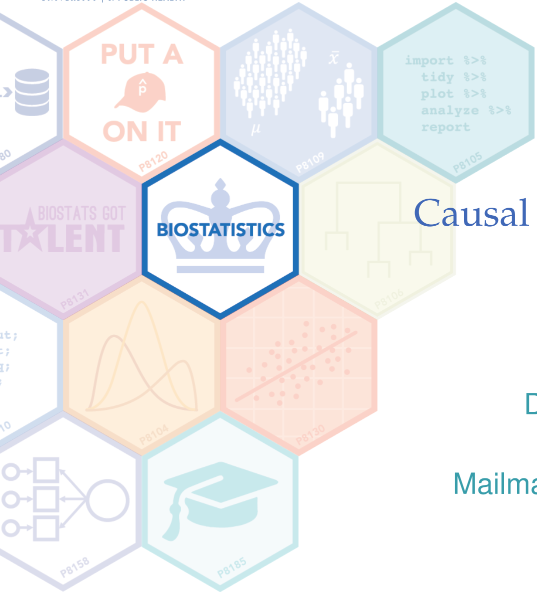
COLUMBIA UNIVERSITY | MAILMAN SCHOOL of PUBLIC HEALTH



Causal Mediation Analysis with STATA

Linda Valeri
Department of Biostatistics
Columbia University
Mailman School of Public Health

February 23, 2023

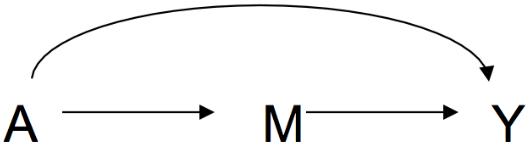
Outline

- Principles of Mediation and Interaction
- Motivating example
- Formal definitions
- Assumptions
- Regression approaches for mediation analysis allowing for exposure-mediator interaction
- med4way command
- Application

Principles of Mediation and Interaction

- Mediation and interaction are two potential mechanisms explaining why and how a causal effect is observed.

- Mediation: is the exposure causing the outcome through changing an intermediate variable (indirect effect) or through other pathways, independent of the intermediate variable (direct effect)?



Principles of Mediation and Interaction

- Mediation and interaction are two potential mechanisms explaining why and how a causal effect is observed

- Interaction: is the effect of the exposure on the outcome different if we intervene changing an additional variable (synergistic or antagonistic interaction is possible)?

- Effect modification: for whom the exposure has an effect?

Example: Income disparities in cancer survival

- Colorectal cancer (CRC) is the third most common cancer in the United States and also one cancer that demonstrates widening mortality disparities across income groups.
- Differential access to screening, leading to disparities in stage at diagnosis has been long hypothesized a determinant of such disparities.
- Mediation: Are the income disparities in survival due to income disparities in stage at diagnosis?
- Interaction: would income disparities in survival be different if stage at diagnosis were fixed at I vs IV?
- Effect modification: are income disparities in survival different among patients with stage I at diagnosis compared to patients with stage IV at diagnosis?

Motivations for studying mediation

1. Scientific understanding and explanation
   E.g. Do genetic variants affect lung cancer through smoking or independently?
2. Confirmation or refutation of theory
   E.g. Does low early SES effect adult health principally by setting an economic trajectory later in life?
3. Limiting the effects of exposure by intervening on a mediator
   E.g. Can we eliminate the effects of antipsychotic medication on mortality by preventing the primary mechanism for mortality?
4. Refinement of Interventions
   - Improving components of an intervention to target mechanism
   - Eliminating costly ineffective components of an intervention
   - Understanding why an intervention failed

Counterfactual approach to mediation

- Defines direct and indirect effects in terms of the counterfactual intervention [i.e. fixing exposure and mediator to a predefined value (controlled), or fixing the exposure to a predefined value and the mediator to the value that naturally follows (natural)].
- Definitions are non-parametric and the total effect decomposes into the sum of natural direct and indirect effect.
- Provide a framework to clarify assumptions for identification of the effects in experimental and observational data.

