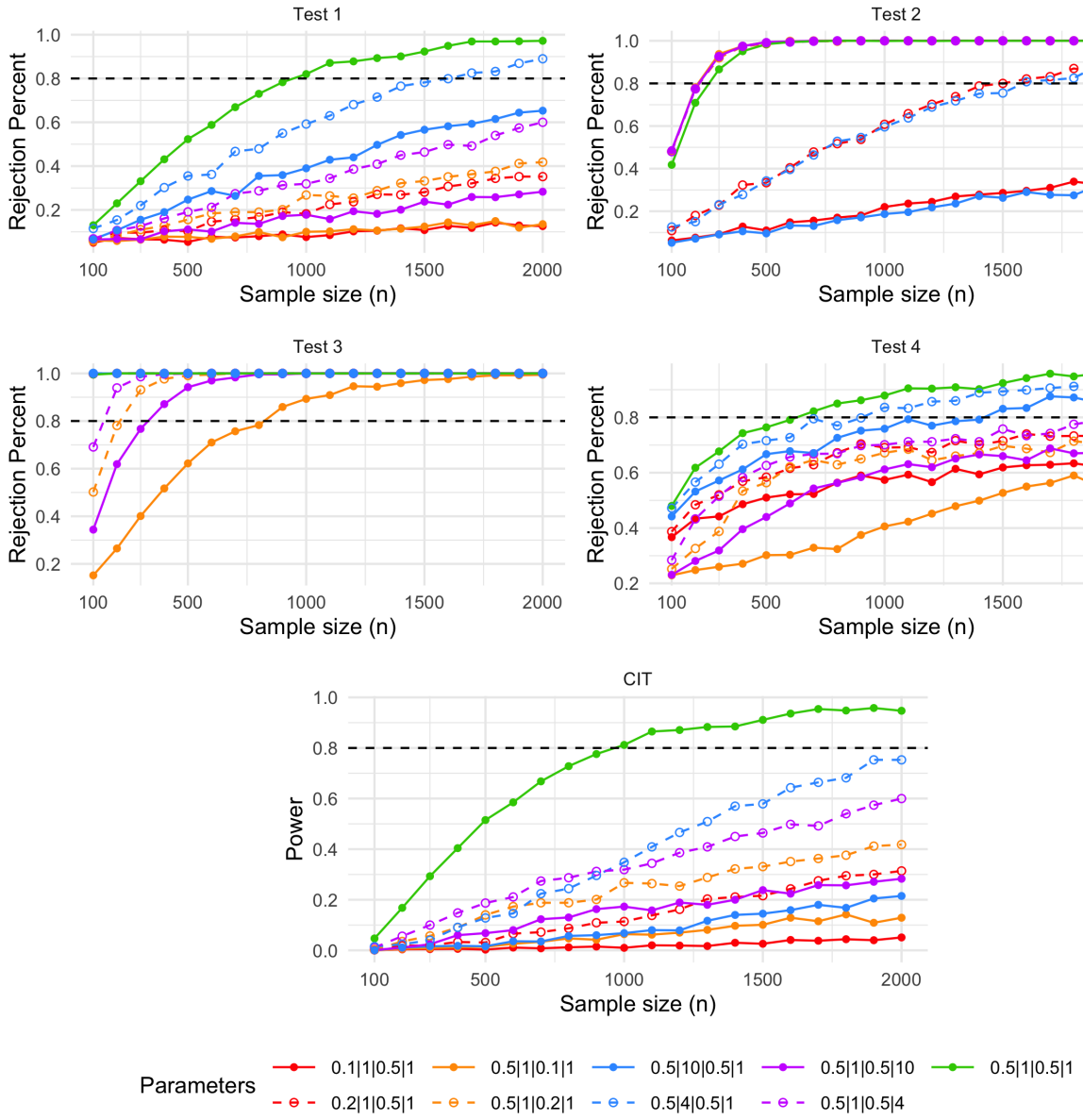


Figure 3.1: Power of CIT and the percentage of rejection for the four component tests under Causal model with different parameter values: $\eta_1|\tau_1^2|\eta_3|\tau_2^2$.

rejection percent of test 2 is only affected by parameters related to X (η_1 and τ_1^2). Unlike test 3 whose rejection percent is above 0.8 for sample size above 800, the rejection percent of

test 2 is lower than 0.4 even with a large sample size when the effect to standard deviation ratio (η_2/τ_1) is small. When $\eta_1 = \eta_3 = 0.5$ and $\tau_1^2 = \tau_2^2 = 1$, sample size needs to reach 1000 for CIT to have a desired power of 0.8. When the effect size (ηs) decreases or the variance (τs) increases, the power of CIT decreases notably. Even with sample size of 2000, none of the other parameter sets reaches sufficient power.

3.2 Reactive Model

$$
\begin{aligned}
G_i &\sim B(1, p) \\
Y_i &\sim \mathcal{N}\left(\eta_4 G_i, \tau_3^2\right) \\
X_i &\sim \mathcal{N}\left(\eta_6 X_i, \tau_4^2\right)
\end{aligned}
$$

The reactive model is generated as above, assuming the error distribution is Gaussian. The effect of G on Y is represented as η_4 with a variance of τ_3^2. Let η_6 denote the effect of Y on X with conditional variance of τ_4^2. In this model, X is the outcome and the effect of G on X is completely mediated through Y.

Under different sets of parameters, the Type 1 error rate of CIT along with the rejection percent of each component test is presented in Figure 3.2. When testing the first null hypothesis, namely Y is independent of G, the rejection percent decreases only when the ratio of effect size of G on Y to its error standard error (η_4/τ_3) is lowered. Based on the definition of the reactive model, it is trivial to reject the first null hypothesis. Thus, even with the extreme low ratio, the rejection percent of component test 1 is well above 0.05 when the true model is reactive model. Similarly, the rejection percent of test 3 is closer to 1 and is only under the influence of the effect size of Y on X (η_6) and its associated error variance (τ_4^2). For component test 4, most of the scenarios also result in a high rejection percent with the exception when effect of Y on X is low. As expected in the Table 1.2, only component test 2 has a low rejection percent, around 0.05, as the null hypothesis that G is independent

