Causal Mediation Analysis with Controlled Indirect Effect

Abstract

Causal mediation analysis, based on several assumptions, allows one to use observational data to estimate the importance of mediating pathways of exposure on outcome. However, current approaches to mediation analysis with multiple mediators either involve assumptions not verifiable by experiments, or joint estimation of multiple mediators precluding the practical design of experiments due to curse of dimensionality, or the difficulty in interpretation when causal dependencies among mediators are present. Akin to the well-established controlled direct effect, we propose controlled indirect effect: mediation analysis for multiple manipulable mediators with causal dependencies. The proposed method is practically relevant because the decomposition of the total effect does not require the cross-world assumptions and focuses on the marginal effects after manipulating one single mediator. We illustrate the approach using simulated data and the framing dataset from political science. Our results provide guidance for ranking mediators to decide which mediator to manipulate to maximally improve the outcome.

1 INTRODUCTION

Inferring causal effects and the mediating pathways from observational and/or experimental data is one of the most important problems in several domains including healthcare and artificial intelligence [Glass et al., 2013]. In animal and some human studies, it is possible to conduct a randomized controlled trial (RCT) to infer the causal effect of a particular intervention on an outcome. RCTs are considered the gold standard of causal inference given their ability to limit/reduce multiple sources of bias [Deaton and Cartwright, 2018]. However, an RCT may not be feasible or ethical for certain interventions. Even when RCT is possible, the mediators are not fully randomized. In these cases, researchers must conduct observational studies instead and adjust for potential biases using statistical methods. Advances in causal inference [Pearl, 2009, Hernán and Robins, 2020, Pingault et al., 2018, Yao et al., 2020] have led to the possibility of studying causal effects in a principled way using observational data to guide decision making. Note that throughout this paper we will use the word "exposure" instead of "intervention" or "treatment". "Exposure" is more general which includes intervention or treatment, or observational factors such as a disease or risk factor. We use the terminology from the potential outcomes framework developed by Neyman, Rubin, and Robins [Neyman, 1923, Rubin, 1978, Robins, 1986]: when the assignment is equal to the observed exposure, the outcome is called "factual outcome"; otherwise it is called "counterfactual outcome"; either of which is called "potential outcome".

Mediation analysis is an important sub-field of causal inference. It aims at measuring the relative importance of each mediating pathway, by decomposing the total effect (TE) into parts including mediation due to mediators, and interactions due to the co-existence of exposure and mediators [VanderWeele, 2015, Imai et al., 2010]. The mediators are defined as those causally affected by the exposure, while also causally affecting the outcome. The categorization of a variable as being a mediator or a confounder is determined by human knowledge or temporal ordering, if any. The total effect can be decomposed in various ways including

(1) controlled direct effect and eliminated effect;

(2) natural direct effect and natural indirect effect [Pearl, 2014]; and

(3) 4-way decomposition: controlled direct effect, reference interaction, mediated interaction, and pure indirect effect [VanderWeele, 2015].

The extension of these approaches into multiple mediators with causal dependency is challenging: For decomposition

(1), the eliminated effect represents all effects other than the controlled direct effect [Robins and Richardson, 2010] which is unclear how each mediator contributes to it. For decomposition (2), although the division into natural direct and indirect effects is simple and can be done even in the presence of interaction, this decomposition involves cross-world effects (nested counterfactuals with different exposures), which does not correspond to randomized experiment performed via realistic interventions on the exposure and/or mediator [Robins and Richardson, 2010]; the identification of these effects also requires a strong "sequential ignorability" assumption which rules out the possibility of assessing each mediator when they are causally dependent [Tingley et al., 2014]. However, Robins et al. [2020] pointed out an alternative interpretation without reference to the cross-world effects where treatment is decomposed into multiple components. For decomposition (3), the pure indirect effect in the case of multiple mediators requires estimating the joint potential outcome, i.e. the potential outcome when the exposure and all mediators would have been set to particular values, which is not practical for experiment design and suffers from the curse of dimensionality when the number of mediators is large.

In this work, we propose a "controlled indirect effect" (CIE) approach to decompose the total effect for multiple manipulable mediators with causal dependencies, which overcomes the above limitations. Note that we limit the scope to binary exposure and mediator. This approach has several advantages including: (1) The decomposition involves terms related to the effect when one single mediator is Intervened. This can mimic clinical practice where a physician might focus on treating one comorbidity at one time, rather than treating all comorbidity jointly. Intervening on a single mediator makes it possible to rank mediators (such as comorbidities) based on their CIE (informally, the change in the outcome if everyone's k-th comorbidity is treated). So that we can make decision to give priority to the top mediators to spend resources (doctor's attention, medication, research). Since CIE can be viewed as the total effect of the mediator on outcome, its effect includes all downstream mediators, which naturally addresses multiple mediators with causal dependency; and (2) there is no cross-world effects in the decomposition, hence CIE can be experimentally validated, such as using the parallel (encouragement) experiment design [Imai and Yamamoto, 2013].

2 RELATED WORK

The existing works mostly focus on the extension of decomposition (2), i.e. the natural direct and indirect effect approach. VanderWeele and Vansteelandt [2014] and VanderWeele et al. [2014] extend it to multiple mediators by considering all mediators jointly as one vector-valued mediator, so that the "sequential ignorability" assumption (no exposure-induced mediator-outcome confounder) still holds. Daniel et al. [2015] still estimates the indirect effect of each mediator (although the "sequential ignorability" assumption is violated), but uses sensitivity analysis to assess the robustness of their results to the violation of the assumption. As we mentioned above, this approach requires cross-world assumption. There are also works focusing on the extension of decomposition (3). Bellavia and Valeri [2018] extends the 4-way decomposition to the finest decomposition that unifies multiple mediators and interactions for causally independent mediators. With more mediators, it becomes incrementally difficult to define, identify, and estimate these components.

Our approach is closer to the interventional effect approach in VanderWeele et al. [2014], Vansteelandt and Daniel [2017], Loh et al. [2019]. The interventional indirect effect is defined as the contrast in the outcome if we fix the exposure, while changing the mediator from a sampled value from the distribution of the mediator among all subjects with one exposure to a sampled value from the distribution from another exposure. However, the sum of interventional direct and indirect effect is not equal to the total effect. In contrast, in our approach, the controlled direct effect and the scaled controlled indirect effect add up to the total effect.

