**Appendix 5. Companion code**

####################################################

##Section 1: Simulating causal effect

####################################################

niterations<-1000

blanks<-rep(NA,niterations)

results\_CE<-data.frame(TrueRR=blanks, TrueRD=blanks,

CrudeRR=blanks, CrudeRD=blanks,

Doomed\_E1=blanks, Causal\_E1=blanks, Immune\_E1=blanks,

Doomed\_E0=blanks, Causal\_E0=blanks, Immune\_E0=blanks)

#Section 1A: Set prevalence of exogenous variables

pZ1<-0.06

pZ2<-0.1

pLowEd<-0.7 #the exposure ("E" in the formulas below)

#Section 1B: Simulate cohort of n=10,000 individuals

sampsize<-10000

set.seed(12345) #set a seed so that results can be replicated

for (i in 1:niterations){

#Simulate exogeneous variables

df<-data.frame(ID=seq(1,sampsize),

Z1=rbinom(sampsize,1,pZ1),

Z2=rbinom(sampsize,1,pZ2),

LowEd=rbinom(sampsize,1,pLowEd),

Dementia=NA)

#Simulate endogenous variables based on SCCs

df$Dementia<-ifelse((df$LowEd==1 & df$Z1==1) | df$Z2==1,1,0) #whether an individual has dementia or not (the outcome, "Y") depends on their value of causal components in the pies whose completion leads to dementia

#Section 1C: Calculate effects

pY\_E0<-sum(df$LowEd==0 & df$Dementia==1)/sum(df$LowEd==0)

pY\_E1<-sum(df$LowEd==1 & df$Dementia==1)/sum(df$LowEd==1)

CrudeRR<-pY\_E1/pY\_E0

CrudeRD<-pY\_E1-pY\_E0

TrueRR <- (pZ1+pZ2-pZ1\*pZ2)/pZ2

TrueRD <- (pZ1+pZ2-pZ1\*pZ2)-pZ2

#For causation scenario we expect Crude measures = True measures because there are no sources of non-exchangeability

#Section 1D: "Under the hood" (define response types based on SCCs)

df$RT<-ifelse(df$Z2==1,"Doomed",ifelse(df$Z1==1, "Causal","Immune") )

RTtab<- t(table(df$LowEd,df$RT)/c(sum(df$LowEd==0),sum(df$LowEd==1)))

RTtab

#Store results

results\_CE$TrueRR[i]<-TrueRR

results\_CE$TrueRD[i]<-TrueRD

results\_CE$CrudeRR[i]<-CrudeRR

results\_CE$CrudeRD[i]<-CrudeRD

results\_CE$Doomed\_E1[i]<-RTtab["Doomed","1"]

results\_CE$Doomed\_E0[i]<-RTtab["Doomed","0"]

results\_CE$Causal\_E1[i]<-RTtab["Causal","1"]

results\_CE$Causal\_E0[i]<-RTtab["Causal","0"]

results\_CE$Immune\_E1[i]<-RTtab["Immune","1"]

results\_CE$Immune\_E0[i]<-RTtab["Immune","0"]

}

#Calculate 95% CI

mean\_CIs<-function(data, estimate){

results<-data.frame(Est=estimate)

results$LCI<-quantile(data[,estimate],probs=.025)

results$mean<-mean(data[,estimate])

results$UCI<-quantile(data[,estimate],probs=.975)

return(results)

}

#Auto-generate a results table

for (i in 1:length(colnames(results\_CE))){

if (i==1){

res\_tab\_CE<-mean\_CIs(results\_CE,colnames(results\_CE)[i])

} else{

res\_tab\_CE<-rbind(res\_tab\_CE, mean\_CIs(results\_CE,colnames(results\_CE)[i]))}

}

#Print table of results:

#Average true and crude RRs and RDs across simulations and their 95% CI

#Average prevalence of response types within exposure groups across simulations and their 95% CI

res\_tab\_CE

#Drop all objects except results tables

rm(list=grep("res\_tab",ls(),value=TRUE,invert=TRUE))

####################################################

##Section 2: Simulating confounding

####################################################

niterations<-1000

blanks<-rep(NA,niterations)

results\_conf<-data.frame(TrueRR=blanks, TrueRD=blanks,

CrudeRR=blanks, CrudeRD=blanks,

AdjRR=blanks, AdjRD=blanks,

Doomed\_E1=blanks,Immune\_E1=blanks,

Doomed\_E0=blanks,Immune\_E0=blanks)