Formal definitions

For $Y(a, m)$ denoting the outcome when we intervene fixing $A = a$ and $M = m$ and $M(a)$ denoting the mediator when we intervene fixing the exposure $A = a$

- Controlled Direct Effect: $CDE(m) = E[Y(a, m) - Y(a^*, m)]$
- Natural Direct Effect: $NDE = E[Y(a, M(a^*)) - Y(a^*, M(a^*)]$
- Natural Indirect Effect: $NIE = E[Y(a, M(a)) - Y(a, M(a^*)]$
- Total Effect= $NDE + NIE$

Natural effects provide information on mechanisms, while controlled effects can be interpreted in terms of interventions (Robins and Greenland, 1992).

Assumptions for identification

Counterfactual objects can not be identified at the individual level (they would require observing an individual in both the real and counterfactual world), but we are able to estimate such effects at the population levels and by making a set of assumptions (Pearl, 2001; VanderWeele, 2009).

1. No unmeasured exposure-outcome confounding given C
2. No unmeasured mediator-outcome confounding given C
3. No unmeasured exposure-mediator confounding given C
4. No mediator-outcome confounder affected by the exposure

Note: assumptions (1) and (3) are satisfied automatically if the exposure is randomized, but not (2) and (4).
Note: estimating the CDE only requires assumptions (1) and (2) to be satisfied.

Counterfactual approach to interaction

When the two exposures A and M are independent, we have the following decomposition (VanderWeele and Tchetgen Tchetgen, 2014):

$$E[Y(m=1)-Y(m=0)] = (E[Y(a=0,m=1)] - E[Y(a=0,m=0)]) +$$
$$(E[Y(a=1,m=1)] - E[Y(a=1,m=0)] - E[Y(a=0,m=1)] + E[Y(a=0,m=0)])P(A)$$

The effect of M on Y decomposes into two parts:

- The effect of M on Y in the absence of A
- The proportion of the effect of M on Y due to interaction with A

Note: this decomposition for the effect of A on Y will have to be modified to account for mediation if A affects M.
Note: need to assume no unmeasured confounding of $M - Y$ and $A - Y$ relationships.

Mediation analysis with exposure-mediator interaction

Let Y denote the continuous outcome, M the continuous intermediate variables, A the exposure and C additional covariates of interest.

$$\begin{aligned} E[Y|a,m,c] &= \theta_0 + \theta_1 a + \theta_2 m + \theta_3 am + \theta_4^{'} c \\ E[M|a,c] &= \beta_0 + \beta_1 a + \beta_2^{'} c \end{aligned}$$

Provided that the models are correctly specified and the identification assumptions (i)-(iv) hold, controlled direct effects, natural direct and indirect effects are derived as (Valeri and VanderWeele, 2013):

$$\begin{aligned} CDE &= (\theta_1 + \theta_3 m)(a - a^*) \\ NDE &= (\theta_1 + \theta_3\beta_0 + \theta_3\beta_1 a^* + \theta_3\beta_2^{'} C)(a - a^*) \\ NIE &= (\theta_2\beta_1 + \theta_3\beta_1 a)(a - a^*) \end{aligned}$$

Estimation and Inference

- Effects can be estimated as function of regression parameters with asymptotic inference using delta method
- A simulation-based approach has been developed by Imai et al (2010) which imputes potential outcomes directly
- This allows to specify much more flexible models for the outcome and the mediator
- Inference can also be obtained via bootstrapping

Unification of Mediation and Interaction

Vanderweele (2014) showed that, by using the counterfactual approach, the total effect can be decomposed into four different components:

1. direct controlled effect
2. pure natural indirect effect
3. interaction alone (reference interaction)
4. mediation and interaction (mediated interaction).

This 4-way decomposition provides the maximum insight on clarifying the contribution of interactive and mediating mechanisms to a given observed total effect.

Unification of Mediation and Interaction

Component	Interpretation
CDE	Treatment effect neither due to mediation nor interaction
INTref	Treatment effect only due to interaction
INTmed	Treatment effect due to both mediation and interaction
NIE	Treatment effect only due to mediation

This 4-way decomposition provides the maximum insight on clarifying the contribution of interactive and mediating mechanisms to a given observed total effect.