3 METHODS

We use Y to denote the outcome, e.g. mortality, cognitive test score, or a physiological measurement. A denotes the exposure (e.g. taking a pill, infection with HIV or coronavirus, or developing a disease such as Alzheimer's). M_k denotes the k-th mediator, e.g. a co-morbid medical condition which worsens the outcome. L denotes the set of covariates, e.g. a patient's age, gender, race, smoking status, and years of education. Here we limit the scope to binary A and M; Y is discrete or continuous; and L is a vector of any type of variable. There are K mediators.

3.1 TOTAL EFFECT DECOMPOSITION

In general, given a causal DAG, for the k-th mediator, the total effect (TE) can be decomposed into controlled direct effect (CDE), and scaled controlled indirect effect (sCIE) which is a function of CIE (proof in Appendix A)

$$\mathrm{TE} = \mathrm{CDE}_k(0) + \mathrm{sCIE}_k \text{ for } k = 1,\dots,K\ , \tag{1}$$

where

$$\mathrm{CDE}_k(0) = Y_k(1,0) - Y_k(0,0)\ ; \tag{2}$$

$$\mathrm{sCIE}_k = M_k(1)\mathrm{CIE}_k(1) - M_k(0)\mathrm{CIE}_k(0)\ ; \tag{3}$$

$$\mathrm{CIE}_k(a) = Y_k(a,1) - Y_k(a,0)\ ; \tag{4}$$

$$Y_k(a,m) = Y\left(a, M_1(a), \cdots, M_{k-1}(a), m, M_{k+1}(a), \cdots, M_K(a)\right)\ ; \tag{5}$$

$$M_k(a) = M_k\left(a, Pa\{M_k\}(a)\right)\ . \tag{6}$$

Here we denote $Y(A=a,M_k=m)$, or simply $Y_k(a,m)$, as the potential outcome of Y when A would have been a, the k-th mediator would have been m, and the other mediators were behaving as if A was a. $\mathrm{CDE}_k(0)$ is the controlled direct effect for the k-th mediator, defined as the contrast in the potential outcome when the exposure changes from 0 to 1, while fixing the k-th mediator to be 0; other mediators were behaving as if A was a. sCIE_k is the scaled controlled indirect effect for the k-th mediator, defined as the controlled indirect effect scaled by the potential outcome of the k-th mediator when fixing the exposure to 1, subtracting the same quantity but when fixing the exposure to 0. $\mathrm{CIE}_k(a)$ is the controlled indirect effect of the k-th mediator, defined as the contrast in the potential outcome when the k-th mediator changes from 0 to 1, while fixing the exposure to a and other mediators were behaving as if A was a. $Pa\{M_k\}(a) = \{M_j(a)\}_{j\in\text{Parent of } M_k}$ which is the set of causal parents of the k-th mediator in the given DAG.

Note that there is no cross-world potential outcome such as $M_k(1,Pa\{M_k\}(0))$. Also note that Equation (6) is a recursive definition: if there is no parent for M_k, it is just $M_k(a)$; if there is a parent mediator M_1 of M_k, $M_k(a)=M_k(a,M_1(a))$; if M_2 is a parent of M_1 and M_1 is a parent of M_k, $M_k(a) = M_k(a,M_1(a),M_2(a,M_1(a)))$; and so forth. If the k-th mediator is not causally affected by the exposure, the a in the parenthesis can be dropped.

We have the following corollary (proof in Appendix B):

Corollary 3.0.1

$$TE = \frac{1}{K}\sum_{k=1}^{K} CDE_k(0) + \frac{1}{K}\sum_{k=1}^{K} sCIE_k\ , \tag{7}$$

which shows that the total effect can also be decomposed as the average of the CDEs of all mediators, and the average of the sCIEs of all mediators, reflecting the average percentage of direct and indirect effects across all mediators. This corollary also provides an alternative way to estimate the total effect, which could serve as a less biased estimate by canceling the model mis-specification biases from each single mediator. This is a trade of precision for accuracy, because when the estimate of the average sCIE is less biased, we lose the information of which particular mediator contributes to the sCIE.

3.2 INTERPRETATION OF THE SCALED CONTROLLED INDIRECT EFFECT

Suppose (omitting subscript k)

$$M(1) = M(0) + \Delta M\ ; \tag{8}$$
$$\mathrm{CIE}(1) = \mathrm{CIE}(0) + \Delta C\ . \tag{9}$$

We can look at the extreme cases

$$\mathrm{sCIE} = \Delta M\cdot \mathrm{CIE}(0) = \Delta M\cdot \mathrm{CIE}(1) \qquad \text{if } \Delta C = 0\ ; \tag{10}$$
$$\mathrm{sCIE} = \Delta C\cdot M(0) = \Delta C\cdot M(1) \qquad \text{if } \Delta M = 0\ . \tag{11}$$

When $\Delta C = 0$, i.e. $\mathrm{CIE}(0) = \mathrm{CIE}(1)$, hence no interaction between the mediator and exposure, sCIE only contains the mediated effect which is the difference in the outcome if that mediator is changed from 0 to 1, scaled by the increase in the probability of the mediator. When $\Delta M = 0$, i.e. $M(0) = M(1)$, hence no mediation, sCIE only contains the interaction between the mediator and the exposure, scaled by the constant probability of the mediator. Therefore when $\Delta M \neq 0$ and $\Delta C \neq 0$, sCIE is a mixture of mediation and interaction effects. In contrast, CIE is the total effect of mediator on the outcome.

3.3 IDENTIFICATION ASSUMPTIONS

There are three assumptions needed to identify $M_k(a)$ and $Y_k(a,m)$, and hence $\mathrm{CDE}_k(0)$, $\mathrm{CIE}_k(a)$, and sCIE_k, from observational data.

1. Consistency assumption: an individual's potential outcome under the observed exposure is equal to the observed outcome

$$M_k^{(i)}(a) = M_k^{(i)} \text{ if } A^{(i)} = a\ ; \tag{12}$$
$$Y_k^{(i)}(a,m) = Y^{(i)} \text{ if } A^{(i)} = a, M_k^{(i)} = m\ . \tag{13}$$

Consistency may be violated if there are multiple versions of exposure [Cole and Frangakis, 2009].