#Section 2A: Set prevalence of exogenous variables

pQ1<-0.8

pQ2<-0.2

pZ1<-0 #setting the causal partner of the exposure to 0 results in a null effect

pZ3<-0.1

pZ4<-0.15

pLowSES<-0.7

#Section 2B: Simulate cohort of n=10,000 individuals

sampsize<-10000

set.seed(12345) #set a seed so that results can be replicated

for (i in 1:niterations){

#Simulate exogeneous variables

df<-data.frame(ID=seq(1,sampsize),

Q1=rbinom(sampsize,1,pQ1),

Q2=rbinom(sampsize,1,pQ2),

LowSES=rbinom(sampsize,1,pLowSES),

Z1=rbinom(sampsize,1,pZ1),

Z3=rbinom(sampsize,1,pZ3),

Z4=rbinom(sampsize,1,pZ4),

LowEd=NA,

Dementia=NA)

#Simulate endogenous variables based on SCCs

df$LowEd<-ifelse((df$LowSES==1 & df$Q1==1) | df$Q2==1,1,0) #in Scenario 2 (confounding), the exposure (LowEd or "E" in the formulas below) is no longer exogenous; each individual's value for LowEd emerges from the values of the components that cause it

df$Dementia<-ifelse((df$LowEd==1 & df$Z1==1) | df$Z3==1 | (df$LowSES==1 & df$Z4==1),1,0) #whether an individual has dementia or not (the outcome, "Y") depends on their value of causal components in the pies whose completion leads to dementia

#Section 2C: Calculate effects

#Truth (true ATE)

TrueRR<-(pZ1+pZ3+pLowSES\*pZ4-pZ1\*pZ3-pZ1\*pLowSES\*pZ4-pZ3\*pLowSES\*pZ4-pZ1\*pZ3\*pLowSES\*pZ4)/(pZ3+pLowSES\*pZ4-pZ3\*pLowSES\*pZ4)

TrueRD<-(pZ1+pZ3+pLowSES\*pZ4-pZ1\*pZ3-pZ1\*pLowSES\*pZ4-pZ3\*pLowSES\*pZ4-pZ1\*pZ3\*pLowSES\*pZ4)-(pZ3+pLowSES\*pZ4-pZ3\*pLowSES\*pZ4)

#Crude estimates

pY\_E0<-sum(df$LowEd==0 & df$Dementia==1)/sum(df$LowEd==0)

pY\_E1<-sum(df$LowEd==1 & df$Dementia==1)/sum(df$LowEd==1)

CrudeRR<-pY\_E1/pY\_E0

CrudeRD<-pY\_E1-pY\_E0

CrudeRR

CrudeRD

#Standardized (adjusted) estimates

pY\_E0\_C0<-sum(df$LowEd==0 & df$LowSES==0 & df$Dementia==1)/sum(df$LowEd==0 & df$LowSES==0)

pY\_E0\_C1<-sum(df$LowEd==0 & df$LowSES==1 & df$Dementia==1)/sum(df$LowEd==0 & df$LowSES==1)

pY\_E1\_C0<-sum(df$LowEd==1 & df$LowSES==0 & df$Dementia==1)/sum(df$LowEd==1 & df$LowSES==0)

pY\_E1\_C1<-sum(df$LowEd==1 & df$LowSES==1 & df$Dementia==1)/sum(df$LowEd==1 & df$LowSES==1)

pC1<-nrow(df[df$LowSES==1,])/nrow(df)

pC0<-1-pC1

pY\_E1\_std<-pY\_E1\_C1\*pC1+pY\_E1\_C0\*pC0

pY\_E0\_std<-pY\_E0\_C1\*pC1+pY\_E0\_C0\*pC0

AdjRR <- pY\_E1\_std/pY\_E0\_std

AdjRD <- pY\_E1\_std-pY\_E0\_std

AdjRR

AdjRD

#Section 2D: "Under the hood" (define response types based on SCCs)

df$RT<-ifelse((df$Z3==1 | (df$LowSES==1 & df$Z4==1 )),"Doomed", ifelse(df$Z1==1, "Causal","Immune") )

RTtab<-t(table(df$LowEd,df$RT)/c(sum(df$LowEd==0),sum(df$LowEd==1)))

RTtab

#Store results

results\_conf$TrueRR[i]<-TrueRR

results\_conf$TrueRD[i]<-TrueRD

results\_conf$CrudeRR[i]<-CrudeRR

results\_conf$CrudeRD[i]<-CrudeRD

results\_conf$AdjRR[i]<-AdjRR

results\_conf$AdjRD[i]<-AdjRD

results\_conf$Doomed\_E1[i]<-RTtab["Doomed","1"]

results\_conf$Doomed\_E0[i]<-RTtab["Doomed","0"]

results\_conf$Immune\_E1[i]<-RTtab["Immune","1"]

results\_conf$Immune\_E0[i]<-RTtab["Immune","0"]

