Supplemental Appendix 1

1) Deforestation causing malaria

Here we show the problem with assuming the causal graph displayed in Fig. 2A instead of the correct causal graph depicted in Fig. 2B in the main manuscript.

- *Simulation setup*

Let malaria, deforestation, and aerosol concentration in county i be denoted by m_i, d_i, and a_i, respectively. We will assume that

$$m_i = \beta_0 + \beta_1 d_i + \beta_2 o_i + e_i$$

$$d_i = \gamma_0 + \gamma_1 o_i + \delta_i$$

where o_i is an omitted variable. These equations show that both d_i and m_i are influenced by o_i. We further assume that deforestation influences aerosol concentration:

$$a_i = \psi_0 + \psi_1 d_i + \epsilon_i$$

In these expressions, $\beta_0, \beta_1, \beta_2, \gamma_0, \gamma_1, \psi_0$, and ψ_1 are regression coefficients. The error terms e_i, δ_i, and ϵ_i are given by:

$$e_i \sim N(0, \sigma_e^2)$$

$$\epsilon_i \sim N(0, \sigma_\epsilon^2)$$

$$\delta_i \sim N(0, \sigma_\delta^2)$$

In relation to the unobserved confounder, we assume that:

$$o_i \sim N(\mu_o, \sigma_o^2)$$

The primary goal of the analysis is to estimate the effect of deforestation on malaria, given by β_1.

- *Derivation of the instrumental variable estimate of β_1 when deforestation influences aerosol concentration*

Using the instrumental variable (IV) methodology, we can estimate β_1 as:

$$\widehat{\beta_1} = \frac{\partial m}{\partial d} = \frac{\frac{\partial m}{\partial a}}{\frac{\partial d}{\partial a}} = \frac{\frac{Cov(m,a)}{Var(a)}}{\frac{Cov(d,a)}{Var(a)}} = \frac{Cov(m,a)}{Cov(d,a)}$$

Notice that

$$Cov(m,a) = Cov(\beta_0 + \beta_1[\gamma_0 + \gamma_1 o_i + \delta_i] + \beta_2 o_i + e_i, \psi_0 + \psi_1[\gamma_0 + \gamma_1 o_i + \delta_i] + \epsilon_i)$$

$$= Cov(\beta_1\gamma_1 o_i + \beta_1\delta_i + \beta_2 o_i, \psi_1\gamma_1 o_i + \psi_1\delta_i)$$

$$= Cov([\beta_1\gamma_1 + \beta_2]o_i, \psi_1\gamma_1 o_i) + Cov(\beta_1\delta_i, \psi_1\delta_i) = [\beta_1\gamma_1 + \beta_2]\psi_1\gamma_1 Var(o_i) + \beta_1\psi_1 Var(\delta_i)$$

Similarly, note that:

$$Cov(d,a) = Cov(\gamma_0 + \gamma_1 o_i + \delta_i, \psi_0 + \psi_1[\gamma_0 + \gamma_1 o_i + \delta_i] + \epsilon_i)$$

$$= Cov(\gamma_1 o_i + \delta_i, \psi_1\gamma_1 o_i + \psi_1\delta_i) = \gamma_1\psi_1\gamma_1 Var(o_i) + \psi_1 Var(\delta_i)$$

Putting these two results together, the IV estimate for β_1 is given by:

$$\widehat{\beta_1} = \frac{\partial m}{\partial d} = \frac{\frac{\partial m}{\partial a}}{\frac{\partial d}{\partial a}} = \frac{[\beta_1\gamma_1 + \beta_2]\psi_1\gamma_1\sigma_o^2 + \beta_1\psi_1\sigma_\delta^2}{\gamma_1\psi_1\gamma_1\sigma_o^2 + \psi_1\sigma_\delta^2} = \frac{[\beta_1\gamma_1 + \beta_2]\gamma_1\sigma_o^2 + \beta_1\sigma_\delta^2}{\gamma_1\gamma_1\sigma_o^2 + \sigma_\delta^2}$$

Clearly the effect of deforestation on malaria estimated using an IV approach ($\widehat{\beta_1}$) can be very different from the true β_1.