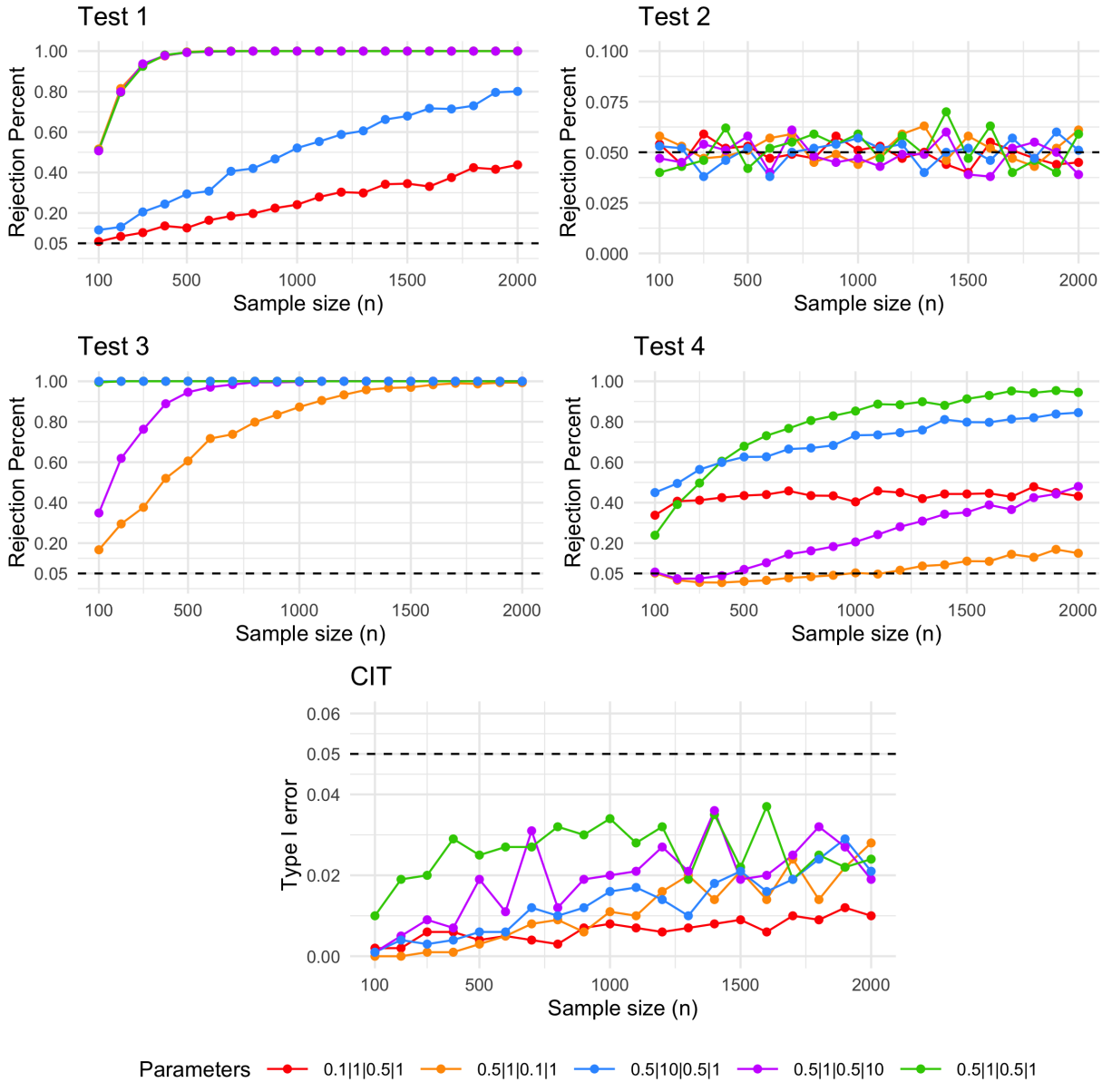


Figure 3.2: Type 1 error of CIT and rejection percent of its component tests under Reactive model with different parameter values: $\eta_4|\tau_3^2|\eta_6|\tau_4^2$

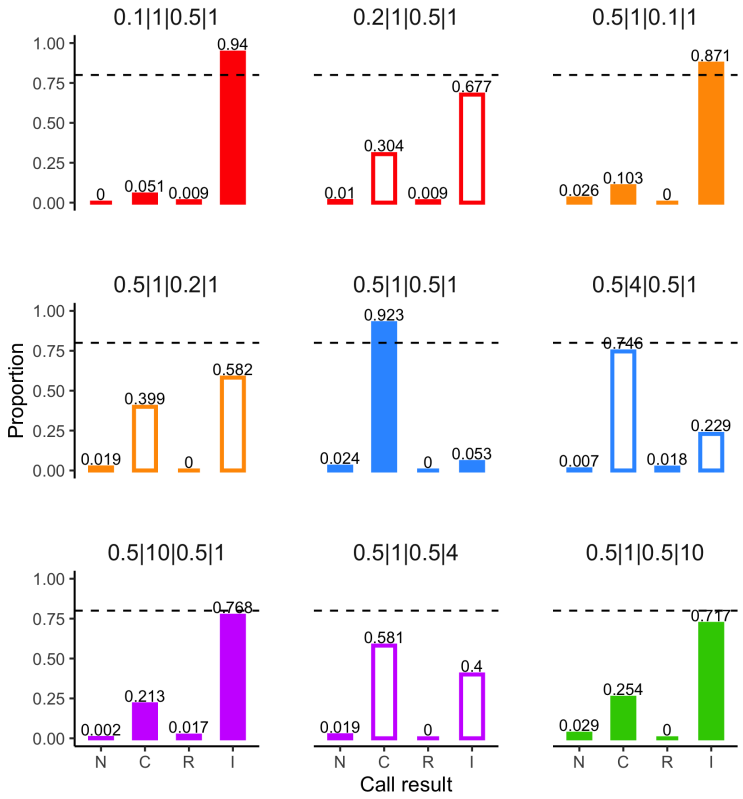
		(partial) causal model	(partial) reactive model
Forward Direction	$G \perp\!\!\!\perp Y$	$E \perp\!\!\!\perp O$	$E \perp\!\!\!\perp M$
	$G \perp\!\!\!\perp X \mid Y$	$E \perp\!\!\!\perp M \mid O$	$E \perp\!\!\!\perp O \mid M$
	$X \perp\!\!\!\perp Y \mid G$	$M \perp\!\!\!\perp O \mid E$	$O \perp\!\!\!\perp M \mid E$
	$G \sim Y \mid X$	$E \sim O \mid M$	$E \sim M \mid O$
Reverse Direction	$G \perp\!\!\!\perp X$	$E \perp\!\!\!\perp M$	$E \perp\!\!\!\perp O$
	$G \perp\!\!\!\perp Y \mid X$	$E \perp\!\!\!\perp O \mid M$	$E \perp\!\!\!\perp M \mid O$
	$X \perp\!\!\!\perp Y \mid G$	$M \perp\!\!\!\perp O \mid E$	$O \perp\!\!\!\perp M \mid E$
	$G \sim X \mid Y$	$E \sim M \mid O$	$E \sim O \mid M$

Table 3.1: Mirror test in the forward and reverse direction for (partial) causal and (partial) reactive model. E denotes for exposure; M denotes for potential intermediate; O denotes for outcome of interest. The null hypotheses are rewritten as its role in the model.

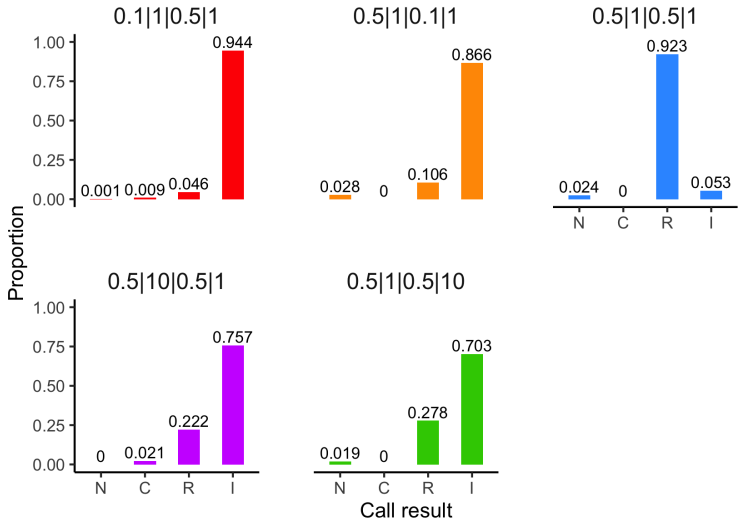
of X conditional on Y is true under reactive model. Since the CIT p-value is the supremum of the p-values from four component tests, test 2 guards the CIT to have a type I error under 0.05 when the underlying true model is reactive model.

3.3 Model selection under Causal model and Reactive model

In the causal model, G is the exposure, X as the potential mediate and Y is the outcome of interest. In the reactive model, the role of X and Y is switched: G still the exposure, Y as potential mediate and X as the outcome of the interest. After rewriting the null hypotheses in forward and reverse direction in term of its role in the models (Table 3.1), causal model tested under causal model is equivalent to reactive model tested under reactive model; causal model tested under reactive model is equivalent to reactive model tested under causal direction . Thus, combining the results from Section 3.1 and 3.2, CIT would incorrectly make an independent model call when the true models are causal or reactive because of low power to have significant p-values in either direction, for example small effects (ηs) or large variability (τs), as shown in Figure 3.3.



(a) Causal model $(\eta_1|\tau_1^2|\eta_3|\tau_2^2)$



(b) Reactive model $(\eta_4|\tau_3^3|\eta_6|\tau_4^2)$

Figure 3.3: Call result of CIT under Causal model and Reactive model with sample size 2000: N for No call; C for causal; R for reactive; I for independent.