}

#Calculate 95% CI

mean\_CIs<-function(data, estimate){

results<-data.frame(Est=estimate)

results$LCI<-quantile(data[,estimate],probs=.025)

results$mean<-mean(data[,estimate])

results$UCI<-quantile(data[,estimate],probs=.975)

return(results)

}

#Auto-generate a results table

for (i in 1:length(colnames(results\_conf))){

if (i==1){

res\_tab\_conf<-mean\_CIs(results\_conf,colnames(results\_conf)[i])

} else{

res\_tab\_conf<-rbind(res\_tab\_conf, mean\_CIs(results\_conf,colnames(results\_conf)[i]))}

}

#Print table of results:

#Average true, crude, and adjusted RRs and RDs across simulations and their 95% CI

#Average prevalence of response types within exposure groups across simulations and their 95% CI (no causal types because simulating under the null)

res\_tab\_conf

#Drop all objects except results tables

rm(list=grep("res\_tab",ls(),value=TRUE,invert=TRUE))

####################################################

##Section 3: Simulating collider bias

####################################################

niterations<-1000

blanks<-rep(NA,niterations)

results\_coll<-data.frame(

TrueRRpop=blanks, TrueRDpop=blanks,

CrudeRRpop=blanks, CrudeRDpop=blanks,

CrudeRRsamp=blanks, CrudeRDsamp=blanks,

AdjRRsamp=blanks, AdjRDsamp=blanks,

Doomed\_E1pop=blanks,Immune\_E1pop=blanks,

Doomed\_E0pop=blanks,Immune\_E0pop=blanks,

Doomed\_E1samp=blanks,Immune\_E1samp=blanks,

Doomed\_E0samp=blanks,Immune\_E0samp=blanks,

nstudy=blanks)

#Section 3A: Set prevalence of exogenous variables

pX1<-0.5

pX2<-0.1

pX3<-0.5

pZ1<-0 #setting the causal partner of the exposure to 0 results in a null effect

pZ5<-0.1

pZ6<-0.4

pAPOE4<-0.25 #APOE4 is a cause of selection into the study sample, this is prevalence in the population

pLowEd<-0.7 #the exposure ("E" in the formulas below) is also a cause of being selected into the study sample

#Section 3B: Simulate cohort of n=10,000 individuals

sampsize<-10000

set.seed(12345) #set a seed so that results can be replicated

for (i in 1:niterations) {

#Simulate exogeneous variables

df<-data.frame(ID=seq(1,sampsize),

X1=rbinom(sampsize,1,pX1),

X2=rbinom(sampsize,1,pX2),

X3=rbinom(sampsize,1,pX3),

APOE4=rbinom(sampsize,1,pAPOE4),

Z1=rbinom(sampsize,1,pZ1),

Z5=rbinom(sampsize,1,pZ5),

Z6=rbinom(sampsize,1,pZ6),

LowEd=rbinom(sampsize,1,pLowEd),

StudyPart=NA,

Dementia=NA)

#Simulate endogenous variables based on SCCs

df$StudyPart<-ifelse((df$LowEd==0 & df$X1==1) | df$X2==1 | (df$APOE4==1 & df$X3==1),1,0) #Recall that high education (i.e., LowEd==0) is a cause of study participation

df$Dementia<-ifelse((df$LowEd==1 & df$Z1==1) | df$Z5==1 | (df$APOE4==1 & df$Z6==1),1,0)

#Section 3C: Calculate effects

#Truth (true ATE)

TrueRRpop<-(pZ1+pZ5+pAPOE4\*pZ6-pZ1\*pZ5-pZ1\*pAPOE4\*pZ6-pZ5\*pAPOE4\*pZ6-pZ1\*pZ5\*pAPOE4\*pZ6)/(pZ5+pAPOE4\*pZ6-pZ5\*pAPOE4\*pZ6)

TrueRDpop<-(pZ1+pZ5+pAPOE4\*pZ6-pZ1\*pZ5-pZ1\*pAPOE4\*pZ6-pZ5\*pAPOE4\*pZ6-pZ1\*pZ5\*pAPOE4\*pZ6)-(pZ5+pAPOE4\*pZ6-pZ5\*pAPOE4\*pZ6)

#With selection bias, we want to compare estimates in the whole population with estimates among the selected sample (our study participants)

#In whole population (dataframe "df")

pY\_E0<-sum(df$LowEd==0 & df$Dementia==1)/sum(df$LowEd==0)

pY\_E1<-sum(df$LowEd==1 & df$Dementia==1)/sum(df$LowEd==1)