2. Positivity assumption: there is a positive probability of receiving every level of exposure for every combination of values of exposure, mediator of interest, and confounding variables in the population. Usually, large sample size can alleviate this assumption. Positivity assumption is an important assumption for weighting based estimation methods such as inverse propensity weight and doubly robust estimation.

3. Ignorability assumption: the exposed and unexposed subjects have equal distributions of potential outcomes when conditioned on confounding variables. This is sometimes referred as exchangeability assumption.We need two ignorability assumptions:

$$M_k(a) \perp\!\!\!\perp A\,|\,L\ ; \tag{14}$$
$$Y_k(a,m) \perp\!\!\!\perp A, M_k\,|\,L\ . \tag{15}$$

These assumptions can be equivalently expressed as the causal DAG is correct. Hence, we can prove the above equations for multiple mediators with causal dependency using d-separation in the single world intervention graph (SWIG) [Richardson and Robins, 2013]. The proof is given in Appendix C. Note that we are not using the natural direct or indirect effect, therefore the much stronger sequential ignorability assumption is not needed [Tingley et al., 2014].

3.4 EFFECT ESTIMATION

CDE, CIE and sCIE are defined as functions of the potential outcomes $M_k(\cdot)$ and $Y(\cdot)$, which need to be estimated from data. Therefore, the unbiasedness property (consistency, zero bias in the limit of infinite data, not to be confused with the consistency assumption in Section 3.3 for causal inference) partially depends on the unbiasedness of $M_k(\cdot)$ and $Y(\cdot)$ (other than other biases such as selection bias or measurement error).

To this end, we use doubly robust estimation [Robins et al., 1994], which entails less biased estimation. The doubly robust property is described by a class of models which admits a doubly robust first order influence function [Robins et al., 2016]. Their influence function has the form of product of two models' influence functions. For example, suppose Y and A are univariate random variables that are dependent on observed data X, the expected product of two conditional expectations $\psi(\theta) = \mathbb{E}_\theta[\mathbb{E}_\theta[Y|X] \cdot \mathbb{E}_\theta[A|X]]$ is a doubly robust estimator [Cui and Tchetgen, 2019]; the other well-known example is the doubly robust estimator for the total effect (average treatment effect), which is unbiased if at least one of the outcome (f function below) or propensity model (g function below) is unbiased.

The doubly robust estimator is written as

$$M_k(a) \approx \frac{1}{N}\sum_{i=1}^{N}\left[f_{M,k}^{(i)} + \frac{\mathbb{1}(A^{(i)}=a)}{g_M^{(i)}}\left(M_k^{(i)} - f_{M,k}^{(i)}\right)\right]; \tag{16}$$

$$Y_k(a,m) \approx \frac{1}{N}\sum_{i=1}^{N}\left[f_{Y,k}^{(i)} + \frac{\mathbb{1}(A^{(i)}=a, M_k^{(i)}=m)}{g_{Y,k}^{(i)}}\left(Y^{(i)} - f_{Y,k}^{(i)}\right)\right], \tag{17}$$

where

$$f_{M,k}^{(i)} = \mathbb{E}\left[M_k | A=a, L^{(i)}\right]; \tag{18}$$

$$f_{Y,k}^{(i)} = \mathbb{E}\left[Y | A=a, M_k=m, L^{(i)}\right]; \tag{19}$$

$$g_M^{(i)} = P(A^{(i)} | L^{(i)}); \tag{20}$$

$$g_{Y,k}^{(i)} = P(M_k^{(i)} | L^{(i)}, A^{(i)}) P(A^{(i)} | L^{(i)}). \tag{21}$$

3.5 MODEL SELECTION AND FITTING

Here we used the principled approach introduced in [Cui and Tchetgen, 2019]. In estimating either TE of exposure on outcome, or CIE (TE of of mediator on outcome), we want to minimize the bias $\mathbb{E}[\psi' - \psi]$, where ψ is the ground truth TE and ψ' is the estimated TE. Directly minimizing this bias is impossible due to unknown ψ. Instead, we minimize a pseudo-risk over different choice of models, where the optimal model choice is least sensitive to perturbations due to model mis-specification. Here we used the mixed minmax solution, which is proved to have a doubly robust property, i.e. zero bias if at least one candidate estimation model is correctly specified. Here, we choose from (1) ℓ_2-norm penalized linear regression or logistic regression; (2) ℓ_2-norm penalized support vector machine (SVM) classifier; (3) random forest; and (4) XGBoost, a type of gradient boosting tree. For the ordinal outcome in the framing dataset introduced later, we used pairwise approach [Liu et al., 2009] to convert ordinal regression problem into binary classification and then solved using the above models.

We used nested cross-validation to fit the models. Nested cross-validation consists of an inner loop and an outer loop. The purpose of the outer loop is to compute an unbiased estimate when applied to data not part of the training set. The purpose of the inner loop was to find the best hyperparameter, C the strength of ℓ_2-norm penalty, to avoid overfitting. The outer loop divided the data into multiple folds. Each fold was used as the testing set, while the other folds were combined and further divided into inner folds. Each inner fold was used as the validation set, while the other inner folds were combined as the training set. The model was trained with a particular C on the training set and evaluated on the validation set. The C with the best average validation performance was chosen and re-fit using the combined training and validation sets. The model was then used to estimate the causal effects on the testing set. The final reported effects were the average effects on the testing sets from the outer loop. The confidence intervals were obtained using bootstrapping which samples the dataset with replacement 1,000 times.

4 RESULTS

4.1 DATASET

The simulated data is generated based on the causal ordering implicated by the DAG in Figure 1 and Figure 2, i.e. $L \longrightarrow A \longrightarrow M \longrightarrow Y$. Each variable is generated as a generalized linear function of its causal parents plus noise. We first randomly generate the coefficients, take the inner product between the coefficients and causal parents plus intercept. The intercept is manually chosen to make the average of the inner product zero. We then added Gaussian noise with standard deviation 1. For binary variables such as A and M, we further applied the sigmoid transformation, and sampled from Bernoulli distribution. The sample size N is 1,000; the number of covariates in L is 10; and the number of mediators in M is 2 or 3 depending on the DAG we study.