Overview of commands for causal mediation analysis in STATA

- paramed: was the first Stata command to be developed for conducting causal mediation analysis allowing for exposure-mediator interaction (Emsley, Liu, Valeri, VanderWeele, 2012).
- medeff: is the Stata command for implementing the imputation approach by Imai. It is based on the R package developed by the Authors, and is computationally intensive (Hicks and Tingley, 2011).
- med4way: implements causal mediation analysis integrating the analysis of survival outcomes and the four-way decomposition of the total effects

Some critical topics that could/should be integrated include:

- Multiple mediators and interactions (Vanderweele & Vansteelandt 2014, Bellavia & Valeri 2019)
- High-dimensional mediation analysis (Blum, Valeri et al. 2020)
- Time-varying exposures, mediators, and confounders (VanderWeele et al. 2017)
- Sensitivity analyses for the counterfactual approach (Vanderweele 2010, Valeri and VanderWeele 2014)
- Multiple imputations

med4way Introduction

- Discacciati, A., Bellavia, A., Lee, J.J., Mazumdar, M., Valeri, L. Med4way: a Stata command to investigate mediating and interactive mechanisms using the four-way effect decomposition. International Journal of Epidemiology. 2019 Feb;48(1):15-20.
- A Stata command for the 4-way decomposition using parametric regression models
- https://github.com/anddis/med4way
- net install med4way, from(”https://raw.githubusercontent.com/anddis/med4way/master/”) replace

Facts

- help med4way
- current version: v2.3.1 - 25jul2019
- uses parametric regression models to estimate the components of the 4-way decomposition of the total effect of an exposure on a outcome in the presence of a mediator with which the exposure may interact.
- This decomposition breaks down the total effect of the exposure on the outcome into components due to mediation alone, to interaction alone, to both mediation and interaction, and to neither mediation nor interaction
- Improves computational speed by using the delta method for calculating standard errors components using the delta method (default) or the bootstrap
- allows continuous, binary, count or survival outcomes, and continuous or binary mediators

Models

Two regression models are fitted:

- model for the outcome (as a function of the exposure, the mediator, their interaction and confounders)
  - linear
  - logistic, log-binomial, Poisson, negative binomial
  - accelerated failure time (exponential, Weibull) and Cox
- model for the mediator (as a function of the exposure and confounders
  - linear
  - logistic
- the causal effects are automatically computed by the command as a function of the regression parameters estimated from the above specified models.

Survival outcome

- In the case of a survival outcome, the outcome variable must be omitted.
- data must be read as survival time, using Stata's stset command,
- med4way is fully integrated with Statas way of handling survival data
- The variable for the interaction between the exposure and the mediator is automatically generated and added to the model for the outcome.

Command

med4way [yvar] avar mvar [cvars], a0() a1() m() yreg() mreg()

- **a0()** specifies the referent level of the exposure;
- **a1(**) specifies the actual level of the exposure;
- **m()** specifies the level of the mediator at which the controlled direct effect is computed (0 recommended to obtain the decomposition)
- **yreg()** specifies the form of the regression model for the outcome
- **mreg()** specifies the form of the regression model for the mediator

Other options

- c() fixes the values of the confounders
- bootstrap
- fulloutput
- casecontrol

SEP disparities in cancer survival

We wish to quantify to which extent the effect of SEP (socio-economic position) inequalities on cancer mortality are mediated through stage at diagnosis (advanced versus non advanced)

- Data for Surveillance Epidemiology End Results (SEER) linked to American Community Survey (ACS) data for patients diagnosed in 1992-2005 and followed up to 2010
- We allow for interaction between the SEP exposure, county median income, and the mediator, stage at diagnosis
- Potential confounders: gender, age at diagnosis, year at diagnosis and state of residence
- Since the outcome is failure time a censoring variable was defined taking value 1 if the individual is censored or value 0 if the event is observed
- The survival outcome is studied using an accelerated failure time model assuming a Weibull distribution

Example: med4way command

```
. stset stset srv_time_mon_pa, failure(censor)

. med4way new_medianincome stage_dich_noinsitu sex rac_reca ///
  age_c date_c ///
  state_2 state_3 state_4 state_5 state_6 state_7 state_8, ///
  a0(25000) a1(75000) m(1) yreg(aft, we) mreg(logistic)

. med4way new_medianincome stage_dich_noinsitu sex rac_reca ///
  age_c date_c ///
  state_2 state_3 state_4 state_5 state_6 state_7 state_8, ///
  a0(25000) a1(75000) m(0) yreg(aft, we) mreg(logistic)
```