#Crude estimates in whole population

CrudeRRpop<-pY\_E1/pY\_E0

CrudeRDpop<-pY\_E1-pY\_E0

CrudeRRpop

CrudeRDpop

#In those included in study sample (create new dataframe, "df\_study")

df\_study<-df[df$StudyPart==1,]

pY\_E0\_Study<-sum(df\_study$LowEd==0 & df\_study$Dementia==1)/sum(df\_study$LowEd==0)

pY\_E1\_Study<-sum(df\_study$LowEd==1 & df\_study$Dementia==1)/sum(df\_study$LowEd==1)

#Crude estimates in study sample

CrudeRR\_Study<-pY\_E1\_Study/pY\_E0\_Study

CrudeRD\_Study<-pY\_E1\_Study-pY\_E0\_Study

CrudeRR\_Study

CrudeRD\_Study

#Standardized (adjusted) estimates: Need to correct estimate in study sample with standardization for APOE4

pY\_E0<-sum(df\_study$LowEd==0 & df\_study$Dementia==1)/sum(df\_study$LowEd==0)

pY\_E1<-sum(df\_study$LowEd==1 & df\_study$Dementia==1)/sum(df\_study$LowEd==1)

pY\_E0\_C0<-sum(df\_study$LowEd==0 & df\_study$APOE4==0 & df\_study$Dementia==1)/sum(df\_study$LowEd==0 & df\_study$APOE4==0)

pY\_E0\_C1<-sum(df\_study$LowEd==0 & df\_study$APOE4==1 & df\_study$Dementia==1)/sum(df\_study$LowEd==0 & df\_study$APOE4==1)

pY\_E1\_C0<-sum(df\_study$LowEd==1 & df\_study$APOE4==0 & df\_study$Dementia==1)/sum(df\_study$LowEd==1 & df\_study$APOE4==0)

pY\_E1\_C1<-sum(df\_study$LowEd==1 & df\_study$APOE4==1 & df\_study$Dementia==1)/sum(df\_study$LowEd==1 & df\_study$APOE4==1)

sum(df\_study$LowEd==1 & df\_study$Dementia==1)/sum(df\_study$LowEd==1)

sum(df\_study$LowEd==0 & df\_study$Dementia==1)/sum(df\_study$LowEd==0)

pC1<-nrow(df\_study[df\_study$APOE4==1,])/nrow(df\_study)

pC0<-1-pC1

pY\_E1\_std<-pY\_E1\_C1\*pC1+pY\_E1\_C0\*pC0

pY\_E0\_std<-pY\_E0\_C1\*pC1+pY\_E0\_C0\*pC0

AdjRR\_Study <- pY\_E1\_std/pY\_E0\_std

AdjRD\_Study <- pY\_E1\_std-pY\_E0\_std

AdjRR\_Study

AdjRD\_Study

#Section 3D: "Under the hood" (define response types based on SCCs)

df$RT<-ifelse((df$Z5==1 | (df$APOE4==1 & df$Z6==1 )),"Doomed", ifelse(df$Z1==1, "Causal","Immune") )

#Dist of RT in population

RTtabpop<-t(table(df$LowEd,df$RT)/c(sum(df$LowEd==0),sum(df$LowEd==1)))

RTtabpop

#Dist of RT in study sample

df\_study$RT<-ifelse((df\_study$Z5==1 | (df\_study$APOE4==1 & df\_study$Z6==1 )),"Doomed", ifelse(df\_study$Z1==1, "Causal","Immune") )

RTtabsamp<-t(table(df\_study$LowEd,df\_study$RT)/c(sum(df\_study$LowEd==0),sum(df\_study$LowEd==1)))

RTtabsamp

#Save results

results\_coll$TrueRRpop[i]<-TrueRRpop

results\_coll$TrueRDpop[i]<-TrueRDpop

results\_coll$CrudeRRpop[i]<-CrudeRRpop

results\_coll$CrudeRDpop[i]<-CrudeRDpop

results\_coll$CrudeRRsamp[i]<-CrudeRR\_Study

results\_coll$CrudeRDsamp[i]<-CrudeRD\_Study

results\_coll$AdjRRsamp[i]<-AdjRR\_Study

results\_coll$AdjRDsamp[i]<-AdjRD\_Study

results\_coll$Doomed\_E1pop[i]<-RTtabpop["Doomed","1"]

results\_coll$Doomed\_E0pop[i]<-RTtabpop["Doomed","0"]

results\_coll$Immune\_E1pop[i]<-RTtabpop["Immune","1"]

results\_coll$Immune\_E0pop[i]<-RTtabpop["Immune","0"]

results\_coll$Doomed\_E1samp[i]<-RTtabsamp["Doomed","1"]

results\_coll$Doomed\_E0samp[i]<-RTtabsamp["Doomed","0"]

results\_coll$Immune\_E1samp[i]<-RTtabsamp["Immune","1"]

results\_coll$Immune\_E0samp[i]<-RTtabsamp["Immune","0"]

results\_coll$nstudy[i]<-nrow(df\_study)