We also used a public dataset "framing" used in the R package "mediation" [Tingley et al., 2014]. The detailed description of the framing data can be found in [Brader et al., 2008]. It is a randomized experiment in which the subjects are shown immigration stories with different framing. The exposure is whether the story is framed positively and features an European immigrant. The covariates include age, gender, education level, and income. The mediators include negative emotion and perceived harm. Emotion measures subjects' negative feeling, and is converted to 1 if more or equal to 8. Perceived harm is with respect to increased immigration, and is converted to 1 if more or equal to 7. The outcome is a four-point scale measuring the attitudes toward increased immigration. There are 265 subjects in this dataset. Note that since the exposure is randomized, we used outcome regression instead of doubly robust estimation for this dataset.

4.2 CAUSALLY INDEPENDENT MEDIATORS

We assume two causally independent mediators as shown in Figure 1. It represents the case that A takes effect on Y through two independent mechanisms M_1 and M_2. Note that here we use the example of 2 mediators, but in general it can be more.

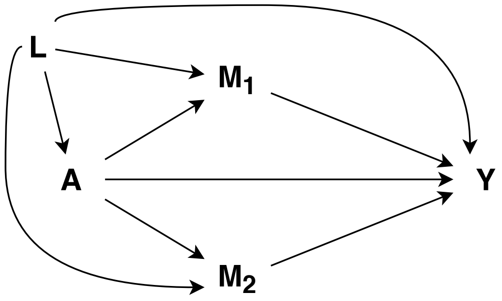

Figure 1: Causal graph with independent mediators. The arrows represent causal influences. M_1 and M_2 are the first and second mediators respectively.

4.2.1 Simulation dataset

The results are shown in Table 1. The model selection method described in Section 3.5 correctly selected linear models (logistic regression and linear regression) for the propensity models and outcome models when estimating $M_k(a)$ and $Y_k(a,m)$. By "correct" we mean that the data is generated using a generalized linear model. Since this is simulated data, we can get the ground truth effects by directly manipulating A and M's. All effects except CDE for M_1 and M_2 and the total effect for M_1 is within the 95% confidence interval. The confidence interval for sCIE is in general wider than that for CIE because sCIE is a function of CIE(a) and $M(a)$ which jointly considers the exposure and mediator.

4.2.2 Framing dataset

For the framing dataset, we used emotion and perceived harm as the two independent mediators. The model selection method selected linear SVM for the outcome models when estimating $Y_k(a,m)$ (outcome regression is used since the exposure is assigned at random). The result is shown in Table 2, which is consistent with the finding that emotion (35.6% sCIE) is a leading mediator compared to perceived harm (18.8% sCIE) when people are making decisions about immigration. But interestingly, the CIE of perceived harm is higher than emotion. In other words, directly reducing perceived harm could be more effective than directly improving the negative emotion (directly intervene the mediator), but it is more difficult to induce perceived harm than to induce negative emotion using different ways of framing (change mediator by intervening the exposure), due to the scaling of mediation effect as well as interaction effect (Equation (10) and (11)). The total effect estimated in [Tingley et al., 2014] Section 6.2 is 0.42 (95% confidence interval [0.17–0.62]). Our estimation is 0.31 to 0.36.

Table 1: The estimated effects and their ground truth value for the simulated data with independent mediators

Effect	**Value**	**True Effect from Simulation**
CDE(M_1)	0.99 [0.75 – 1.24]	1.15
sCIE(M_1)	0.27 [0.12 – 0.42]	0.075
TE(M_1)	1.27 [1.00 – 1.53]	1.23
CIE0(M_1)	4.04 [3.72 – 4.35]	4
CIE1(M_1)	4.10 [3.78 – 4.39]	4
CDE(M_2)	0.95 [0.73 – 1.17]	1.07
sCIE(M_2)	0.36 [0.22 – 0.53]	0.16
TE(M_2)	1.31 [1.09 – 1.60]	1.23
CIE0(M_2)	4.71 [4.44 – 5.00]	5
CIE1(M_2)	4.84 [4.56 – 5.11]	5

Table 2: The estimated effects for the framing dataset when assuming independent mediators

Effect	M_1 **Emotion**	M_2 **Perceived Harm**
CDE	0.16 [-0.12 – 0.77]	0.19 [-0.12 – 0.86]
sCIE	0.20 [-0.014 – 0.29]	0.12 [-0.13 – 0.20]
TE	0.36 [0.026 – 0.81]	0.31 [-0.038 – 0.79]
CIE0	0.74 [0.41 – 1.22]	1.03 [0.58 – 1.43]
CIE1	0.79 [0.38 – 1.10]	1.05 [0.51 – 1.29]

4.3 CAUSALLY DEPENDENT MEDIATORS

We assume three causally dependent mediators as shown in Figure 2. It represents the case that A has an effect on Y

through three mechanisms M_1, M_2, and M_3, while M_1 also causes M_2 and M_3, and M_2 also causes M_3. Note that here we use the example of 3 mediators, but in general it can be more and other causal dependencies as long as there are no cycles.

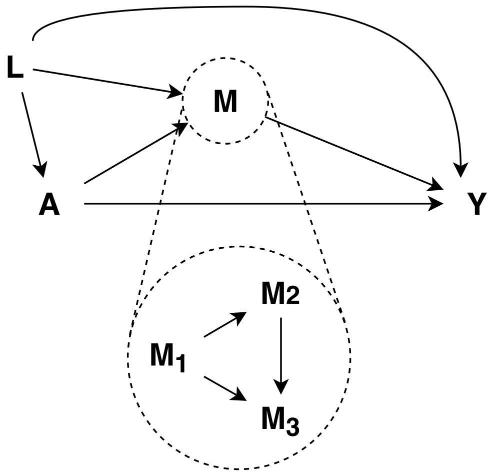

Figure 2: Causal graph with dependent mediators. The arrows represent causal influences. Here we use an example of 3 mediators. M_1, M_2, and M_3 are the mediators which are causally affected as in the figure. Both L and A causally affect each mediator; each mediator causally affect the outcome Y.