Example: med4way command

```
Summary

  Outcome    (yvar):  srv_time_mon_pa
  Exposure   (avar):  new_medianincome
  Mediator   (mvar):  stage_dich_noinsitu
  Covariates (cvars): sex rac_reca age_c date_c state_2
  state_3 state_4 state_5 state_6 state_7 state_8

  Model for the outcome  (yreg): aft, weibull
  Model for the mediator (mreg): logistic

  Referent exposure level (a0):              25000
  Actual exposure level   (a1):              75000
  Mediator level for the decomposition (m): 1
```

Example: outcome model

```
Weibull AFT regression

No. of subjects =       93,797      Number of obs     =      93,797
No. of failures =       62,738
Time at risk    =      6716569
                                    LR chi2(14)       =    35775.51
Log likelihood  =   -136942.97      Prob > chi2       =      0.0000

------------------------------------------------------------------------------------
                           _t |    Coef. Std. Err.      z  P>|z|  [95% Conf. Interval]
------------------------------+-----------------------------------------------------
              new_medianincome | 0.000006 0.000001    9.08 0.000  0.000005    0.000007
           stage_dich_noinsitu |-1.403908 0.042142  -33.31 0.000 -1.486505   -1.321310
_new_medianinXstage_dich_~000 |-0.000008 0.000001   -8.63 0.000 -0.000009   -0.000006
                           sex | 0.176502 0.009440   18.70 0.000  0.157999    0.195005
                      rac_reca |-0.213340 0.015855  -13.46 0.000 -0.244415   -0.182265
                         age_c |-0.048775 0.000425 -114.65 0.000 -0.049609   -0.047941
                         _cons | 4.647863 0.037425  124.19 0.000  4.574513    4.721214

------------------------------+-----------------------------------------------------
                         /ln_p | -0.154948 0.003305 -46.89 0.000  -0.161425   -0.148471
------------------------------+-----------------------------------------------------
                             p |  0.856460 0.002830                0.850930    0.862025
                           1/p |  1.167597 0.003859                1.160059    1.175184
------------------------------------------------------------------------------------
```

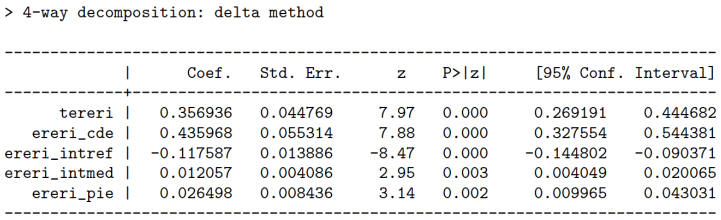
Example: mediator model

```
 Model for the mediator

Iteration 0:   log likelihood = -55851.309
Iteration 1:   log likelihood = -55678.911
Iteration 2:   log likelihood = -55678.204
Iteration 3:   log likelihood = -55678.204

Logistic regression                             Number of obs     =    100,000
                                                LR chi2(12)       =     346.21
                                                Prob > chi2       =     0.0000
Log likelihood = -55678.204                     Pseudo R2         =     0.0031

------------------------------------------------------------------------------------
stage_dich_noinsitu |       Coef.   Std. Err.       z    P>|z|     [95% Conf. Interval]
--------------------+---------------------------------------------------------------
   new_medianincome |   -0.000003    0.000001   -3.17    0.002    -0.000005   -0.000001
                sex |    0.000180    0.014822    0.01    0.990    -0.028872    0.029231
           rac_reca |    0.331178    0.023816   13.91    0.000     0.284501    0.377856
              age_c |    0.003913    0.000572    6.84    0.000     0.002792    0.005035
              _cons |   -1.041371    0.053634  -19.42    0.000    -1.146492   -0.936249
------------------------------------------------------------------------------------
```

Interpretation models

- The survival outcome model yielded a positive effect of income on survival, a negative effect of stage
- A negative, significant interaction between tumor stage at diagnosis and county household median income was detected
- The logistic regression analysis showed a negative association between the SEP measure and stage at diagnosis