3.4 Partial mediation Model

$$G_i \sim B(1, p)$$
$$X_i \sim \mathcal{N}\left(\eta_1 G_i, \tau_1^2\right)$$
$$Y_i \sim \mathcal{N}\left(\eta_2 G_1 + \eta_3 X_i, \tau_2^2\right)$$

The partial mediation model is generated as above with the normal disturbance assumption. η_1 is denoted as the effect of G on X and τ_1^2 as the variance of X conditioned on G, and the effect of X on Y is η_3 with conditional variance of $Y|X$ equals τ_2^2. Unlike complete mediation model, there is also direct effect of G on Y, denoted as η_2. Under different sets of parameters, the power of CIT along with the rejection percent of each component test is presented in Figure 3.4. The marginal effect of G on Y in partial mediation model equals $\eta_2 + \eta_1\eta_3$, which is positively associated with the rejection percent of component test 1. When the marginal effect of G on Y equals zero (as $\eta_1 = 0.5$, $\eta_2 = -0.25$ and $\eta_3 = 0.5$), the rejection percent is close to 0 regardless the sample size. In this scenario, CIT has zero power to detect mediation. Also, the larger the disturbance is, the lower the rejection percent of test 1 is. For component test 3 which tests the association between X and Y condition on G, the rejection percent is compromised only when the ratio between the effect size of X on Y and error standard deviation around Y (η_3/τ_2) is small. Notice that both cases with low rejection in test 1 or 3 also result a low rejection percent of the fourth hypothesis that G is associated with $Y|X$. As shown in the derivation (Section 2.2.2), when

$$\frac{\eta_1}{\eta_2\eta_3} = \frac{\tau_1^2}{\tau_2^2}$$

there is no power to reject the null hypothesis of component test 2 as G is independent of X conditioned on Y.

All the above parameter values that fails to rejection one or more component tests would lead to a false negative result of CIT. These observations confirm the derivation I made in

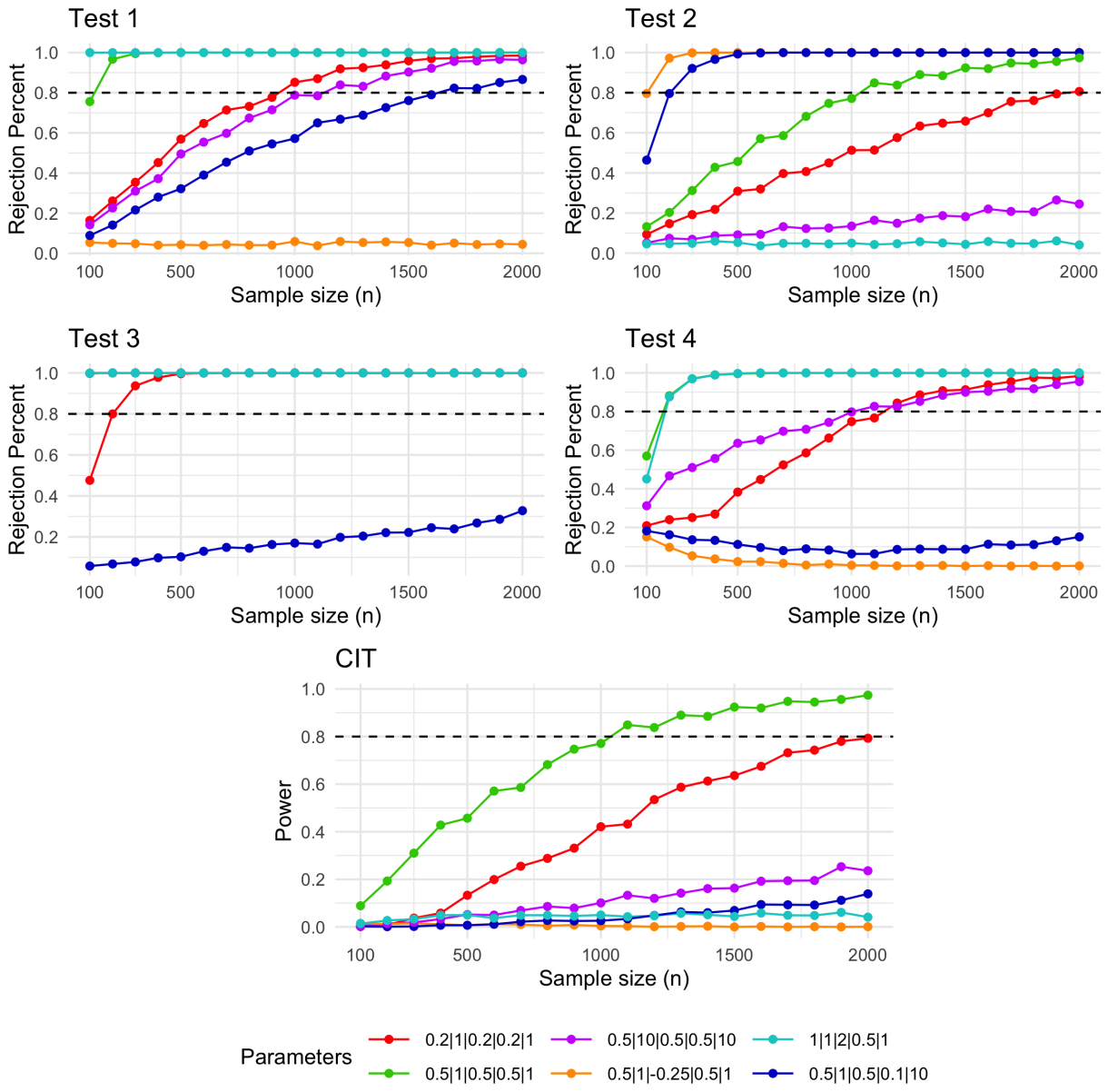


Figure 3.4: Power of CIT and rejection percent of its component tests under partial Mediation model with different parameter values: $\eta_1|\tau_1^2|\eta_2|\eta_3|\tau_2^2$

Chapter 2, that there are scenarios of practically plausible partial mediation, where CIT will have zero power to detect even for large samples. The test is inconsistent under these

numerical coincidences. More importantly, CIT will have low power when these effects do not exactly cancel each other but have opposite sign or subtracting each other.

3.5 Partial Reactive Model

$$
\begin{aligned}
G_i &\sim B(1, p) \\
Y_i &\sim \mathcal{N}\left(\eta_4 G_i, \tau_3^2\right) \\
X_i &\sim \mathcal{N}\left(\eta_5 G_1 + \eta_6 Y_i, \tau_4^2\right)
\end{aligned}
$$

The partial reactive model is generated as above with the normal error assumption. The effect of G on Y is represented as η_4 with a variance of τ_3^2 and η_6 denotes the effect of Y on X with the conditional variance of τ_4^2. There is direct effect of G on X, denoted as η_5.

Under different sets of parameters , the type I error rate of CIT along with the rejection percent of each component test is presented in Figure 3.5. As expected, the first three component tests would have high rejection percent as the partial reactive model meets the first three conditions: 1) G is associated with Y; 2) G is associated with $X|Y$; 3) X is associated with $Y|G$. The rejection percent is negatively associated with ratio of the effect size and error standard error: η_4/τ_3 for test 1, η_5/τ_4 and η_6/τ_4 for test 2 and 3. When the ratio is close to 0, it could result in a low rejection percent.

As shown in the derivation (Section 2.4), when $\eta_4\tau_4^2 = \eta_5\eta_6\tau_3^2$, the rejection percent of component test 4 is above 0.8 regardless of the sample size. Thus, CIT would incorrectly return a false positive result. More importantly, when the effects do not exactly equal to each other, the increasing sample size has the adverse effect of increasing confidence in the wrong direction (Figure 3.5). With a large sample size, CIT would return a false positive rate 100% even although the true model is partial reactive model. This is another serious drawback of CIT that there are plausible scenarios CIT will declare the mediation relationship but the truth is not.