4.3.1 Simulation dataset

In Table 3 we show the result. The model selection method described in Section 3.5 again correctly selected linear models (logistic regression and linear regression) for the propensity models and outcome models when estimating $M_k(a)$ and $Y_k(a,m)$. Since this is simulated data, we can get the ground truth effects by directly manipulating A and M's. The true effects are within the 95% confidence interval for M_1 and M_3. M_2 tends to overestimate the indirect effect and underestimate the direct effect.

4.3.2 Framing dataset

The assumption here is that the subject first perceives harm based on the immigration story, which leads to anxiety and then other negative emotions. Therefore in this case M_1 is perceived harm; M_2 is anxiety; and M_3 is the negative emotion.

In Table 4, we can observe that anxiety and emotion have higher scaled CIE (sCIE) than perceived harm, but lower unscaled CIE's. The difference between sCIE and CIE is that CIE only considers the effect on outcome when directly intervening the mediator; while sCIE in addition considers how much the mediator can be changed by intervening the exposure. The result that perceived harm has relatively low sCIE and high CDE and the opposite for anxiety and emotion indicates that perceived harm creates most of its impact through anxiety and other negative emotions as its downstream.

Table 3: The estimated effects and their ground truth value for the simulated data with dependent mediators

Effect	**Value**	**True Effect from Simulation**
CDE(M_1)	1.17 [0.87 – 1.43]	1.26
sCIE(M_1)	0.14 [0.028 – 0.26]	0.10
TE(M_1)	1.31 [1.04 – 1.59]	1.36
CIE0(M_1)	2.82 [2.56 – 3.09]	3.75
CIE1(M_1)	2.83 [2.57 – 3.10]	3.81
CDE(M_2)	1.24 [0.44 – 1.59]	1.15
sCIE(M_2)	0.21 [0.001 – 0.64]	0.18
TE(M_2)	1.45 [0.74 – 1.79]	1.33
CIE0(M_2)	1.77 [1.27 – 2.19]	2.18
CIE1(M_2)	1.77 [1.40 – 2.46]	2.2
CDE(M_3)	1.24 [0.89 – 1.70]	1.21
sCIE(M_3)	0.13 [-0.041 – 0.24]	0.1
TE(M_3)	1.37 [1.02 – 1.71]	1.31
CIE0(M_3)	1.16 [0.84 – 1.50]	1
CIE1(M_3)	1.06 [0.66 – 1.40]	1

Table 4: The estimated effects for the framing dataset when assuming dependent mediators

Effect	M_1 **Perceived Harm**	M_2 **Anxiety**	M_3 **Emotion**
CDE	0.57 [0.12 – 1.29]	0.56 [0.14 – 0.91]	0.63 [0.14 – 1.14]
sCIE	0.01 [-0.30 – 0.27]	0.16 [0.04 – 0.40]	0.11 [-0.05 – 0.26]
TE	0.58 [0.26 – 1.06]	0.71 [0.30 – 1.15]	0.74 [0.30 – 1.16]
CIE0	1.86 [1.40 – 2.08]	0.67 [0.24 – 1.21]	0.82 [0.17 – 1.38]
CIE1	1.43 [0.78 – 1.80]	0.62 [0.17 – 1.10]	0.67 [0.12 – 1.13]

5 DISCUSSION

We have presented the controlled indirect effect approach for mediation analysis with multiple manipulable mediators and causal dependency. Our contributions are detailed in the following aspects: Since the decomposition does not require the cross-world considerations, the effects are directly related to what would happen, say, if the doctors took a particular course of action to treat one comorbidity (mediator). Doing this without the cross-world considerations is also a lead-in to confirmation of a hypothesis in a clinical trial. It is also practical since the controlled indirect effect focuses on the effect of manipulating (treating) one single mediator rather than all of them jointly.

Alternative interpretations of the scaled controlled indirect effect We have

$$\begin{aligned} \text{sCIE} &= (M(0)+\Delta M)(\text{CIE}(0)+\Delta C) - M(0)\text{CIE}(0) \\ &= \Delta C \cdot M(0) + \Delta M \cdot \text{CIE}(0) + \Delta M \cdot \Delta C \\ &= \Delta C \cdot M(0) + \Delta M \cdot \text{CIE}(1)\ . \end{aligned} \tag{22}$$

The last equation shows the consequences of reducing ΔM. The meaning of reducing ΔM is intuitive. If a certain medication or preventive measure reduces the risk of the mediator by a known percent, that percent multiplied by CIE(1) is the amount of outcome prevented, averaged across the exposed population. On the other hand, sCIE can also be viewed as the effect of the mediator of interest on the outcome in the exposed population ($\Delta M \cdot \text{CIE}(1)$) beyond the baseline level of exposure-mediator interaction in the unexposed population ($\Delta C \cdot M(0)$).

Cross validation and model estimation in causal inference Regularization and using cross validation to select the regularization strength is in general not advised in effect estimation, since the loss function of regularized models do not respect the target causal effect. The idea of using perturbation as a pseudo-risk, as used in Section 3.5, represents a possible direction. Other possibilities include adding regularization terms that improves the consistency assumption, such as minimizing the difference between the factual branch of model-based potential outcome vs. the observed value, such as in Matching After Learning To Stretch (MALTS) [Parikh et al., 2018], however, finding good quality matched groups becomes the key, if possible.

Extension to path-specific analysis Path-specific analysis is an extension to mediation analysis by looking at the effect mediated by a path (a bundle of nodes and edges) [Shpitser, 2013]. Longitudinal setting represents a typical use case in path-specific analysis [Nabi et al., 2018]. The idea is TE $= (Y(a) - Y_\pi) + (Y_\pi - Y(a'))$, where Y_π is the effect specific to path π. The decomposition is analogous to TE $= [Y(a,M(a)) - Y(a,M(a'))] + [Y(a,M(a') - Y(a',M(a'))] =$ NIE + NDE, which still requires cross-world counterfactuals. In contrast, our approach represents the "controlled effect" flavor, rather than the "natural effect" flavor, which is TE = CDE + sCIE = CDE(0) + f(CIE(1), CIE(0)), and may be extended to TE $= [Y(a,do(M_\pi = 0)) - Y(a',do(M_\pi = 0))] + f[Y(a,do(M_\pi = m)) - Y(a,do(M_\pi = m')), Y(a',do(M_\pi = m)) - Y(a',do(M_\pi = m'))]$. The "controlled effect" flavor has the advantages that CIE directly simulates what if the mediator path are intervened, answering a realizable question: what if I intervene the mediating path? The disadvantage is that CIE is TE of the mediator on outcome, which is subject to unmeasured confounding. The "natural effect" flavor has the advantage that it deals with unmeasured confounding, but NIE and NDE cannot be interpreted in a realizable way. As an important future work, extension of the controlled flavor into path-specific effects is needed.