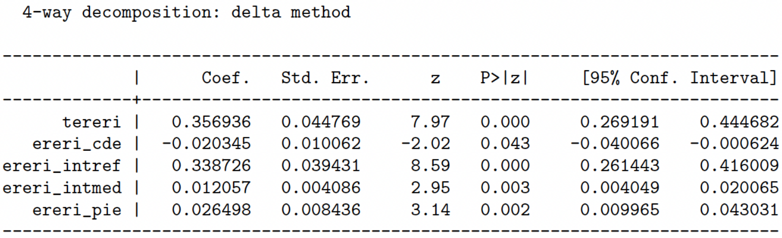
Example: total effect



```
> 4-way decomposition: delta method

------------------------------------------------------------------------------
             |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      tereri |   0.356936   0.044769     7.97   0.000     0.269191    0.444682
   ereri_cde |   0.435968   0.055314     7.88   0.000     0.327554    0.544381
ereri_intref |  -0.117587   0.013886    -8.47   0.000    -0.144802   -0.090371
ereri_intmed |   0.012057   0.004086     2.95   0.003     0.004049    0.020065
   ereri_pie |   0.026498   0.008436     3.14   0.002     0.009965    0.043031
------------------------------------------------------------------------------
```

The total effect of 1.35 indicates that the mean survival time of individuals if they lived in counties with median income of 75,000 would be higher than if they lived in counties with median income of 25,000 in the mean survival ratio scale (total excess relative risk = TE-1 = 0.35)

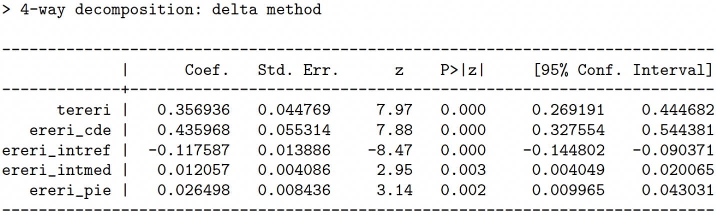
Example: m=1



```
  4-way decomposition: delta method

------------------------------------------------------------------------------
             |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      tereri |   0.356936   0.044769     7.97   0.000     0.269191    0.444682
   ereri_cde |  -0.020345   0.010062    -2.02   0.043    -0.040066   -0.000624
ereri_intref |   0.338726   0.039431     8.59   0.000     0.261443    0.416009
ereri_intmed |   0.012057   0.004086     2.95   0.003     0.004049    0.020065
   ereri_pie |   0.026498   0.008436     3.14   0.002     0.009965    0.043031
------------------------------------------------------------------------------
```

The controlled direct effect (ereri_cde), controlling the mediator at level m=1 reveals that, had we intervened setting stage at diagnosis to be advanced for all individuals, mean survival time of individuals living in counties with median income of 75,000 would be lower than that of individuals living in counties with median income of 25,000, with a relative excess risk due to controlled direct effect estimated as -0.02

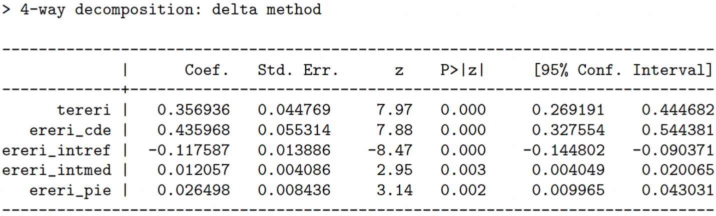
Example: m=0, Only direct effect



```
> 4-way decomposition: delta method

-----------------------------------------------------------------------------
             |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+---------------------------------------------------------------
      tereri |   0.356936   0.044769     7.97   0.000     0.269191    0.444682
   ereri_cde |   0.435968   0.055314     7.88   0.000     0.327554    0.544381
ereri_intref |  -0.117587   0.013886    -8.47   0.000    -0.144802   -0.090371
ereri_intmed |   0.012057   0.004086     2.95   0.003     0.004049    0.020065
   ereri_pie |   0.026498   0.008436     3.14   0.002     0.009965    0.043031
-----------------------------------------------------------------------------
```

The controlled direct effect (ereri_cde), controlling the mediator at level m=0 reveals that, had we intervened setting stage at diagnosis to be not advanced for all individuals, mean survival time of individuals living in counties with median income of 75,000 would be higher than that of individuals living in counties with median income of 25,000, with a relative excess risk due to controlled direct effect estimated as 0.43.