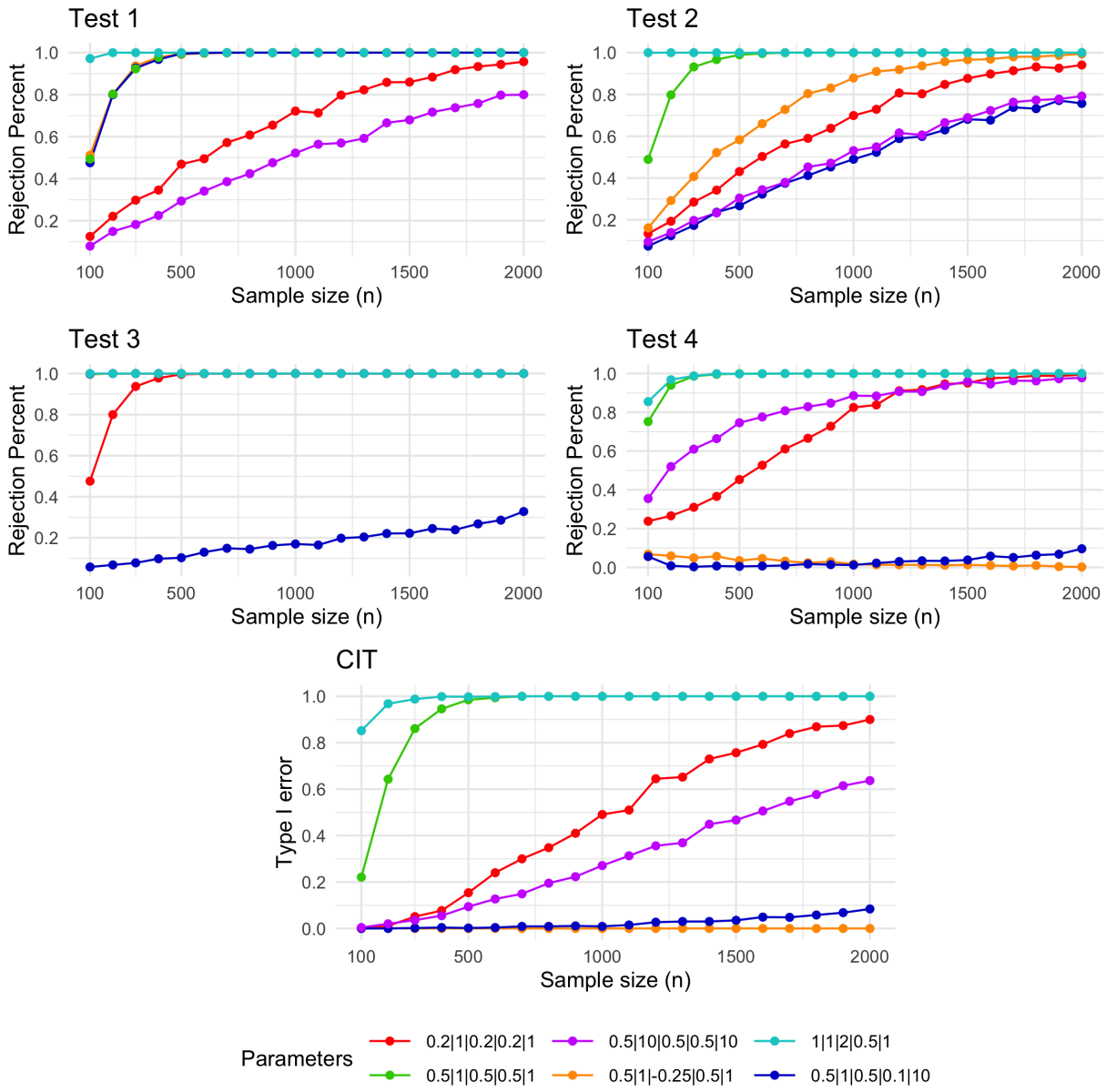
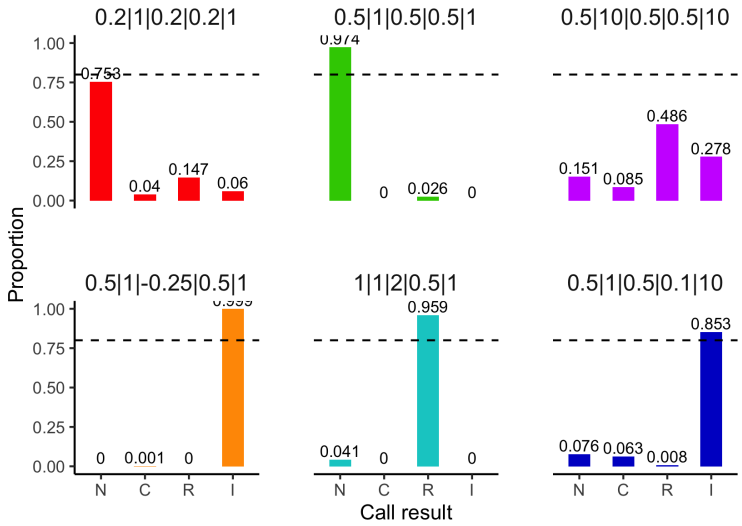


Figure 3.5: Type I error of CIT and rejection percent of its component tests under partial Reactive model with different parameter values: $\eta_4|\tau_3^2|\eta_5|\eta_6|\tau_4^2$

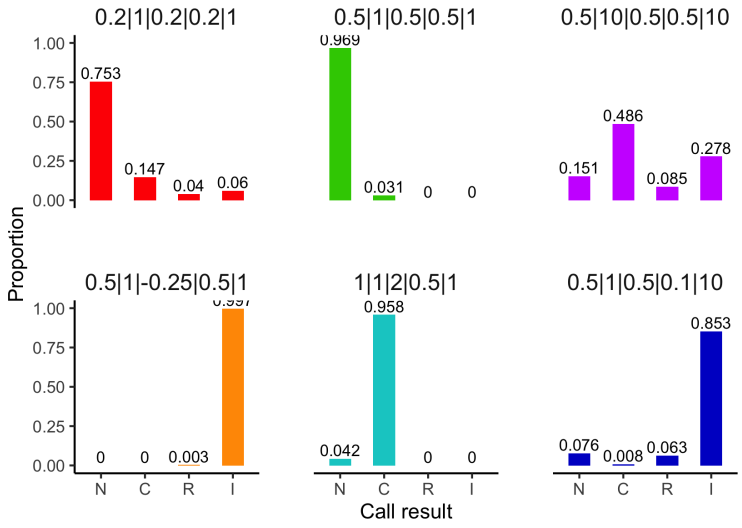
3.6 Model selection under Partial Mediation model and Partial Reactive model

When applying CIT in both forward and reverse directions to obtain a call of causal direction, partial mediation model and partial reactive model have mirror result (Table 3.1): partial mediation model tested under causal direction equivalent to partial reactive model tested under reactive direction; partial mediation model tested under reactive direction equivalent to partial reactive model tested under causal direction. Thus, combining the result from Figure 3.4 and 3.5, one can obtained the call result of CIT under partial mediation model and partial reactive model (Figure 3.6). The most problematic case is that the partial mediation model being misclassified as reactive model or partial reactive model misclassified as causal model. This would happen when $\eta_1/(\eta_2\eta_3) = \tau_1^2/\tau_2^2$ in the partial mediation model or $\eta_4/(\eta_5\eta_6) = \tau_3^2/\tau_4^2$ in the partial reactive model.