Significance to aritificial intelligence Causal inference and mediation analysis make up an under-represented but scientifically valuable field in artificial intelligence and machine learning applied to healthcare. They help healthcare practitioners and researchers understand the underlying data-generating mechanisms by prospectively or retrospectively observing patients. In general, machine learning algorithms that take causality into account have great potential to guide decision-making in healthcare based not on association but on causality, improving the algorithm performance and transferability to different settings since the causal mechanisms are stable [Peters et al., 2017].

Significance to healthcare The approach developed in this paper provides a new method for mediation analysis that, when applied to a healthcare problem, can provide insight into the consequences of preventing or treating a comorbidity that mediates the effect of a particular disease on a particular outcome. In fact, much of medicine is devoted to mitigating the effects of a disease by treating a resultant comorbidity. Our method provides a principled way to quantify the possible effect or clinical benefit, of such a mitigation strategy on downstream clinical outcomes. Additionally, by allowing causal dependencies among multiple mediators, this provides flexibility for a healthcare provider to consider the mediator of interest in clinically realistic scenario.

Limitations First, our analysis is limited to the case where both A and M are binary (0 or 1) making it restrictive in applications. Although it is a helpful simplification to indicate if the mediator (comorbidity) is treated or not, in reality comorbidities can be reduced without being fully treated. Second, we have not considered other types of contrast. In the present work we have focused on the difference between two potential outcomes. But depending on the data type of Y and M, different decomposition equations need to be derived and validated.

Mediation analysis requires a relatively large sample size. This is because mediation analysis divides the data into multiple strata, i.e. samples with and without the presence of each mediator in both exposed and unexposed groups. And there should be enough samples in each stratum to reduce sampling bias. In the case of nested cross-validation, the sample size should be even larger to make sure each fold in the inner loop has enough samples. The fact that our approach deals with each mediator one by one reduces the need for large sample size so that the samples need not grow with the number of mediators. This is helpful but does not completely resolve this limitation. Monte Carlo based power analysis can be done by generating the data using models estimated from actual data, up to the point significance is shown [Schoemann et al., 2017].

6 CONCLUSION

The proposed approach can be used to assess the importance of multiple manipulable mediators with causal dependencies. In the case of healthcare problems where the mediators are comorbidities or side-effects of certain exposures, our approach provides principled guidance for choosing which mediator to treat in order to optimize the healthcare outcome.

References

Andrea Bellavia and Linda Valeri. Decomposition of the total effect in the presence of multiple mediators and interactions. *American Journal of Epidemiology*, 187(6): 1311–1318, 2018.

Ted Brader, Nicholas A Valentino, and Elizabeth Suhay. What triggers public opposition to immigration? anxiety, group cues, and immigration threat. *American Journal of Political Science*, 52(4):959–978, 2008.

Stephen R Cole and Constantine E Frangakis. The consistency statement in causal inference: a definition or an assumption? *Epidemiology*, 20(1):3–5, 2009.

Yifan Cui and Eric Tchetgen Tchetgen. Selective machine learning of doubly robust functionals, 2019.

RM Daniel, BL De Stavola, SN Cousens, and Stijn Vansteelandt. Causal mediation analysis with multiple mediators. *Biometrics*, 71(1):1–14, 2015.

Angus Deaton and Nancy Cartwright. Understanding and misunderstanding randomized controlled trials. *Social Science & Medicine*, 210:2–21, 2018.

Thomas A Glass, Steven N Goodman, Miguel A Hernán, and Jonathan M Samet. Causal inference in public health. *Annual review of public health*, 34:61–75, 2013.

MA Hernán and JM Robins. *Causal Inference: What If*. Boca Raton, FL. Chapman & Hall/CRC, 2020.

Kosuke Imai and Teppei Yamamoto. Identification and sensitivity analysis for multiple causal mechanisms: Revisiting evidence from framing experiments. *Political Analysis*, 21(2):141–171, 2013.

Kosuke Imai, Luke Keele, and Dustin Tingley. A general approach to causal mediation analysis. *Psychological methods*, 15(4):309, 2010.

Tie-Yan Liu et al. Learning to rank for information retrieval. *Foundations and Trends® in Information Retrieval*, 3(3): 225–331, 2009.

Wen Wei Loh, Beatrijs Moerkerke, Tom Loeys, and Stijn Vansteelandt. Interventional effect models for multiple mediators. *arXiv preprint arXiv:1907.08415*, 2019.

Razieh Nabi, Phyllis Kanki, and Ilya Shpitser. Estimation of personalized effects associated with causal pathways. In *Conference on Uncertainty in Artificial Intelligence*, volume 2018. NIH Public Access, 2018.

Jerzy S Neyman. On the application of probability theory to agricultural experiments. *Annals of Agricultural Sciences*, 10:1–51, 1923.

Harsh Parikh, Cynthia Rudin, and Alexander Volfovsky. Malts: Matching after learning to stretch. *arXiv preprint arXiv:1811.07415*, 2018.

Judea Pearl. *Causality*. Cambridge University Press, 2009.

Judea Pearl. Interpretation and identification of causal mediation. *Psychological methods*, 19(4):459, 2014.

Jonas Peters, Dominik Janzing, and Bernhard Schölkopf. *Elements of causal inference: foundations and learning algorithms*. MIT press, 2017.

Jean-Baptiste Pingault, Paul F O'reilly, Tabea Schoeler, George B Ploubidis, Frühling Rijsdijk, and Frank Dudbridge. Using genetic data to strengthen causal inference in observational research. *Nature Reviews Genetics*, 19 (9):566–580, 2018.