- *Simulations*

To illustrate how this IV approach can generate parameter estimates of the opposite sign as the true parameter, we rely on a simple numerical example using the R [1] script below. In this example, we are trying to estimate $\beta_1 = 1$, which indicates that as deforestation increases, malaria cases also increase. However, using the IV approach adopted by [2], we obtain $\widehat{\beta_1} = -0.96$. In other words, the estimate of β_1 using this IV approach would incorrectly suggest that as deforestation increases, malaria cases decrease. This script enables users to change parameters to determine how those changes affect the difference between the IV estimate $\widehat{\beta_1}$ and the true parameter β_1.

```
rm(list=ls(all=TRUE))
set.seed(1)
nobs=100000

#main parameters of the simulation
b0=0
b1=1 #positive effect of deforestation on malaria
b2=-2
g0=0
g1=1
g2=1
p0=0
p1=1
var.o=5
var.delta=0.1
var.e=0.5
var.epsilon=0.1

#simulate data
o=rnorm(nobs,mean=-1,sd=sqrt(var.o))          #omitted variable
d=rnorm(nobs,mean=g0+g1*o,sd=sqrt(var.delta))   #deforestation variable
a=rnorm(nobs,mean=p0+p1*d,sd=sqrt(var.epsilon)) #instrumental variable
m=rnorm(nobs,mean=b0+b1*d+b2*o,sd=sqrt(var.e))  #response variable

#naive estimate based on a standard regression
res=lm(m~d)
res$coefficients[2]

# IV estimate for beta_1 using empirical approach 1
tmp=lm(m~a)
slope.ma=tmp$coefficient[2]
tmp=lm(d~a)
slope.da=tmp$coefficient[2]
slope.ma/slope.da

# IV estimate for beta_1 using empirical approach 2
res=lm(d~a)
d.pred=predict(res)
res1=lm(m~d.pred)
res1$coefficients[2]

# IV estimate for beta_1 using the theoretical results from this Appendix
part1=(b1*g1+b2)*g1*var.o
part2=b1*var.delta
part3=(g1^2)*var.o
part4=var.delta
(part1+part2)/(part3+part4)
```

2) Malaria causing deforestation

Here we show the problem with assuming the causal graph displayed in Fig. 2C instead of the potentially more accurate causal graph depicted in Fig. 2D in the main manuscript.

- *Simulation setup*

Let deforestation, malaria, and optimal temperature in county i be denoted by d_i, m_i, and t_i. We will assume that

$$d_i = \beta_0 + \beta_1 m_i + \beta_2 o_i + e_i$$

$$m_i = \gamma_0 + \gamma_1 o_i + \gamma_2 t_i + \delta_i$$

where o_i is an omitted variable. These equations show that both d_i and m_i are influenced by o_i but that t_i only influences m_i. We further assume that the omitted variable is influenced by temperature:

$$o_i = \psi_0 + \psi_1 t_i + \epsilon_i$$

In these expressions, $\beta_0, \beta_1, \beta_2, \gamma_0, \gamma_1, \psi_0$, and ψ_1 are regression coefficients. The error terms e_i, δ_i, and ϵ_i are given by:

$$e_i \sim N(0, \sigma_e^2)$$

$$\epsilon_i \sim N(0, \sigma_\epsilon^2)$$

$$\delta_i \sim N(0, \sigma_\delta^2)$$

In relation to temperature, we assume that:

$$t_i \sim N(\mu_t, \sigma_t^2)$$

The primary goal of the analysis is to estimate the effect of malaria on deforestation, given by β_1.