Note that the component test 3 is equivalent in both causal and reactive direction, whether X and Y are associated conditioned on G. If the parameter values that fails the component test 3, it would automatically result an independent call as both CIT p-values would be non-significant. Another scenario which would be misclassified as the independent model is when the direct effect of G cancels out the mediated (indirect) effect of G, $\eta_2 = -\eta_1\eta_3$ (or $\eta_5 = -\eta_4\eta_6$). This suggests that marginally G is independent of Y (or X), thus G must be associated with $Y|X$ (or $X|Y$) and with X (or Y) marginally. It further implies that G is associated with $X|Y$ (or $Y|X$). Thus, there is no power to rejection the fourth null hypotheses in both directions and result in an independent call. An interesting result of the model selection is no call, both the forward CIT p-value and reverse CIT p-value are significant: ηs= 0.2(or 0.5) and τ^2s= 1 in Figure 3.6. At first glance, it seems impossible as the component test 4 in one direction is opposite to the component test 2 in the other direction. For example, the fourth null hypothesis in mediation direction is $G \sim Y|X$, and the second null hypothesis in reactive direction is $G \perp\!\!\!\perp Y|X$. By the set-up of the component



(a) Partial Mediation model $(\eta_1|\tau_1^2|\eta_2|\eta_3|\tau_2^2)$



(b) Partial Reactive model $(\eta_4|\tau_3^2|\eta_5|\eta_6|\tau_4^2)$

Figure 3.6: Call result of CIT under Partial Mediation model and Partial Reactive model with sample size 2000: N for No call; C for causal; R for reactive; I for independent.

test 4, the null hypothesis is rejected if the parameter is closer to zero than what would be expected under independent model. Thus, one possible scenario resulting in no call would be that there is a conditional association between G and $Y|X$ to reject the second null hypothesis in the reactive direction, but its effect size is small so that the fourth null hypothesis in the mediation direction is also rejected. In this case, although CIT is able to detect the presence of mediation, it could not distinguish the direction of the mediation which usually is the question the researchers want to answer when applying CIT

3.7 Independent Model

$$
\begin{aligned}
G_i &\sim B(1, p) \\
X_i &\sim \mathcal{N}\left(\eta_7 G_i, \tau_5^2\right) \\
Y_i &\sim \mathcal{N}\left(\eta_8 G_i, \tau_6^2\right)
\end{aligned}
$$

The independent model is generated as above with the normal disturbance assumption. η_7 is denoted as the effect of G on X and τ_5^2 as is error variance, where the effect of G on Y is η_8 with error variance of τ_6^2.

Under different sets of parameters, the type I error rates of CIT along with the rejection percent of each component test are presented in Figure 3.7. As expected the rejection percent of the component test 3 would be around 0.05, since X is independent of Y conditioned on G by the definition of the independent model. It guarantees that the CIT p-value would have a desired type I error when the underlying model is the independent model. Because of the equivalent third test in both forward and reverse directions, CIT p-values will not be significant in either direction and therefore would correctly make an independent call (Figure 3.8).

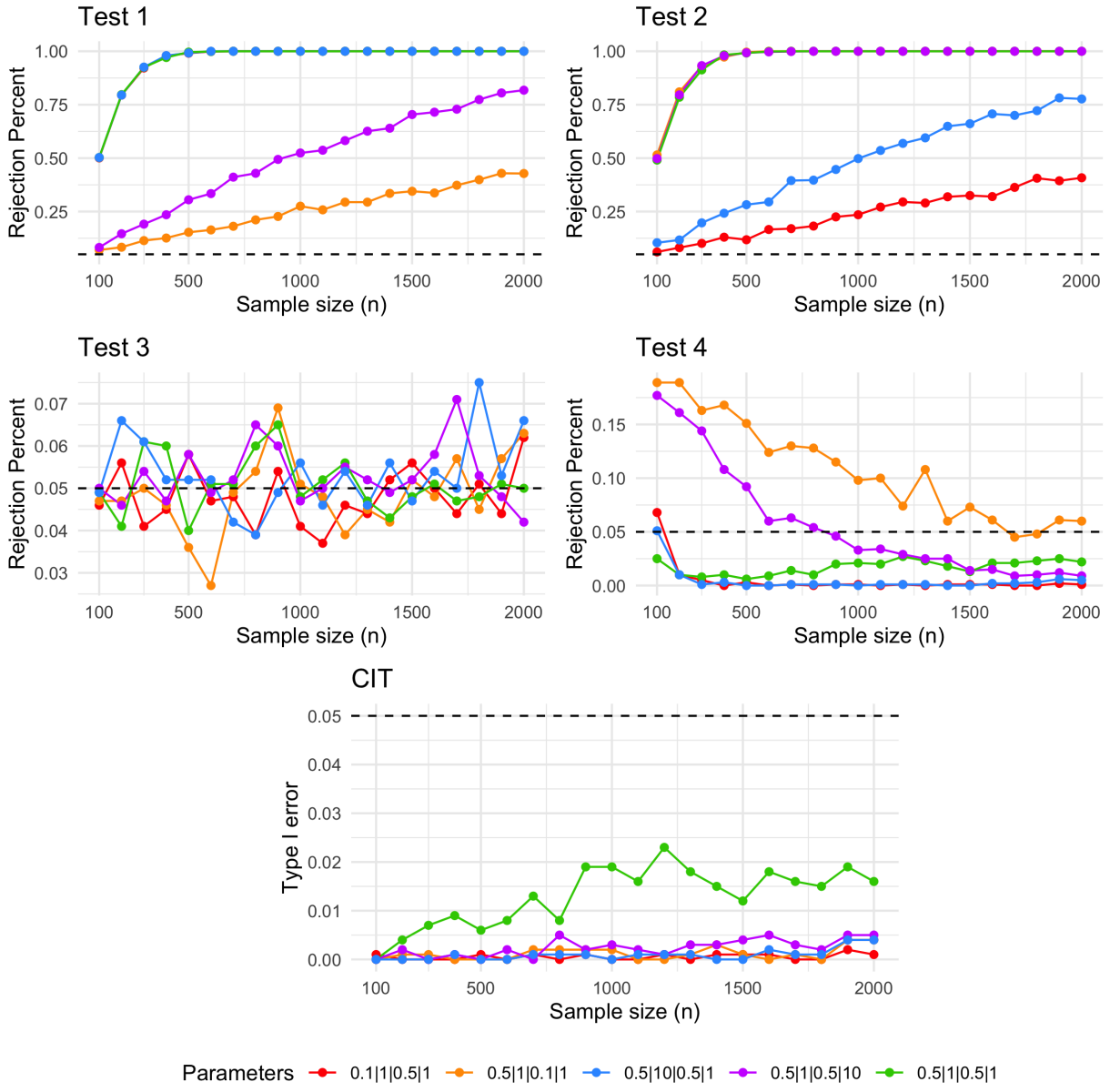


Figure 3.7: Type I error of CIT and rejection percent of four component tests under Independent model with different parameter values: $\eta_7|\tau_5^2|\eta_8|\tau_6^2$.

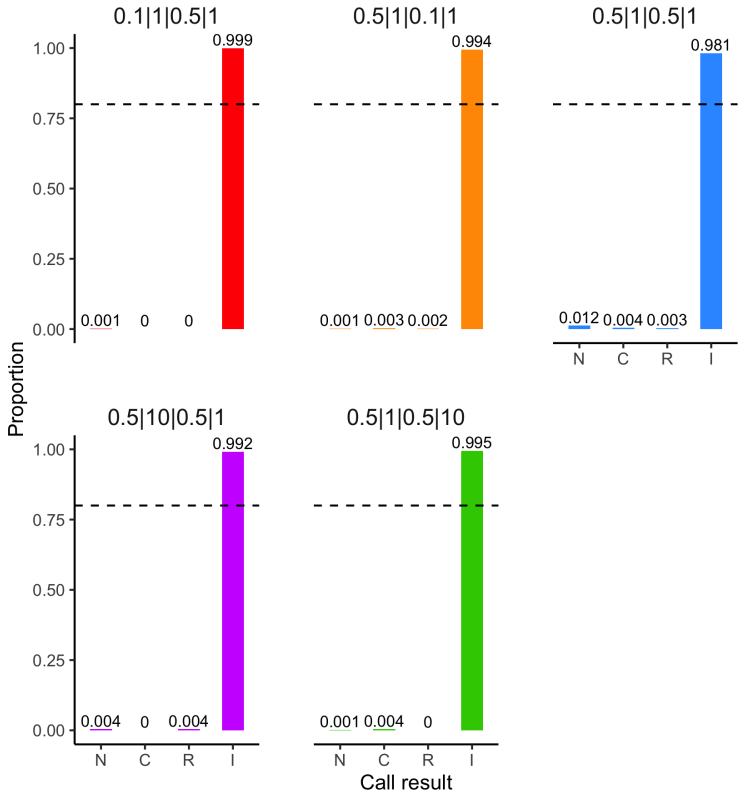


Figure 3.8: Call result of CIT under Independent model ($\eta_7|\tau_5^2|\eta_8|\tau_6^2$) with sample size 2000: N for No call; C for causal; R for reactive; I for independent.