}

#Calculate 95% CI

mean\_CIs<-function(data, estimate){

results<-data.frame(Est=estimate)

results$LCI<-quantile(data[,estimate],probs=.025)

results$mean<-mean(data[,estimate])

results$UCI<-quantile(data[,estimate],probs=.975)

return(results)

}

#Auto-generate a results table

for (i in 1:length(colnames(results\_coll))){

if (i==1){

res\_tab\_coll<-mean\_CIs(results\_coll,colnames(results\_coll)[i])

} else{

res\_tab\_coll<-rbind(res\_tab\_coll, mean\_CIs(results\_coll,colnames(results\_coll)[i]))}

}

#Print table of results:

#Population average true, crude, and adjusted RRs and RDs across simulations and their 95% CI

#Sample average crude and adjusted RRs and RDs across simulations and their 95% CI

#Average prevalence of response types within exposure groups in population and sample across simulations and their 95% CI (no causal types because simulating under the null)

res\_tab\_coll

#Drop all objects except results tables

rm(list=grep("res\_tab",ls(),value=TRUE,invert=TRUE))

####################################################

##Section 4: Simulating Effect Modification

####################################################

niterations<-1000

blanks<-rep(NA,niterations)

results\_EM<-data.frame(

TrueRR=blanks, TrueRD=blanks,

TrueRR\_APOE1=blanks, TrueRD\_APOE1=blanks,

TrueRR\_APOE0=blanks, TrueRD\_APOE0=blanks,

CrudeRR=blanks, CrudeRD=blanks,

CrudeRR\_APOE1=blanks, CrudeRD\_APOE1=blanks,

CrudeRR\_APOE0=blanks, CrudeRD\_APOE0=blanks,

Doomed\_E1=blanks,Causal\_E1=blanks, Immune\_E1=blanks,

Doomed\_E0=blanks,Causal\_E0=blanks, Immune\_E0=blanks,

Doomed\_E1\_APOE1=blanks,Causal\_E1\_APOE1=blanks, Immune\_E1\_APOE1=blanks,

Doomed\_E0\_APOE1=blanks,Causal\_E0\_APOE1=blanks, Immune\_E0\_APOE1=blanks,

Doomed\_E1\_APOE0=blanks,Causal\_E1\_APOE0=blanks, Immune\_E1\_APOE0=blanks,

Doomed\_E0\_APOE0=blanks,Causal\_E0\_APOE0=blanks, Immune\_E0\_APOE0=blanks)

#Section 4A: Set prevalence of exogenous variables:

pZ1a<-0.04

pZ2<-0.1

pZ1b<-0.1

pAPOE<-0.25 #the effect modifier of explicit interest

pLowEd<-0.7 #the exposure ("E" in the formulas below)

#Section 4B: Simulate cohort of n=10,000 individuals

sampsize<-10000

set.seed(12345) #set a seed so that results can be replicated

for (i in 1:niterations){

#Simulate exogeneous variables

df<-data.frame(ID=seq(1,sampsize),

Z1a=rbinom(sampsize,1,pZ1a),

Z2=rbinom(sampsize,1,pZ2),

Z1b=rbinom(sampsize,1,pZ1b),

LowEd=rbinom(sampsize,1,pLowEd),

APOE=rbinom(sampsize,1,pAPOE),

Dementia=NA)

#Simulate endogenous variables based on SCCs

df$Dementia<-ifelse((df$LowEd==1 & df$Z1a==1) | df$Z2==1 | (df$LowEd==1 & df$APOE==1 & df$Z1b==1) ,1,0) #whether an individual has dementia or not (the outcome, "Y") depends on their value of causal components in the pies whose completion leads to dementia

#Section 4C: Calculate effects

#True effects (in the overall sample and within strata of modifier, APOE)

TrueRR<-(1\*pZ1a+pZ2+1\*pZ1b\*pAPOE-

1\*pZ1a\*pZ2-1\*pZ1a\*1\*pZ1b\*pAPOE-pZ2\*1\*pZ1b\*pAPOE+

1\*pZ1a\*pZ2\*1\*pZ1b\*pAPOE)/(pZ2)

TrueRD<-(1\*pZ1a+pZ2+1\*pZ1b\*pAPOE-

1\*pZ1a\*pZ2-1\*pZ1a\*1\*pZ1b\*pAPOE-pZ2\*1\*pZ1b\*pAPOE+

1\*pZ1a\*pZ2\*1\*pZ1b\*pAPOE)-(pZ2)