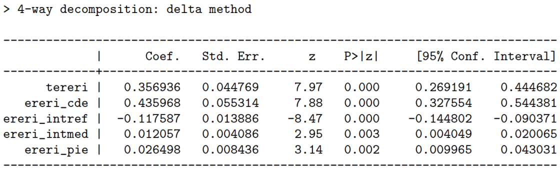
Example: Interaction only



```
> 4-way decomposition: delta method

------------------------------------------------------------------------------
             |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      tereri |   0.356936   0.044769     7.97   0.000     0.269191    0.444682
   ereri_cde |   0.435968   0.055314     7.88   0.000     0.327554    0.544381
ereri_intref |  -0.117587   0.013886    -8.47   0.000    -0.144802   -0.090371
ereri_intmed |   0.012057   0.004086     2.95   0.003     0.004049    0.020065
   ereri_pie |   0.026498   0.008436     3.14   0.002     0.009965    0.043031
------------------------------------------------------------------------------
```

An estimate of -0.11 for relative excess risk due to interaction only (ereri_intref) indicates that income disparities would be lower had individuals been diagnosed with advanced stage at diagnosis.

Interpretation Mediation and Interaction



```
> 4-way decomposition: delta method

------------------------------------------------------------------------------
             |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      tereri |   0.356936   0.044769     7.97   0.000     0.269191    0.444682
   ereri_cde |   0.435968   0.055314     7.88   0.000     0.327554    0.544381
ereri_intref |  -0.117587   0.013886    -8.47   0.000    -0.144802   -0.090371
ereri_intmed |   0.012057   0.004086     2.95   0.003     0.004049    0.020065
   ereri_pie |   0.026498   0.008436     3.14   0.002     0.009965    0.043031
------------------------------------------------------------------------------
```

- Since one of the components of the 4-way decomposition is negative, it is not meaningful to report the proportion of effect due to each component
- Still we can comment that the direct effect and interaction only effect displayed the larger effect sizes.
- The relative excess risk due to mediated interaction and pure indirect effect were estimated as 0.012 and 0.026 respectively, indicating that the effect of SEP on survival is only partly explained by its effect on stage at diagnosis.

Example: Conclusion

- In explaining the mechanisms that lead to income disparities in survival outcomes stage at diagnosis appears to be involved via both interactive and mediating mechanisms.
- The former role seems more important that the latter.
- mediating and interactive mechanisms operate in opposite directions.
- The results of this example should be interpreted with caution as several biases might be present:
    - aggregate measure of socio-economic position potential ecologic bias.
    - no-unmeasured confounding assumptions might be violated
    - stage at diagnosis might be misclassified

Discussion

- Introduced counterfactual framework for mediation analysis in the presence of interaction
- med4way is the first Stata command for the decomposition of the total effect in mediating and interactive effects allowing for continuous, binary, count and survival outcome and for a continuous or binary mediator.
- The command accommodates both cohort and case-control designs
- Methods for missing data and measurement error that might induce bias in the analyses are being incorporated
- We developed a decomposition of the total effect in the presence of multiple mediators.

References

- Bellavia A, **Valeri L**. Decomposition of the total eect in the presence of multiple mediators and interactions. American journal of epidemiology. 2018
- Discacciati, A., Bellavia, A., Lee, J.J., Mazumdar, M., **Valeri L**. Med4way: a Stata command to investigate mediating and interactive mechanisms using the four-way effect decomposition. International Journal of Epidemiology. 2019 Feb;48(1):15-20.
- Emsley R, Liu H, **Valeri L**, VanderWeele TJ. PARAMED: Stata module to perform causal mediation analysis using parametric regression models.
- Hicks R, Tingley D. Causal mediation analysis. The Stata Journal. 2011
- Imai K, et al. A general approach to causal mediation analysis. Psyc. methods. 2010
- **Valeri L**. and VanderWeele, T.J. (2013) Mediation analysis allowing for exposure-mediator interactions and causal interpretation: theoretical assumptions and implementation with SAS and SPSS macros. Psychological Methods, 18:137-150.
- VanderWeele TJ. Bias formulas for sensitivity analysis for direct and indirect effects. Epidemiology. 2010
- VanderWeele TJ, Tchetgen EJ. Mediation analysis with time varying exposures and mediators. Journal of the Royal Statistical Society. Series B. 2017
- VanderWeele T, Vansteelandt S. Mediation analysis with multiple mediators. Epidemiologic methods. 2014

Thank you!

lv2424@columbia.edu
@valeritweety