Thomas S Richardson and James M Robins. Single world intervention graphs (swigs): A unification of the counterfactual and graphical approaches to causality. *Center for the Statistics and the Social Sciences, University of Washington Series. Working Paper*, 128(30):2013, 2013.

James Robins. A new approach to causal inference in mortality studies with a sustained exposure period—application to control of the healthy worker survivor effect. *Mathematical modelling*, 7(9-12):1393–1512, 1986.

James Robins, Lingling Li, Eric Tchetgen Tchetgen, and Aad van der Vaart. Technical report: Higher order influence functions and minimax estimation of nonlinear functionals. *arXiv preprint arXiv:1601.05820*, 2016.

James M Robins and Thomas S Richardson. Alternative graphical causal models and the identification of direct effects. *Causality and psychopathology: Finding the determinants of disorders and their cures*, pages 103–158, 2010.

James M Robins, Andrea Rotnitzky, and Lue Ping Zhao. Estimation of regression coefficients when some regressors are not always observed. *Journal of the American statistical Association*, 89(427):846–866, 1994.

James M Robins, Thomas S Richardson, and Ilya Shpitser. An interventionist approach to mediation analysis. *arXiv preprint arXiv:2008.06019*, 2020.

Donald B Rubin. Bayesian inference for causal effects: The role of randomization. *The Annals of statistics*, pages 34–58, 1978.

Alexander M Schoemann, Aaron J Boulton, and Stephen D Short. Determining power and sample size for simple and complex mediation models. *Social Psychological and Personality Science*, 8(4):379–386, 2017.

Ilya Shpitser. Counterfactual graphical models for longitudinal mediation analysis with unobserved confounding. *Cognitive science*, 37(6):1011–1035, 2013.

Dustin Tingley, Teppei Yamamoto, Kentaro Hirose, Luke Keele, and Kosuke Imai. Mediation: R package for causal mediation analysis. 2014.

Tyler VanderWeele. *Explanation in causal inference: methods for mediation and interaction*. Oxford University Press, 2015.

Tyler VanderWeele and Stijn Vansteelandt. Mediation analysis with multiple mediators. *Epidemiologic methods*, 2 (1):95–115, 2014.

Tyler J VanderWeele, Stijn Vansteelandt, and James M Robins. Effect decomposition in the presence of an exposure-induced mediator-outcome confounder. *Epidemiology (Cambridge, Mass.)*, 25(2):300, 2014.

Stijn Vansteelandt and Rhian M Daniel. Interventional effects for mediation analysis with multiple mediators. *Epidemiology (Cambridge, Mass.)*, 28(2):258, 2017.

Liuyi Yao, Zhixuan Chu, Sheng Li, Yaliang Li, Jing Gao, and Aidong Zhang. A survey on causal inference, 2020.

SUPPLEMENTARY MATERIAL

A PROOF OF EQUATION (1)

We have the total effect as

$$\begin{aligned} \text{TE} &= Y(a=1) - Y(a=0)\\ &= Y(1) - Y(0)\\ &= Y\left(1, M_1(1, Pa\{M_1\}(1)), \cdots, M_K(1, Pa\{M_K\}(1))\right)\\ &\qquad - Y\left(0, M_1(0, Pa\{M_1\}(0)), \cdots, M_K(0, Pa\{M_K\}(0))\right) . \end{aligned} \tag{23}$$

Expanding the k-th mediator, we have (see next page)

Similarly, we have

$$\begin{aligned} Y(0) &= Y\left(0, M_1(0, Pa\{M_1\}(0)), \cdots, M_K(0, Pa\{M_K\}(0))\right)\\ &= Y\left(0, M_1(0, Pa\{M_1\}(0)), \cdots, 0, \cdots, M_K(0, Pa\{M_K\}(0))\right) + M_k(0, Pa\{M_k\}(0)) \text{CIE}_k(0) . \end{aligned} \tag{25}$$

Therefore,

$$\begin{aligned} TE &= Y(1) - Y(0)\\ &= \Big[Y\left(1, M_1(1, Pa\{M_1\}(1)), \cdots, 0, \cdots, M_K(1, Pa\{M_K\}(1))\right)\\ &\qquad - Y\left(0, M_1(0, Pa\{M_1\}(0)), \cdots, 0, \cdots, M_K(0, Pa\{M_K\}(0))\right)\Big]\\ &\qquad + \Big[M_k(1, Pa\{M_k\}(1)) \text{CIE}_k(1) - M_k(0, Pa\{M_k\}(0)) \text{CIE}_k(0)\Big]\\ &= \underbrace{\Big[Y_k(1,0) - Y_k(0,0)\Big]}_{\text{CDE}_k(0)} + \underbrace{\Big[M_k(1)\text{CIE}_k(1) - M_k(0)\text{CIE}_k(0)\Big]}_{\text{sCIE}_k} . \end{aligned} \tag{26}$$

B PROOF OF COROLLARY 3.0.1

Equation (24) and (25) are general equations obtained by expanding the k-th mediator. We repeat this for all mediators $1, \ldots, K$, so that

$$TE = \text{CDE}_1(0) + \text{sCIE}_1 ; \tag{27}$$

$$\cdots$$

$$TE = \text{CDE}_K(0) + \text{sCIE}_K . \tag{28}$$

Therefore,

$$TE = \frac{1}{K} \sum_{k=1}^{K} \text{CDE}_k(0) + \frac{1}{K} \sum_{k=1}^{K} \text{sCIE}_k . \tag{29}$$

C PROOF OF IGNORABILITY

We can graphically prove Equation (14) $M_k(a) \perp\!\!\!\perp A \,|\, L$ by constructing the single world intervention graph (SWIG) as in Figure 3b. The conditional independence is true since all connections between $M_k(a)$ and A must go through L, which is blocked by conditioning on L based on d-separation.