TrueRR\_APOE1<-(1\*pZ1a+pZ2+1\*pZ1b\*1-

1\*pZ1a\*pZ2-1\*pZ1a\*1\*pZ1b\*1-pZ2\*1\*pZ1b\*1+

1\*pZ1a\*pZ2\*1\*pZ1b\*1)/(pZ2)

TrueRD\_APOE1<-(1\*pZ1a+pZ2+1\*pZ1b\*1-

1\*pZ1a\*pZ2-1\*pZ1a\*1\*pZ1b\*1-pZ2\*1\*pZ1b\*1+

1\*pZ1a\*pZ2\*1\*pZ1b\*1)-(pZ2)

TrueRR\_APOE0<-(1\*pZ1a+pZ2+1\*pZ1b\*0-

1\*pZ1a\*pZ2-1\*pZ1a\*1\*pZ1b\*0-pZ2\*1\*pZ1b\*0+

1\*pZ1a\*pZ2\*1\*pZ1b\*0)/(pZ2)

TrueRD\_APOE0<-(1\*pZ1a+pZ2+1\*pZ1b\*0-

1\*pZ1a\*pZ2-1\*pZ1a\*1\*pZ1b\*0-pZ2\*1\*pZ1b\*0+

1\*pZ1a\*pZ2\*1\*pZ1b\*0)-(pZ2)

#Crude estimates

pY\_E0<-sum(df$LowEd==0 & df$Dementia==1)/sum(df$LowEd==0)

pY\_E1<-sum(df$LowEd==1 & df$Dementia==1)/sum(df$LowEd==1)

CrudeRR<-pY\_E1/pY\_E0

CrudeRD<-pY\_E1-pY\_E0

#Calculate crude estimates in strata of APOE

pY\_E0\_APOE1<-sum(df$LowEd==0 & df$Dementia==1 & df$APOE==1)/sum(df$LowEd==0 & df$APOE==1)

pY\_E1\_APOE1<-sum(df$LowEd==1 & df$Dementia==1 & df$APOE==1)/sum(df$LowEd==1 & df$APOE==1)

pY\_E0\_APOE0<-sum(df$LowEd==0 & df$Dementia==1 & df$APOE==0)/sum(df$LowEd==0 & df$APOE==0)

pY\_E1\_APOE0<-sum(df$LowEd==1 & df$Dementia==1 & df$APOE==0)/sum(df$LowEd==1 & df$APOE==0)

CrudeRR\_APOE1<-pY\_E1\_APOE1/pY\_E0\_APOE1

CrudeRD\_APOE1<-pY\_E1\_APOE1-pY\_E0\_APOE1

CrudeRR\_APOE1

CrudeRD\_APOE1

CrudeRR\_APOE0<-pY\_E1\_APOE0/pY\_E0\_APOE0

CrudeRD\_APOE0<-pY\_E1\_APOE0-pY\_E0\_APOE0

CrudeRR\_APOE0

CrudeRD\_APOE0

#For effect modification scenario we expect Crude measures = True measures because there are no sources of non-exchangeability

#Section 4D: "Under the hood" (define response types based on SCCs)

#Response types in total study sample

df$RT<-ifelse(df$Z2==1,"Doomed",ifelse(df$Z1a==1 | (df$APOE==1 & df$Z1b==1), "Causal","Immune") )

RTtab<-t(table(df$LowEd,df$RT)/c(sum(df$LowEd==0),sum(df$LowEd==1)))

RTtab

#Response types among those with APOE in the study sample

df\_APOE1<-df[df$APOE==1,]

RTtab\_APOE1<-t(table(df\_APOE1$LowEd,df\_APOE1$RT)/c(sum(df\_APOE1$LowEd==0),sum(df\_APOE1$LowEd==1)))

RTtab\_APOE1

#Response types among those without APOE in the study sample

df\_APOE0<-df[df$APOE==0,]

RTtab\_APOE0<-t(table(df\_APOE0$LowEd,df\_APOE0$RT)/c(sum(df\_APOE0$LowEd==0),sum(df\_APOE0$LowEd==1)))