- *Derivation of the instrumental variable estimate of β_1 when temperature influences the omitted variable*

Using the instrumental variable (IV) methodology, we can estimate β_1 as:

$$\widehat{\beta_1} = \frac{\partial d}{\partial m} = \frac{\frac{\partial d}{\partial t}}{\frac{\partial m}{\partial t}} = \frac{\frac{Cov(d,t)}{Var(t)}}{\frac{Cov(m,t)}{Var(t)}} = \frac{Cov(d,t)}{Cov(m,t)}$$

Notice that

$$Cov(d,t) = Cov(\beta_0 + \beta_1[\gamma_0 + \gamma_1[\psi_0 + \psi_1 t_i + \epsilon_i] + \gamma_2 t_i + \delta_i] + \beta_2[\psi_0 + \psi_1 t_i + \epsilon_i] + e_i, t_i)$$

$$= Cov(\beta_1\gamma_1\psi_1 t_i + \beta_1\gamma_2 t_i + \beta_2\psi_1 t_i, t_i)$$

$$= [\beta_1\gamma_1\psi_1 + \beta_1\gamma_2 + \beta_2\psi_1]Var(t_i)$$

Similarly, note that:

$$Cov(m,t) = Cov(\gamma_0 + \gamma_1 o_i + \gamma_2 t_i + \delta_i, t_i) = Cov(\gamma_1\psi_1 t_i + \gamma_2 t_i, t_i) = [\gamma_1\psi_1 + \gamma_2]Var(t_i)$$

Therefore, we have that the IV estimate for β_1 is given by:

$$\widehat{\beta_1} = \frac{\partial d}{\partial m} = \frac{\frac{\partial d}{\partial t}}{\frac{\partial m}{\partial t}} = \frac{\beta_1[\gamma_1\psi_1 + \gamma_2] + \beta_2\psi_1}{\gamma_1\psi_1 + \gamma_2}$$

Clearly the effect of deforestation on malaria estimated using an IV approach ($\widehat{\beta_1}$) can be very different from the true β_1. Importantly, if $\psi_1 = 0$ (temperature does not affect the omitted variable), then clearly the IV approach works as intended because $\widehat{\beta_1} = \beta_1$

- *Numerical example*

To illustrate how this IV approach can generate parameter estimates of the opposite sign as the true parameter, we rely on a simple numerical example. In this example, we are trying to estimate $\beta_1 = 1$, which indicates that as malaria cases increase, deforestation increases. However, using the IV approach adopted by MacDonald and Mordecai [2], we obtain $\widehat{\beta_1} = -1$. In other words, the estimate of β_1 using this IV approach would incorrectly suggest that as malaria cases increase, deforestation decreases.

- *Simulations*

The results provide above can be confirmed with the script given below. This script was created in R [1] and enables users to change parameters to determine how those changes affect the difference between the IV estimate $\widehat{\beta_1}$ and the true parameter β_1.

```
rm(list=ls(all=TRUE))
set.seed(1)
nobs=100000

#main parameters of the simulation
b0=0
b1=1    #positive effect of malaria on deforestation
b2=2
g0=0
g1=-2
g2=1
p0=0
p1=1
var.t=5
var.delta=0.1
var.e=0.5
var.epsilon=0.1

#simulate data
t=rnorm(nobs,mean=2,sd=sqrt(var.t))              #temperature
o=rnorm(nobs,mean=p0+p1*t,sd=sqrt(var.epsilon))    #omitted variable
m=rnorm(nobs,mean=g0+g1*o+g2*t,sd=sqrt(var.delta))  #malaria variable
d=rnorm(nobs,mean=b0+b1*m+b2*o,sd=sqrt(var.e))     #deforestation variable

#naive estimate based on a standard regression
res=lm(d~m)
res$coefficients[2]

# IV estimate for beta_1 using empirical approach 1
tmp=lm(d~t)
slope.dt=tmp$coefficient[2]
tmp=lm(m~t)
slope.mt=tmp$coefficient[2]
slope.dt/slope.mt

# IV estimate for beta_1 using empirical approach 2
res=lm(m~t)
m.pred=predict(res)
res1=lm(d~m.pred)
res1$coefficients[2]

# IV estimate for beta_1 using the theoretical results from this Appendix
part1=b1*(g1*p1+g2)
part2=b2*p1
part3=g1*p1
part4=g2
(part1+part2)/(part3+part4)
```

References

1. R Core Team, 2020. R: A Language and Environment for Statistical Computing. Available at: https://www.R-project.org/. Accessed.

2. MacDonald AJ, Mordecai EA, 2019. Amazon deforestation drives malaria transmission, and malaria burden reduces forest clearing. Proceedings of the National Academy of Science 116: 22212-22218.