We can also graphically prove Equation (15) $Y_k(a,m) \perp\!\!\!\perp A, M_k \,|\, L$ by constructing the SWIG as in Figure 3c. The conditional independence is true since all connections between $Y_k(a,m)$ and A, M_k must go through L, which is blocked by conditioning on L based on d-separation.

$$
\begin{aligned}
Y(1) &= Y\left(1, M_1(1, Pa\{M_1\}(1)), \cdots, M_K(1, Pa\{M_K\}(1))\right) \\
&= Y\left(1, M_1(1, Pa\{M_1\}(1)), \cdots, 1, \cdots, M_K(1, Pa\{M_K\}(1))\right) M_k(1, Pa\{M_k\}(1)) \\
&\quad + Y\left(1, M_1(1, Pa\{M_1\}(1)), \cdots, 0, \cdots, M_K(1, Pa\{M_K\}(1))\right) \left(1 - M_k(1, Pa\{M_k\}(1))\right) \\
&= Y\left(1, M_1(1, Pa\{M_1\}(1)), \cdots, 0, \cdots, M_K(1, Pa\{M_K\}(1))\right) \\
&\quad + M_k(1, Pa\{M_k\}(1)) \Big[Y\left(1, M_1(1, Pa\{M_1\}(1)), \cdots, 1, \cdots, M_K(1, Pa\{M_K\}(1))\right) \\
&\qquad\qquad\qquad\qquad - Y\left(1, M_1(1, Pa\{M_1\}(1)), \cdots, 0, \cdots, M_K(1, Pa\{M_K\}(1))\right) \Big] \\
&= Y\left(1, M_1(1, Pa\{M_1\}(1)), \cdots, 0, \cdots, M_K(1, Pa\{M_K\}(1))\right) + M_k(1, Pa\{M_k\}(1)) \mathrm{CIE}_k(1)\ .
\end{aligned}
\tag{24}
$$

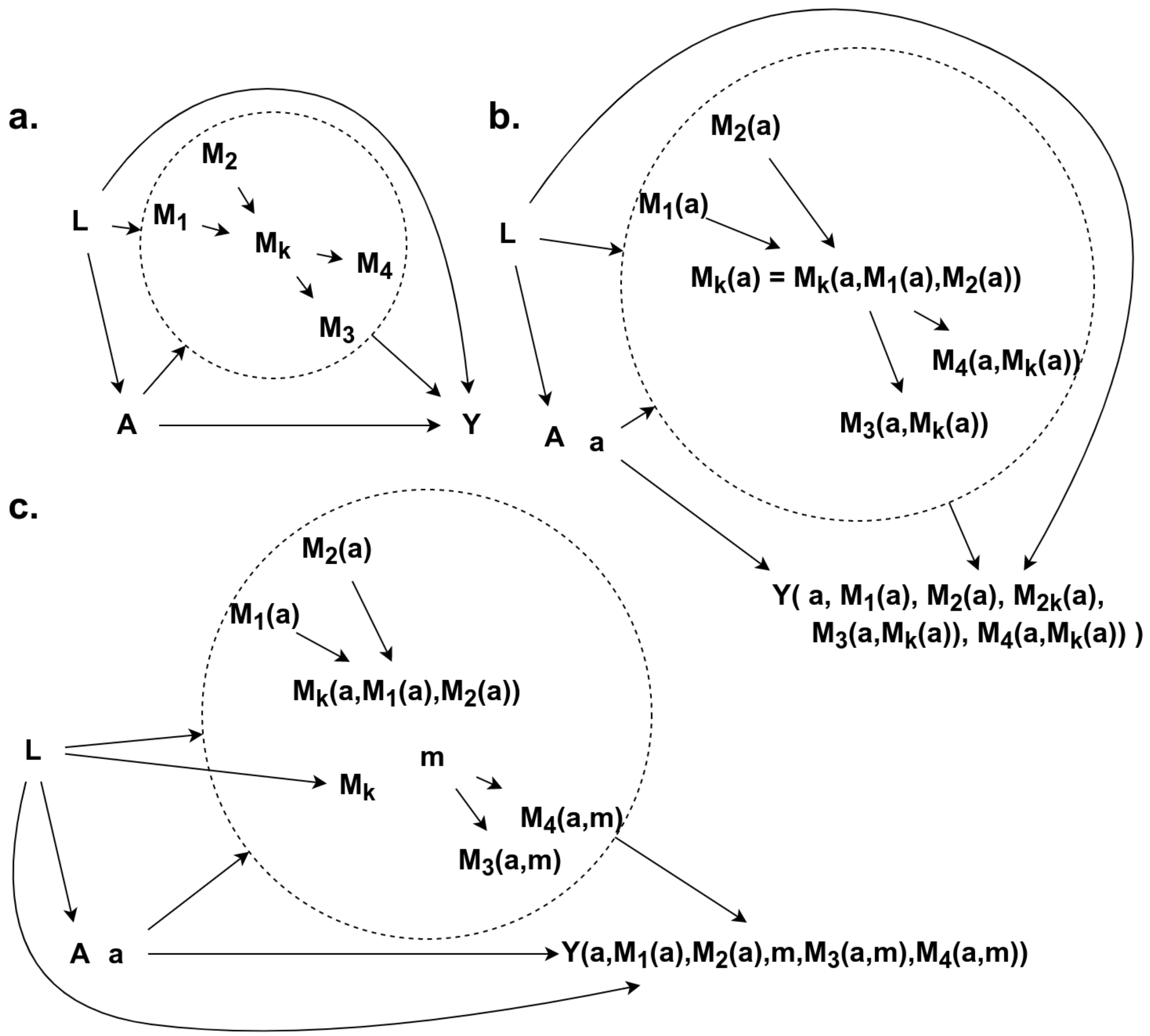

Figure 3: (a) A general causal graph where the mediators in the dashed circle represent multiple mediators with causal dependence. Both L and A causally affect each mediator; each mediator causally affect the outcome Y. Here we study the k-th mediator M_k, which has M_1 and M_2 as its parents and M_3 and M_4 as its children. (b) The SWIG of panel a when intervening A to a, so that the exposure value a and the observed A are separated; and the mediators becomes potential outcome for a. We ignored the arrows pointing into the outcome. (c) The SWIG of panel a when intervening A to a and M_k to k. Note that there are three versions of M_k: M_k is the observed value when no intervention is applied; $M_k(a,M_1(a),M_2(a))$ is the potential outcome of M_k when intervening A to a; and m is the intervened value of M_k.