RTtab\_APOE0

#Save results

results\_EM$TrueRR[i]<-TrueRR

results\_EM$TrueRD[i]<-TrueRD

results\_EM$TrueRR\_APOE1[i]<-TrueRR\_APOE1

results\_EM$TrueRD\_APOE1[i]<-TrueRD\_APOE1

results\_EM$TrueRR\_APOE0[i]<-TrueRR\_APOE0

results\_EM$TrueRD\_APOE0[i]<-TrueRD\_APOE0

results\_EM$CrudeRR[i]<-CrudeRR

results\_EM$CrudeRD[i]<-CrudeRD

results\_EM$CrudeRR\_APOE1[i]<-CrudeRR\_APOE1

results\_EM$CrudeRD\_APOE1[i]<-CrudeRD\_APOE1

results\_EM$CrudeRR\_APOE0[i]<-CrudeRR\_APOE0

results\_EM$CrudeRD\_APOE0[i]<-CrudeRD\_APOE0

results\_EM$Doomed\_E1[i]<-RTtab["Doomed","1"]

results\_EM$Doomed\_E0[i]<-RTtab["Doomed","0"]

results\_EM$Causal\_E1[i]<-RTtab["Causal","1"]

results\_EM$Causal\_E0[i]<-RTtab["Causal","0"]

results\_EM$Immune\_E1[i]<-RTtab["Immune","1"]

results\_EM$Immune\_E0[i]<-RTtab["Immune","0"]

results\_EM$Doomed\_E1\_APOE1[i]<-RTtab\_APOE1["Doomed","1"]

results\_EM$Doomed\_E0\_APOE1[i]<-RTtab\_APOE1["Doomed","0"]

results\_EM$Causal\_E1\_APOE1[i]<-RTtab\_APOE1["Causal","1"]

results\_EM$Causal\_E0\_APOE1[i]<-RTtab\_APOE1["Causal","0"]

results\_EM$Immune\_E1\_APOE1[i]<-RTtab\_APOE1["Immune","1"]

results\_EM$Immune\_E0\_APOE1[i]<-RTtab\_APOE1["Immune","0"]

results\_EM$Doomed\_E1\_APOE0[i]<-RTtab\_APOE0["Doomed","1"]

results\_EM$Doomed\_E0\_APOE0[i]<-RTtab\_APOE0["Doomed","0"]

results\_EM$Causal\_E0\_APOE0[i]<-RTtab\_APOE0["Causal","0"]

results\_EM$Causal\_E1\_APOE0[i]<-RTtab\_APOE0["Causal","1"]

results\_EM$Immune\_E1\_APOE0[i]<-RTtab\_APOE0["Immune","1"]

results\_EM$Immune\_E0\_APOE0[i]<-RTtab\_APOE0["Immune","0"]

}

#Calculate 95% CI

mean\_CIs<-function(data, estimate){

results<-data.frame(Est=estimate)

results$LCI<-quantile(data[,estimate],probs=.025)

results$mean<-mean(data[,estimate])

results$UCI<-quantile(data[,estimate],probs=.975)

return(results)

}

#Auto-generate a results table

for (i in 1:length(colnames(results\_EM))){

if (i==1){

res\_tab\_EM<-mean\_CIs(results\_EM,colnames(results\_EM)[i])

} else{

res\_tab\_EM<-rbind(res\_tab\_EM, mean\_CIs(results\_EM,colnames(results\_EM)[i]))}

}

res\_tab\_EM

#Drop all objects except results tables

rm(list=grep("res\_tab",ls(),value=TRUE,invert=TRUE))

####################################################

##Section 2b: Simulating confounding, non-null

####################################################

niterations<-1000

blanks<-rep(NA,niterations)

results\_confb<-data.frame(TrueRR=blanks, TrueRD=blanks,

CrudeRR=blanks, CrudeRD=blanks,

AdjRR=blanks, AdjRD=blanks,

Doomed\_E1=blanks,Causal\_E1=blanks, Immune\_E1=blanks,

Doomed\_E0=blanks,Causal\_E0=blanks, Immune\_E0=blanks)

#Section 2A: Set prevalence of exogenous variables

pQ1<-0.8

pQ2<-0.05

pZ1<-0.04

pZ3<-0.1

pZ4<-0.05

pLowSES<-0.7

#Section 2B: Simulate cohort of n=10,000 individuals

sampsize<-10000

set.seed(12345) #set a seed so that results can be replicated

for (i in 1:niterations){

#Simulate exogeneous variables

df<-data.frame(ID=seq(1,sampsize),

Q1=rbinom(sampsize,1,pQ1),

Q2=rbinom(sampsize,1,pQ2),

LowSES=rbinom(sampsize,1,pLowSES),

Z1=rbinom(sampsize,1,pZ1),

Z3=rbinom(sampsize,1,pZ3),

Z4=rbinom(sampsize,1,pZ4),

LowEd=NA,

Dementia=NA)

#Simulate endogenous variables based on SCCs

df$LowEd<-ifelse((df$LowSES==1 & df$Q1==1) | df$Q2==1,1,0) #in Scenario 2 (confounding), the exposure (LowEd or "E" in the formulas below) is no longer exogenous; each individual's value for LowEd emerges from the values of the components that cause it

df$Dementia<-ifelse((df$LowEd==1 & df$Z1==1) | df$Z3==1 | (df$LowSES==1 & df$Z4==1),1,0) #whether an individual has dementia or not (the outcome, "Y") depends on their value of causal components in the pies whose completion leads to dementia