Chapter 4
APPLICATION TO TCGA DATA

In this Chapter, I will apply CIT to TCGA prostate cancer data to test its performance. The Cancer Genome Atlas (TCGA), a collaboration between NCI and the National Human Genome Research Institute, molecularly characterized over 20,000 primary cancer and matched normal samples spanning 33 cancer types[1]. It produces a rich database including genomic, epigenomic, transcriptomic, and proteomic data, that can be studied to better diagnose, treat and prevent cancer. Prostate cancer is the most commonly diagnosed and second most deadly cancer in men in the US, with over 248,000 estimated new cases and over 34,000 deaths in 2021[24]. Studies of twin and worldwide populations have indicated that genetic is responsible for prostate cancer susceptibility[5, 11, 32]. Through GWAS, the largest number of risk variants have been identified in prostate cancer compared to any of the other common cancers[5]. However, most of the prostate cancer risk SNPs are located in the non-coding regions of DNA, leaving their molecular function unknown[2, 5].

rs12653946 located at chromosome 5 has been identified in multiple GWAS studies as a risk locus for prostate cancer susceptibility in multiethnic populations [6, 12, 18, 25]. Its genotype modulates the expression of nearby Iroquois-class homeodomain protein 4 (*IRX4*) gene: an addition of a susceptible allele of rs12653946 is associated with the decreasing expression of *IRX4* [16, 31]. *IRX4* is a member of homeobox gene family which acts as transcription factors. In addition to expressing in developing central nervous system and skin, *IRX4* is predominantly expressed in the cardiac ventricles and an important mediator of ventricular differentiation during cardiac development [4]. *IRX4* has been described as a tumor suppressor in prostate cancer with the interaction of vitamin D receptor [16]. A

recent study identified 12 *IRX4* transcripts and 4 *IRX4* protein isoforms in prostate cancer cells, which induce distinct functional programming that could contribute to prostate cancer hallmarks [6]. The transcriptional regulation of *IRX4* is still not completely clear in prostate caner. The SNP might exert an influence on DNA expression via DNA methylation, a crucial factor in the development and progression of prostate cancer[8, 14]. It has been shown in the blood tissue that DNA methylation and expression may share a common causal variant and mechanism[17]. The goal of this analysis is to identify the role of DNA methylation in the regulation of gene expression in prostate cancer, specifically the role of DNA methylation at each CpG site in the IRX 4 gene region.

The TCGA data I analyze here is collected from 492 primary Prostate Cancer cases. Each observation contains the genotype of SNP rs12653946, the expression level of gene *IRX4*, and methylation level of 19 CpG sites. All TCGA data were downloaded from the TCGA portal. CIT is applied to the data in the forward (SNP→CpG→gene) and the reverse (SNP→gene→CpG) direction for 19 {SNP, CpG, gene} triplets. Hence, the following linear regression models are fitted:

$$\mathbb{E}[\text{gene}] = \alpha_1 + \beta_1 \text{SNP}$$

$$\mathbb{E}[\text{gene}] = \alpha_2 + \beta_2 \text{SNP} + \beta_3 \text{CpG}$$

$$\mathbb{E}[\text{CpG}] = \alpha_3 + \beta_4 \text{SNP}$$

$$\mathbb{E}[\text{CpG}] = \alpha_4 + \beta_5 \text{SNP} + \beta_6 \text{gene}$$

After obtaining the two p-values from the forward and reverse direction, the CIT call result is obtained as described in section 1.2.

Table 4.1 shows that only cg14823763 site has "no call" result with the p-values from both directions are very significant, while the call result for the other 18 CpG sites are independent. The parameter estimate and corresponding p-value for testing $\beta_i = 0$ is shown

CpG	Chromosome (Position)	Target Gene	Regulatory Region	Forward p-value	Reverse p-value	Call Result
cg00264591	5(1883514)	IRX4	TSS1500	0.976	0.946	Independent
cg02238012	5(1880310)	IRX4	Body	1.000	0.994	Independent
cg02641288	5(1883011)	IRX4	TSS200	0.997	0.999	Independent
cg03398865	5(1879554)	IRX4	Body	1.000	1.000	Independent
cg03963198	5(1882871)	IRX4; IRX4	5'UTR; 1stExon	0.843	0.897	Independent
cg04180086	5(1878499)	IRX4	Body	0.794	0.709	Independent
cg07167946	5(1878205)	IRX4	Body	0.428	0.496	Independent
cg07733257	5(1882520)	IRX4	Body	0.775	0.803	Independent
cg07882671	5(1883005)	IRX4	TSS200	0.997	0.980	Independent
cg13974394	5(1882188)	IRX4	Body	0.757	0.870	Independent
cg14823763	5(1884365)	IRX4	TSS1500	1.823×10^{-10}	2.423×10^{-8}	No Call
cg14970046	5(1880021)	IRX4	Body	1.000	1.000	Independent
cg17774559	5(1879698)	IRX4	Body	0.996	0.988	Independent
cg17838420	5(1877906)	IRX4	3'UTR	1.000	0.981	Independent
cg18009496	5(1883940)	IRX4	TSS1500	0.809	0.796	Independent
cg18172877	5(1878672)	IRX4	Body	0.695	0.717	Independent
cg23121993	5(1883954)	IRX4	TSS1500	0.546	0.556	Independent
cg24876960	5(1883214)	IRX4	TSS1500	0.973	0.986	Independent
cg24937747	5(1882902)	IRX4	TSS200	0.759	0.850	Independent

Table 4.1: CIT result for TCGA dataset. Forward direction is SNP→CpG→gene and reverse direction is SNP→gene→CpG. Regulatory region represent gene region feature category describing the CpG position from UCSC: TSS200=0–200 bases upstream of the transcriptional start site (TSS); TSS1500=200–1500 bases upstream of the TSS; 5'UTR=Within the 5' untranslated region, between the TSS and the ATG start site; Body=Between the ATG and stop codon, irrespective of the presence of introns, exons, TSS, or promoters; 3'UTR=Between the stop codon and poly A signal.