#Section 2C: Calculate effects

#Truth (true ATE)

TrueRR<-(pZ1+pZ3+pLowSES\*pZ4-pZ1\*pZ3-pZ1\*pLowSES\*pZ4-pZ3\*pLowSES\*pZ4-pZ1\*pZ3\*pLowSES\*pZ4)/(pZ3+pLowSES\*pZ4-pZ3\*pLowSES\*pZ4)

TrueRD<-(pZ1+pZ3+pLowSES\*pZ4-pZ1\*pZ3-pZ1\*pLowSES\*pZ4-pZ3\*pLowSES\*pZ4-pZ1\*pZ3\*pLowSES\*pZ4)-(pZ3+pLowSES\*pZ4-pZ3\*pLowSES\*pZ4)

#Crude estimates

pY\_E0<-sum(df$LowEd==0 & df$Dementia==1)/sum(df$LowEd==0)

pY\_E1<-sum(df$LowEd==1 & df$Dementia==1)/sum(df$LowEd==1)

CrudeRR<-pY\_E1/pY\_E0

CrudeRD<-pY\_E1-pY\_E0

CrudeRR

CrudeRD

#Standardized (adjusted) estimates

pY\_E0\_C0<-sum(df$LowEd==0 & df$LowSES==0 & df$Dementia==1)/sum(df$LowEd==0 & df$LowSES==0)

pY\_E0\_C1<-sum(df$LowEd==0 & df$LowSES==1 & df$Dementia==1)/sum(df$LowEd==0 & df$LowSES==1)

pY\_E1\_C0<-sum(df$LowEd==1 & df$LowSES==0 & df$Dementia==1)/sum(df$LowEd==1 & df$LowSES==0)

pY\_E1\_C1<-sum(df$LowEd==1 & df$LowSES==1 & df$Dementia==1)/sum(df$LowEd==1 & df$LowSES==1)

pC1<-nrow(df[df$LowSES==1,])/nrow(df)

pC0<-1-pC1

pY\_E1\_std<-pY\_E1\_C1\*pC1+pY\_E1\_C0\*pC0

pY\_E0\_std<-pY\_E0\_C1\*pC1+pY\_E0\_C0\*pC0

AdjRR <- pY\_E1\_std/pY\_E0\_std

AdjRD <- pY\_E1\_std-pY\_E0\_std

AdjRR

AdjRD

#Section 2D: "Under the hood" (define response types based on SCCs)

df$RT<-ifelse((df$Z3==1 | (df$LowSES==1 & df$Z4==1 )),"Doomed", ifelse(df$Z1==1, "Causal","Immune") )

RTtab<-t(table(df$LowEd,df$RT)/c(sum(df$LowEd==0),sum(df$LowEd==1)))

RTtab

#Store results

results\_confb$TrueRR[i]<-TrueRR

results\_confb$TrueRD[i]<-TrueRD

results\_confb$CrudeRR[i]<-CrudeRR

results\_confb$CrudeRD[i]<-CrudeRD

results\_confb$AdjRR[i]<-AdjRR

results\_confb$AdjRD[i]<-AdjRD

results\_confb$Doomed\_E1[i]<-RTtab["Doomed","1"]

results\_confb$Doomed\_E0[i]<-RTtab["Doomed","0"]

results\_confb$Causal\_E1[i]<-RTtab["Causal","1"]

results\_confb$Causal\_E0[i]<-RTtab["Causal","0"]

results\_confb$Immune\_E1[i]<-RTtab["Immune","1"]

results\_confb$Immune\_E0[i]<-RTtab["Immune","0"]

}

#Calculate 95% CI

mean\_CIs<-function(data, estimate){

results<-data.frame(Est=estimate)

results$LCI<-quantile(data[,estimate],probs=.025)

results$mean<-mean(data[,estimate])

results$UCI<-quantile(data[,estimate],probs=.975)

return(results)

}

#Auto-generate a results table

for (i in 1:length(colnames(results\_confb))){

if (i==1){

res\_tab\_confb<-mean\_CIs(results\_confb,colnames(results\_confb)[i])

} else{

res\_tab\_confb<-rbind(res\_tab\_confb, mean\_CIs(results\_confb,colnames(results\_confb)[i]))}

}

#Print table of results:

#Average true, crude, and adjusted RRs and RDs across simulations and their 95% CI

#Average prevalence of response types within exposure groups across simulations and their 95% CI (no causal types because simulating under the null)

res\_tab\_confb

#Drop all objects except results tables

rm(list=grep("res\_tab",ls(),value=TRUE,invert=TRUE))