CpG	β_1(p-value)	β_2(p-value)	β_3(p-value)	β_4(p-value)	β_5(p-value)	β_6(p-value)
cg00264591		-0.78 (5.71×10^{-37})	-13.4 (0.0581)	-0.000593 (0.0999)	-0.78 (5.71×10^{-37})	-0.000546 (0.0581)
cg02238012		-0.78 (1.38×10^{-37})	0.945 (6.03×10^{-04})	0.00832 (0.367)	-0.78 (1.38×10^{-37})	0.0252 (6.03×10^{-04})
cg02641288		-0.765 (4.14×10^{-37})	-5.79 (1.63×10^{-06})	0.00128 (0.539)	-0.765 (4.14×10^{-37})	-0.00793 (1.63×10^{-06})
cg03398865		-0.754 (1.02×10^{-37})	2.6 (4.39×10^{-12})	-0.00715 (0.281)	-0.754 (1.02×10^{-37})	0.0359 (4.39×10^{-12})
cg03963198		-0.767 (1.63×10^{-36})	-3.61 (9.46×10^{-04})	0.00154 (0.508)	-0.767 (1.63×10^{-36})	-0.00612 (9.46×10^{-04})
cg04180086		-0.75 (2.02×10^{-36})	2.22 (8.87×10^{-09})	-0.0101 (0.121)	-0.75 (2.02×10^{-36})	0.0295 (8.87×10^{-09})
cg07167946		-0.741 (5.77×10^{-36})	1.67 (4.27×10^{-10})	-0.0188 (0.0447)	-0.741 (5.77×10^{-36})	0.0458 (4.27×10^{-10})
cg07733257		-0.775 (2.18×10^{-36})	-0.548 (0.487)	-0.00565 (0.0813)	-0.775 (2.18×10^{-36})	-0.0018 (0.487)
cg07882671		-0.773 (1.16×10^{-36})	-2.95 (0.0246)	-0.000228 (0.906)	-0.773 (1.16×10^{-36})	-0.00349 (0.0246)
cg13974394	-0.772 (2.24×10^{-36})	-0.77 (1.85×10^{-36})	0.736 (0.0124)	-0.00376 (0.663)	-0.77 (1.85×10^{-36})	0.0173 (0.0124)
cg14823763		-0.531 (1.51×10^{-15})	-1.95 (1.9×10^{-11})	0.124 (1.36×10^{-39})	-0.531 (1.51×10^{-15})	-0.0452 (1.9×10^{-11})
cg14970046		-0.768 (3.09×10^{-38})	1.53 (1.17×10^{-09})	-0.00295 (0.767)	-0.768 (3.09×10^{-38})	0.0477 (1.17×10^{-09})
cg17774559		-0.754 (3.24×10^{-37})	3.2 (3.84×10^{-10})	-0.00562 (0.252)	-0.754 (3.24×10^{-37})	0.0241 (3.84×10^{-10})
cg17838420		-0.757 (2.99×10^{-37})	2.4 (2.48×10^{-09})	-0.00624 (0.317)	-0.757 (2.99×10^{-37})	0.0293 (2.48×10^{-09})
cg18009496		-0.77 (1.88×10^{-36})	-3.13 (0.0249)	0.000627 (0.731)	-0.77 (1.88×10^{-36})	-0.00327 (0.0249)
cg18172877		-0.752 (2.42×10^{-36})	2.02 (6.72×10^{-08})	-0.0102 (0.132)	-0.752 (2.42×10^{-36})	0.0287 (6.72×10^{-08})
cg23121993		-0.765 (3.05×10^{-36})	-2.6 (0.00184)	0.00297 (0.331)	-0.765 (3.05×10^{-36})	-0.00758 (0.00184)
cg24876960		-0.771 (9.15×10^{-37})	-4.85 (0.00322)	0.00018 (0.907)	-0.771 (9.15×10^{-37})	-0.00363 (0.00322)
cg24937747		-0.763 (1.67×10^{-36})	-2.5 (6.64×10^{-05})	0.0037 (0.36)	-0.763 (1.67×10^{-36})	-0.0128 (6.64×10^{-05})

Table 4.2: β estimate with p-value from linear regression models for 19 triplets {SNP, CpG, gene}

in Table 4.2. The association between SNP and DNA methylation presents at cg14823763 with a very significant p-value, while the β_4 estimate at other CpG locations are close to 0 and most are non-significant. With all significantly non-zero βs, there may be a partial mediation relationship among the triplets {SNP, cg14823763, gene}. However, the direction of the mediation is unclear based on the CIT result: whether methylation level of CpG site mediates the genetic effect on gene expression or methylation is cased by DNA expression can not be determined, same as the simulation result in section 3.6

The partial mediation relation among SNP-CpG-gene expression is expected to be common with the intricate metabolic pathway of DNA methylation in prostate cancer involving a large group of enzymes and metabolites [14]. There is also evidence for both mediation effects of DNA methylation on gene expression as well as the influence of gene expression on methylation[9, 26]. Thus, one should be cautious when using CIT to study the role of DNA methylation in gene expression regulation. When the true relationship is partial mediation or partial reactive, CIT may not be able to answer the question of interest when applied in both directions, as shown in this data application. More importantly, CIT may return a false positive result when only applied in the forward direction.

Chapter 5
CONCLUSION

In this thesis, I studied the performance of the CIT under five models of tri-variate relationship: causal mediation, partial mediation, reactive, partial reactive, and independent models, as a test of the forward mediation relationship and as a procedure to call the direction of the mediation. When the true model is reactive or independent model, both theoretical derivation and simulation shows that CIT will achieve the correct test inference as the type I error of CIT is close to 0.05 since it is guarded by component test 2 or 3. With a large sample size, CIT would also correctly call the forward mediation direction when the underlying model is causal mediation. The threat to the validity of CIT comes from partial mediation or partial reactive models. I showed that there are two scenarios of partial mediation models that CIT has no power to detect the mediation even for large sample size: 1) the direct effect and indirect effect of the exposure on outcome are in the opposite direction and canceled out in the overall genetic effect; 2) parameter values satisfy $\eta_1/(\eta_2\eta_3) = \tau_1^2/\tau_2^2$ so that CIT will make a reactive call when applied in both forward and reverse directions. I also showed a scenario where CIT will declare the mediation relationship but the truth is not: when $\eta_4/(\eta_5\eta_6) = \tau_3^2/\tau_4^2$ under partial reverse model.

As a procedure to call the direction of mediation, CIT may not reliably select the direction of the mediation if both forward and reverse p-values are significant: it is plausible biologically that the underlying model is a partial mediation or partial reactive model as shown in data application. Our simulation results confirm the inconsistency of CIT call in these scenarios.

Only considered the simplest models without involving hidden confounding variables

or multiple mediators, my study shows that CIT is inconsistent under certain numerical coincidence and may not reliably distinguish the direction of the mediation. It will be crucial to improve CIT so that it would detect the direction of the mediation correctly even under partial mediation or partial reactive models. Future work might also include investigation of the performance of CIT in more comprehensive settings, considering more variables and possibly measurement error of the mediator.

BIBLIOGRAPHY

[1] The cancer genome atlas program.

[2] S. Benafif and R. Eeles. Genetic predisposition to prostate cancer. *British Medical Bulletin*, 120(1):75–89, 12 2016.

[3] Yohan Bossé and Christopher I Amos. A decade of gwas results in lung cancer. *Cancer Epidemiol Biomarkers Prev*, 27(4):363–379, 2018.

[4] Benoit G. Bruneau, Zheng-Zheng Bao, Makoto Tanaka, Jean-Jacques Schott, Seigo Izumo, Constance L. Cepko, J.G. Seidman, and Christine E. Seidman. Cardiac expression of the ventricle-specific homeobox gene irx4 is modulated by nkx2-5 and dhand. *Developmental Biology*, 217(2):266–277, 2000.

[5] Rosalind Eeles, Chee Goh, Elena Castro, Elizabeth Bancroft, Michelle Guy, Ali Amin Al Olama, Douglas Easton, and Zsofia Kote-Jarai. The genetic epidemiology of prostate cancer and its clinical implications. *Nature reviews. Urology*, 11(1):18–31, 2014.

[6] Achala Fernando, Chamikara Liyanage, Afshin Moradi, Panchadsaram Janaththani, and Jyotsna Batra. Identification and characterization of alternatively spliced transcript isoforms of irx4 in prostate cancer. *Genes*, 12(5), 2021.

[7] Luis M Franco, Kristine L Bucasas, Janet M Wells, Diane Niño, Xueqing Wang, Gladys E Zapata, Nancy Arden, Alexander Renwick, Peng Yu, John M Quarles, Molly S Bray, Robert B Couch, John W Belmont, and Chad A Shaw. Integrative genomic analysis of the human immune response to influenza vaccination. *eLife*, 2:e00299, 2013.

[8] Wolfgang Goering, Michael Kloth, and Wolfgang A. Schulz. *DNA Methylation Changes in Prostate Cancer*, pages 47–66. Humana Press, Totowa, NJ, 2012.

[9] Maria Gutierrez-Arcelus, Tuuli Lappalainen, Stephen B Montgomery, Alfonso Buil, Halit Ongen, Alisa Yurovsky, Julien Bryois, Thomas Giger, Luciana Romano, Alexandra Planchon, Emilie Falconnet, Deborah Bielser, Maryline Gagnebin, Ismael Padioleau, Christelle Borel, Audrey Letourneau, Periklis Makrythanasis, Michel Guipponi, Corinne Gehrig, Stylianos E Antonarakis, and Emmanouil T Dermitzakis. Passive and active dna methylation and the interplay with genetic variation in gene regulation. *eLife*, 2:e00523, jun 2013.

[10] Ryan Hunt, Zuben E Sauna, Suresh V Ambudkar, Michael M Gottesman, and Chava Kimchi-Sarfaty. Silent (synonymous) snps: Should we care about them? 578:23–39, 2009.

[11] Paul Lichtenstein, Niels V. Holm, Pia K. Verkasalo, Anastasia Iliadou, Jaakko Kaprio, Markku Koskenvuo, Eero Pukkala, Axel Skytthe, and Kari Hemminki. Environmental and heritable factors in the causation of cancer — analyses of cohorts of twins from sweden, denmark, and finland. *New England Journal of Medicine*, 343(2):78–85, 2000. PMID: 10891514.

[12] Sara Lindström, Fredrick R. Schumacher, Daniele Campa, Demetrius Albanes, Gerald Andriole, Sonja I. Berndt, H. Bas Bueno-de Mesquita, Stephen J. Chanock, W. Ryan Diver, J. Michael Ganziano, Susan M. Gapstur, Edward Giovannucci, Christopher A. Haiman, Brian Henderson, David J. Hunter, Mattias Johansson, Laurence N. Kolonel, Loic Le Marchand, Jing Ma, Meir Stampfer, Victoria L. Stevens, Dimitrios Trichopoulos, Jarmo Virtamo, Walter C. Willett, Meredith Yeager, Ann W. Hsing, and Peter Kraft. Replication of five prostate cancer loci identified in an asian population—results from the nci breast and prostate cancer cohort consortium (bpc3). *Cancer Epidemiology and Prevention Biomarkers*, 21(1):212–216, 2012.

[13] Yun Liu, Martin J Aryee, Leonid Padyukov, M Daniele Fallin, Espen Hesselberg, Arni Runarsson, Lovisa Reinius, Nathalie Acevedo, Margaret Taub, Marcus Ronninger, Klementy Shchetynsky, Annika Scheynius, Juha Kere, Lars Alfredsson, Lars Klareskog, Tomas J Ekström, and Andrew P Feinberg. Epigenome-wide association data implicate dna methylation as an intermediary of genetic risk in rheumatoid arthritis. *Nature biotechnology*, 31(2):142–147, 2013.

[14] Charles E. Massie, Ian G. Mills, and Andy G. Lynch. The importance of dna methylation in prostate cancer development. *The Journal of Steroid Biochemistry and Molecular Biology*, 166:1–15, 2017. Special Issue on Prostate Cancer.

[15] Joshua Millstein, Bin Zhang, Jun Zhu, and Eric E Schadt. Disentangling molecular relationships with a causal inference test. *BMC Genetics*, 10(1):23–37, 2009.

[16] Hai Ha Nguyen, Ryo Takata, Tomoaki Fujioka, Yusuke Nakamura, Hidewaki Nakagawa, Shusuke Akamatsu, Daichi Shigemizu, Tatsuhiko Tsunoda, Mutsuo Furihata, Atsushi Takahashi, Michiaki Kubo, Naoyuki Kamatani, and Osamu Ogawa. Irx4 at 5p15 suppresses prostate cancer growth through the interaction with vitamin d receptor, conferring prostate cancer susceptibility. *Human molecular genetics*, 21(9):2076–2085, 2012.

[17] Brandon L Pierce, Lin Tong, Maria Argos, Kathryn Demanelis, Farzana Jasmine, Muhammad Rakibuz-Zaman, Golam Sarwar, Md Tariqul Islam, Hasan Shahriar, Tariqul

Islam, Mahfuzar Rahman, Md Yunus, Muhammad G Kibriya, Lin S Chen, and Habibul Ahsan. Co-occurring expression and methylation qtls allow detection of common causal variants and shared biological mechanisms. *Nature communications*, 9(1):804–804, 2018.

[18] Jun Qi, Lu Tian, Zhuo Chen, Li Wang, Sha Tao, Xin Gu, Rong Na, Yang Jiao, Jian Kang, Siqun Zheng, Jianfeng Xu, and Jielin Sun. Genetic variants in 2q31 and 5p15 are associated with aggressive benign prostatic hyperplasia in a chinese population. *The Prostate*, 73(11):1182–1190, 2013.

[19] David C Rees, Thomas N Williams, and Mark T Gladwin. Sickle-cell disease. *The Lancet*, 376(9757):2018–2031, 2010.

[20] Alvin C Rencher and G. Bruce Schaalje. *Linear Models in Statistics*. John Wiley & Sons, Incorporated, Newy York, 2008.

[21] Francis Robert and Jerry Pelletier. Exploring the impact of single-nucleotide polymorphisms on translation. *Frontiers in genetics*, 9:507–507, 2018.

[22] Stephen W. Scherer and Peter M. Visscher. *Genome-Wide Association Studies: From Polymorphism to Personalized Medicine*. Cambridge University Press, 2016.

[23] Barkur S Shastry. Snps: Impact on gene function and phenotype. *Single Nucleotide Polymorphisms*, 578:3–22, 2009.

[24] Rebecca L. Siegel, Kimberly D. Miller, Hannah E. Fuchs, and Ahmedin Jemal. Cancer statistics, 2021. *CA: A Cancer Journal for Clinicians*, 71(1):7–33, 2021.

[25] Ryo Takata, Shusuke Akamatsu, Michiaki Kubo, Atsushi Takahashi, Naoya Hosono, Takahisa Kawaguchi, Tatsuhiko Tsunoda, Johji Inazawa, Naoyuki Kamatani, Osamu Ogawa, Tomoaki Fujioka, Yusuke Nakamura, and Hidewaki Nakagawa. Genome-wide association study identifies five new susceptibility loci for prostate cancer in the japanese population. *Nature genetics*, 42(9):751–754, 2010.

[26] Sergio Villicaña and Jordana T. Bell. Genetic impacts on dna methylation: research findings and future perspectives. *Genome Biology*, 22(1):127, 2021.

[27] Peter M. Visscher, Naomi R. Wray, Qian Zhang, Pamela Sklar, Mark I. McCarthy, Matthew A. Brown, and Jian Yang. 10 years of gwas discovery: Biology, function, and translation. *The American Journal of Human Genetics*, 101(1):5 – 22, 2017.

[28] Petr Volkov, Anders H Olsson, Linn Gillberg, Sine W Jørgensen, Charlotte Brøns, Karl-Fredrik Eriksson, Leif Groop, Per-Anders Jansson, Emma Nilsson, Tina Rönn, Allan Vaag, and Charlotte Ling. A genome-wide mqtl analysis in human adipose tissue identifies genetic variants associated with dna methylation, gene expression and metabolic traits. *PloS one*, 11(6):e0157776, 2016.

[29] Sanford Weisberg. *Applied linear regression*. Wiley series in probability and statistics. Wiley, Hoboken, New Jersey, 4th ed. edition, 2014.

[30] Stefan Wellek. *Testing Statistical Hypotheses of Equivalence and Noninferiority*. CRC Press, 2010.

[31] Xing Xu, Wasay M Hussain, Joseph Vijai, Kenneth Offit, Mark A Rubin, Francesca Demichelis, and Robert J Klein. Variants at irx4 as prostate cancer expression quantitative trait loci. *European Journal of Human Genetics*, 22(4):558–563, 2014.

[32] Charnita M Zeigler-Johnson, Hanna Rennert, R. Devi Mittal, Mohamed Jalloh, Rajeev Sachdeva, S. Bruce Malkowicz, Anil Mandhani, B Mittal, Serigne M Gueye, and Timothy R Rebbeck. Evaluation of prostate cancer characteristics in four populations worldwide. *Canadian journal of urology*, 15(3):4056–4064, 2008.